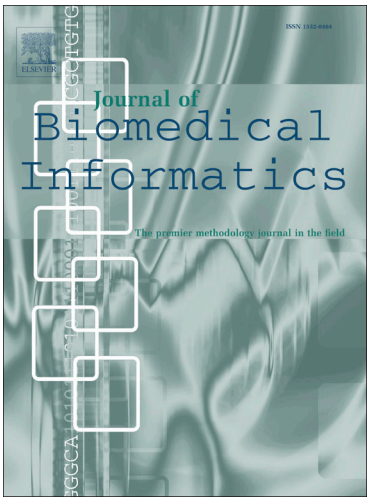


A frame semantic overview of NLP-based information extraction for cancer-related EHR notes

Surabhi Datta, Elmer V. Bernstam, Kirk Roberts

PII: S1532-0464(19)30221-7
DOI: https://doi.org/10.1016/j.jbi.2019.103301
Reference: YJBIN 103301

To appear in: *Journal of Biomedical Informatics*

Received Date: 2 April 2019
Revised Date: 4 September 2019
Accepted Date: 3 October 2019

Please cite this article as: Datta, S., V. Bernstam, E., Roberts, K., A frame semantic overview of NLP-based information extraction for cancer-related EHR notes, *Journal of Biomedical Informatics* (2019), doi: https://doi.org/10.1016/j.jbi.2019.103301

A frame semantic overview of NLP-based information extraction for cancer-related EHR notes

Surabhi Datta[1], Elmer V Bernstam[1,2], Kirk Roberts[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX USA
[2]Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, TX USA

Corresponding Author:
Kirk Roberts, PhD
7000 Fannin St #600
Houston TX 77030
kirk.roberts@uth.tmc.edu

ABSTRACT

***Objective*:** There is a lot of information about cancer in Electronic Health Record (EHR) notes that can be useful for biomedical research provided natural language processing (NLP) methods are available to extract and structure this information. In this paper, we present a scoping review of existing clinical NLP literature for cancer.

***Methods*:** We identified studies describing an NLP method to extract specific cancer-related information from EHR sources from PubMed, Google Scholar, ACL Anthology, and existing reviews. Two exclusion criteria were used in this study. We excluded articles where the extraction techniques used were too broad to be represented as frames (e.g., document classification) and also where very low-level extraction methods were used (e.g. simply identifying clinical concepts). 78 articles were included in the final review. We organized this information according to frame semantic principles to help identify common areas of overlap and potential gaps.

***Results*:** Frames were created from the reviewed articles pertaining to cancer information such as cancer diagnosis, tumor description, cancer procedure, breast cancer diagnosis, prostate cancer diagnosis and pain in prostate cancer patients. These frames included both a definition as well as specific frame elements (i.e. extractable attributes). We found that cancer diagnosis was the most common frame among the reviewed papers (36 out of 78), with recent work focusing on extracting information related to treatment and breast cancer diagnosis.

***Conclusion*:** The list of common frames described in this paper identifies important cancer-related information extracted by existing NLP techniques and serves as a useful resource for future researchers requiring cancer information extracted from EHR notes. We also argue, due to the heavy duplication of cancer NLP systems, that a general purpose resource of annotated cancer frames and corresponding NLP tools would be valuable.

INTRODUCTION

Unstructured, free-text clinical data about cancer patients is increasingly available in clinical notes in Electronic Health Records (EHRs) and other systems. There is increasing interest in utilizing this data for biomedical research [1][2] [3][4][5][6][7][8]. Furthermore, with the emergence of precision medicine, the need to extract more and more detailed information from a patient's EHR notes becomes ever greater. Many publications argue for the need to capture important clinical information on cancer, ranging from information about tissue specimens [9] to disease-related and outcome information [10], thus facilitating translational research by associating molecular information with disease phenotype [11]. Many of these information types are found nearly exclusively in unstructured or semi-structured text format in EHRs. For example, information associated with biomarkers and cancer prognosis are frequently stored in free-text surgical pathology reports [12]. Some work focuses on extracting information to improve cancer screening efficiency [1] [7][13][14][15]. All such work has focused on deriving cancer-related information automatically using various natural language processing (NLP) techniques. Although a majority of the methods applying deep learning in cancer research are for processing images (e.g. mammograms) [16][17][18][19] and gene expression profiles [20][21][22], more recently, deep-learning based NLP systems are gaining prominence for cancer information extraction from EHRs [23][24][25] [26][27]. For example, Gao et al. [26] implemented a hierarchical attention network for extracting some of the crucial clinical oncology data elements such as primary cancer site and histological grade which are gathered by cancer registries.

However, many of these researchers put sizable effort into designing and implementing NLP systems that extract similar information types. In this scoping review, we investigate which cancer information types have been extracted with NLP techniques. We organize the extracted information into frames, based on the linguistic theory of 'frame semantics'. Frames provide a convenient, flexible, and linguistically-motivated representation for information as complex and diverse as that related to cancer. Our aim is to provide a list of cancer-related frames in existing work that would be valuable to the scientific community.

Frame semantics is a linguistic theory that postulates the meaning of most words is understood in relation to a conceptual frame in which entities take part. E.g., the meaning of sell in the "*Jerry sold a car to Chuck*" evokes a frame related to COMMERCE, which includes four elements: BUYER, SELLER, MONEY, and GOODS, though not all elements are required (as with MONEY here). Frames can represent all events types, from the simple PLACING frame ("*Maria put money in her account*") with elements AGENT, THEME, and GOAL, to the more complex HIRING ("*He hired Downing as his coach in Hawai'i*") with elements EMPLOYER, EMPLOYEE, POSITION, and PLACE. Frames can also encode relations (e.g., KINSHIP: "*Joe's brother John*") and states (e.g., BEING_IN_OPERATION: "*the centrifuge is operational*"). For a medical frame example, consider that a radiology report may contain tumor information such as "*There is a single lesion in segment 7 measuring 1.5 x 1.9 cm*". This evokes a TUMOR DESCRIPTION frame with three elements present: COUNT ("*Single*"), ANATOMICAL SITE ("*segment 7*"), and SIZE ("*1.5 x 1.9 cm*").

Currently, the Berkeley FrameNet database [28] contains more than 1,200 frames with an average of about 10 elements per frame. Manual annotations of more than 200,000 frame instances provide a unique level of detail about how these elements can appear in sentences. The frames are connected by more than 1,800 frame relations, forming a lattice structure. As exemplified by the above examples, however, FrameNet has minimal coverage of biomedical information. This review may be viewed, then, as part of a proposed set of frames specifically targeting cancer information in EHR notes. These frames are not constructed with the intent of adding them to Berkeley FrameNet. Rather, these can be thought of as an auxiliary frame specification, which is commonly done [29][30]. On the other hand, frame semantics simply provides a useful mechanism for organizing the wide variety of information found in the papers covered by this review. We make no claim these frames are the 'best' way to organize cancer information for non-NLP purposes. These frames are, however, a natural endpoint for a scoping review, as they help to succinctly define the scope of cancer information that has been extracted from EHR free text.

MATERIALS AND METHODS

The goal of this study is to review papers on NLP for EHR notes related to cancer, determine

what (parts of) frame(s) the researchers aim to extract, then express the complete set of frames in a consistent way across all the identified papers. The focus is thus the type of cancer data that is extracted using NLP, rather than on the NLP method used or its performance. PubMed and Google Scholar were searched using the keywords 'natural language processing' or 'NLP' anywhere in the article and one of the three keywords from 'cancer', 'tumor', and 'oncology' in the title. We also searched the ACL Anthology using only the cancer related keywords ('cancer', 'tumor', or 'oncology') in the title. Results were limited to papers published after January 1, 2000 in order to capture the current generation of statistical and machine learning-based NLP (though we did not exclude articles after this date based on method). The last inclusion date for the search was September 20, 2018. Additionally, citations from two cancer NLP reviews [31][32] were used. Finally, relevant citations in the reviewed papers were iteratively added to the review process.

We obtained a total of 899 articles using the above search process (108 via PubMed, 738 via Google Scholar, 16 via ACL Anthology, and 37 via the existing reviews). After de-duplication, 703 articles were chosen for title/abstract screening.

Screening Process

We selected relevant articles based on whether the title or abstract contained the description of an NLP method to extract cancer-related information from EHR notes, or suggests cancer-related information is extracted from free text EHR sources. Other types of unstructured data (e.g., images, waveforms) were excluded. Two raters (SD, KR) independently assessed whether the title or abstract suggested that the data source was not related to EHR notes (e.g., literature [33], social media [34][35]). Papers describing NLP on data sources other than EHR notes were excluded. If the data source was not clear, or the raters could not agree during reconciliation, the paper was kept for full-text review.

For the 173 papers that passed the initial title/abstract screening, two further exclusion criteria were applied to the full text. First, document-level text classification methods were excluded. Not only are these not considered information extraction NLP methods and are not amenable to frame representation, they are often application specific and not re-purposable. For example, an

article by Garla et al. [36] describing a binary classifier to determine whether a clinical report is about a potential malignant liver lesion requiring follow up would not be considered for this study. This is a very common use case of NLP for cancer research, but these methods do not attempt to identify specific cancer-related data elements, and are thus not amenable to frames. On the other hand, an article that extracts references to malignant liver lesions and factors (e.g., phrases) that may impact the need for follow up would be relevant for this study since those factors could be organized by frame. Second, articles were excluded where very unspecific extraction methods were used like concept recognition and named entity recognition techniques (such as the use of MetaMap [37] or cTAKES [38]) that identify phrases but do not attempt to connect these to their wider context [39][40][41]. For example, an article by Xie et al. [41] that does not differentiate between an asserted and a negated concept, such as extracting the phrase "breast cancer" regardless of the context being "has breast cancer" vs. "has no breast cancer", is not sufficiently semantically expressive and would thus be excluded. Similarly, an article describing the use of a general-purpose concept extraction tool on mammography reports would not be considered relevant for frame representation. These types of NLP articles can be viewed as describing the basic building blocks upon which a frame-based information extraction system can be built, but they do not capture sufficient contextual information to be properly described as frames. In cases where the exclusion decision could not be made from the title and abstract alone, the full text was screened. In total, title/abstract screening resulted in removing 530 articles and the full-text review of the remaining 173 articles eliminated a further 95 articles. Finally 78 articles were included in this scoping review. Figure 1 shows the overall process including numbers removed at each stage.

Frame Construction

For each of the 78 articles relevant to frame semantics for cancer in EHR notes, one or more frames were constructed to represent the information types extracted by the described NLP system. Note that NLP systems are not typically described in a frame semantic manner nor use the terminology of frames. Instead, this had to be inferred based on the descriptions. For papers that focus on one or two data types, this is a straightforward process that results in a single, small frame description or a single element within a larger frame. For papers that focus on multiple data types (at least 7 papers describe over 5 different data types), the data types were organized

into one or more frames. Multiple frames were used when the information was conceptually different (e.g., the description of a tumor versus the description of a cancer procedure). However, this was an iterative, collaborative process: batches of 20 papers were considered at a time, and two reviewers proposed initial frames for each of the papers. Disagreements were reconciled, after which the next batch was considered, which may have resulted in new frames as well as modifications to frames created in previous batches. A common set of frames was created describing important cancer-related information extracted using NLP techniques. Finally, both a cancer informatics expert (EVB) as well as a practicing oncologist with informatics experience (FMB) reviewed the frames to ensure medical validity.

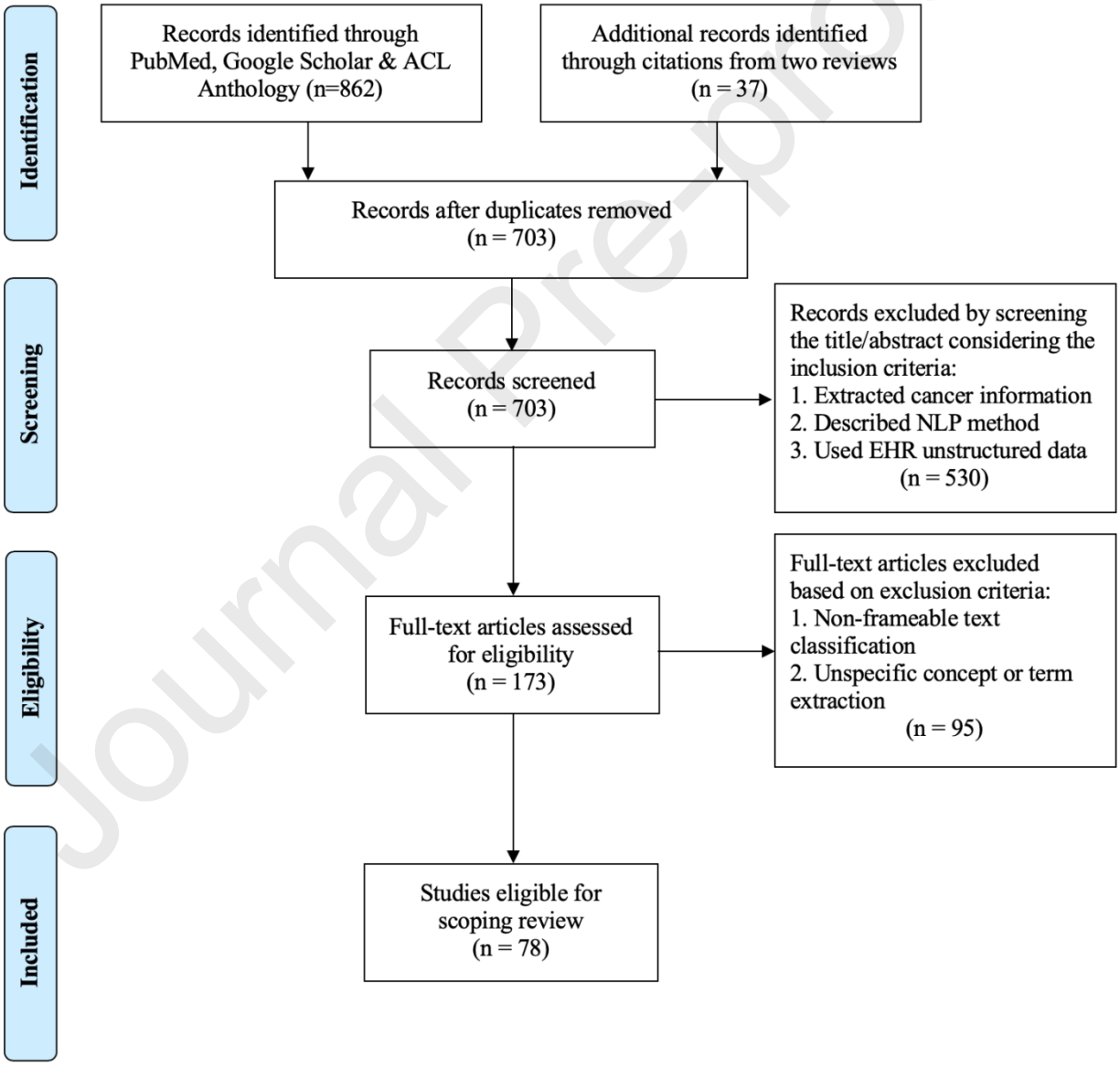


Figure 1: PRISMA diagram for study selection

RESULTS

Frame Descriptions

Each constructed frame has a name along with a definition (see Figure 2). Moreover, each frame is described using some attributes or elements, known as 'frame elements'. The frame elements correspond to the specific items extracted by one or more NLP system included in this study. Each of these elements is defined briefly along with identifying the corresponding articles that extract that frame element (Table 1). For example, 'CANCER DIAGNOSIS' is a frame, which is defined as 'an event of a patient being diagnosed with a cancer', and described using frame elements such as ANATOMICAL SITE, HISTOLOGY, and DIAGNOSIS STATUS.

Frame Relations

The constructed frames additionally have relations between them based on those found in Berkeley FrameNet [28]. First is the 'parent-child' or 'inheritance' relation ('inherits' in Figure 2). More specifically, every child frame is a specific version of the parent frame. For example, 'CANCER DIAGNOSIS' frame is inherited by multiple frames such as 'BREAST CANCER DIAGNOSIS' and 'BLADDER CANCER DIAGNOSIS'. The elements in 'CANCER DIAGNOSIS' frame are generic to diagnosis of any cancer type, whereas 'BREAST CANCER DIAGNOSIS' frame contains extra breast cancer-specific elements such as RECEPTOR STATUS and RECEPTOR PERCENT. Similarly, an element such as MUSCULARIS PROPRIA is more prevalent in bladder cancer diagnosis cases.

The second relation is 'an element of' where a frame is an element described in another frame. This is analogous to the 'Subframe' relation proposed in the Berkeley FrameNet project. As mentioned in Table 1, TNM CLASSIFICATION is an element present in both 'CANCER DIAGNOSIS' and 'TUMOR DESCRIPTION' frames, while at the same time, 'TNM CLASSIFICATION' is described as a separate frame altogether. A separate frame for a frame element was created if the element has multiple attributes and at least one included paper extracts such attributes. For example, AAlAbdulsalam et al. [2] extracts elements which are very specific to TNM stage such as the actual stage designation (e.g. T1, N1b), staging method (e.g. clinical or

pathological staging), and negation. This information can be better understood if a separate frame 'TNM CLASSIFICATION' is created to contain TNM staging related details.

The third frame relation is 'associated with', which indicates any kind of association or dependency between two frames. For example, the 'CANCER FINDING' frame is associated with the 'THERAPEUTIC PROCEDURE' frame as the findings in clinical reports often aid in selecting the best therapy for treating cancer patients.

Although existing work extracted cancer information for numerous cancer types, we only created child frames for those types where any type-specific information was extracted. There were papers related to cancer types such as lung cancer [42][43], liver cancer [44] and ovarian cancer [45], but no frames were created for these types as they all extract the general cancer-related elements described in the 'CANCER DIAGNOSIS' frame.

Table 1 summarizes the frames created from the 78 selected papers, with corresponding per-element references. Supplementary Table 1 lists the per-frame references. The frame descriptions and the overall relationships between the frames are shown in Figure 2. Frames with similar purpose (e.g., diagnosis, imaging, assessment) are assigned the same colors in Figure 2. The frames in Table 1 are organized in a manner such that the ones which are similar (also corresponding to the colors in Figure 2) appear together. Among the set of similar frames, a broader frame is followed by the other related frames. Note that we start the table with the 'CANCER DIAGNOSIS' frame.

Table 1: Frames (including frame elements) along with the associated articles. In case a frame element is described as a frame, the frame name is placed in brackets.

Frame	Frame Elements	References
CANCER DIAGNOSIS	NAME: the cancer type. Often duplicated as the ANATOMICAL SITE or HISTOLOGY, but here specifically refers to the term used to declare a diagnosis	[44], [46], [47], [48], [49], [50], [51], [52], [53]
	ANATOMICAL SITE: the location description of the finding (including primary and metastatic sites)	[47], [54], [44], [55], [56], [57], [25], [27], [58], [59]
	HISTOLOGY: histological description (e.g. carcinoma)	[46], [54], [60], [57], [55], [56], [4], [27], [61], [45], [59]
	GRADE: appearance of the cancerous cells, can be frame with further information [GRADING VALUE]	[46], [54], [56], [4], [50], [27], [61], [62], [45], [63], [64]
	INVASION TYPE: patterns of invasive growth, migration type of cell(s), or invasion mechanism (e.g., collective migration, individual cell migration)	[54]
	TUMOR BLOCK: tissue cores removed from regions of interest in paraffin-embedded tissues (e.g. 0.6 mm in diameter)	[54]
	TISSUE BANK: identifiers about location of tissue samples within an institution	[54]
	STATUS: whether confirmed, suspected and there is no evidence of finding (e.g. probable, definite, without)	[44], [59]
	RECURRENT STATUS: the value of recurrent status	[44], [65], [59]
	TEMPORAL INFORMATION: refers to information about time (e.g., year, month, and date, 2007-08-04)	[44], [59]
	SPECIMEN TYPE: the type of specimen involved in diagnosis	[55]
	LATERALITY: describes the side of a paired organ associated with origin of the primary cancer	[55], [56], [25], [50], [66], [27], [53]
	TUMOR SIZE: how large across the tumor is at its widest point (part of cancer staging)	[54], [55], [56], [25], [50], [67], [61], [62], [64], [59]
	TNM STAGE: cancer staging system, can be a separate frame with further	[57], [55], [2], [3], [25], [68], [69], [62], [52], [42],

Frame	Frame Elements	References
	information [TNM CLASSIFICATION]	[63]
	EXTENSION: direct extension of tumor	[55]
	UNCERTAINTY: used to differentiate clinical suspicions from conclusive findings (e.g., possible, likely)	[70], [59]
	NEGATION: existence/negation of diagnosis (e.g., no, positive)	[70], [59]
	DISEASE: disease related concepts related to diagnosis (disease stage and severity)	[11], [25]
	DISEASE EXTENT: determine extent of disease (e.g., non-invasive, invasive, or metastatic)	[70], [71], [59]
	STAGE: the overall stage of cancer (e.g. Stage 0, Stage I)	[72], [54], [44], [56], [25], [71], [42], [45], [63], [59]
	EXISTENCE: existence description of the finding	[47]
	TEMPORAL MODIFIER: temporal modifiers of the finding	[47]
	ASSOCIATION: associations with other findings (may or may not be related to cancer) (e.g. causal, differential interpretation, and co-occurring)	[47]
BREAST CANCER DIAGNOSIS	STATUS OF CANCER TYPES: presence or absence of various types of breast cancer (e.g. ductal carcinoma in situ, invasive lobular carcinoma)	[49], [66], [51], [67], [24]
	RECEPTOR NAME: estrogen, progesterone, human epidermal growth factor 2	[73], [74], [25]
	RECEPTOR STATUS: positive/negative	[74], [50], [68], [66], [71], [67], [62]
	RECEPTOR STATUS NEGATION: negation of the RECEPTOR STATUS	[74]
	RECEPTOR PERCENT: number of cells out of 100 that stained positive for a receptor	[50], [62]
	EXTRACAPSULAR AXILLARY NODAL EXTENSION STATUS: presence/absence of extracapsular extension in axillary lymph nodes	[66]
	ISOLATED CANCER CELLS IN LYMPH	[66]

Frame	Frame Elements	References
	NODES STATUS: presence/absence of isolated cancer cells in sentinel lymph nodes	
	MENOPAUSAL STATUS: status of menopause	[71]
	SCARFF-BLOOM-RICHARDSON (SBR) STAGE -prognostic factor in breast cancer, associated with cell proliferation, also a consistent indicator of response to chemotherapy	[67]
	CONTRALATERAL EVENT: event of detecting a tumor in the opposite breast which was diagnosed more than 6 months following the detection of the first cancer	[53]
	MEDIASTINAL AND/OR STERNAL INVOLVEMENT: metastatic to sternum or mediastinum	[75]
COLORECTAL CANCER DIAGNOSIS	POLYP TYPE: type of polyp present at the time of colonoscopy procedure (e.g., advanced conventional adenomas)	[61]
BLADDER CANCER DIAGNOSIS	INVASION STATUS: presence or absence of invasion	[46], [60], [4]
	DEPTH OF INVASION: measured from the basement membrane of epithelium from which the tumor is considered to arise, to the deepest point of invasion (e.g., superficial and muscle invasion)	[46], [60], [4], [52]
	MUSCULARIS PROPRIA: presence or absence of muscle in the specimen	[46], [60], [4]
	CARCINOMA IN SITU: statements regarding presence of carcinoma in situ	[60], [4], [69], [63]
SKIN CANCER DIAGNOSIS	CLARK LEVEL: described in separate frame [CLARK LEVEL]	[5]
	BRESLOW DEPTH: described in separate frame [BRESLOW DEPTH]	[5]
PROSTATE CANCER DIAGNOSIS	GLEASON SCORE: described in separate frame [GLEASON SCORE]	[5], [70], [3]
	PSA: prostate-specific antigen value	[3]
PAIN IN PROSTATE CANCER	MENTION: mention of pain related term	[76]
	PAIN SEVERITY: associated with a severity level from the four-tiered pain	[76]

Frame	Frame Elements	References
PATIENTS	scale (no pain - category 0; some pain - category 1; controlled pain - category 2; severe pain - category 3)	
	INTERNAL DATES: pain start and end dates	[76]
	LOCATION: body location of pain	[76]
	NEGATION: negated mentions of pain related terms	[76]
COMORBIDITY DIAGNOSIS	NAME: name of the comorbidity diagnosis (e.g. Liver cirrhosis)	[44]
	DIAGNOSIS STATUS: Whether confirmed, suspected and there is no evidence of finding (e.g. probable, definite, without)	[44]
	CHILD-PUGH STAGING: e.g. Child-Pugh score: class A)	[44]
	TEMPORAL INFORMATION: refers to information about time	[44]
	REPORT TYPE: type of the report (e.g. Computed Tomography)	[44]
FAMILY HISTORY	FAMILY MEMBER: associated member	[77]
	DIAGNOSIS: associated cancer diagnosis	[77]
	RELATION: Relation between the family member and diagnosis	[77]
	NEGATION: negated mention of history	[77]
KI-67 EXPRESSION	KI-67 SCORE: a cancer proliferation marker that is expressed during cell growth and division, but is absent in the cell resting phase	[6], [78], [50], [67], [64]
PERFORMANCE STATUS	NAME: name of the performance status scale, used as a prognostic tool (e.g. ECOG, Karnofsky)	[57], [42]
	SCORE: measure of functional status (ECOG score ranges from 0-5, whereas Karnofsky Scale' ranges from 0-100)	[42]
GRADING VALUE	SCALE: cancer is usually graded on a scale of 1-3 (lower number indicates cancer cells look more similar to normal cells)	[5], [54]
	TYPE: grading categories (e.g. Grade 1 – well differentiated)	[54], [56], [4], [45], [63]
CLARK LEVEL	VALUE: describes the level of anatomical invasion of the melanoma	[5]

Frame	Frame Elements	References
	in the skin	
BRESLOW DEPTH	VALUE: the distance between the upper layer of the epidermis and the deepest point of tumor penetration	[5]
GLEASON SCORE	SCORE: grading system used to determine the aggressiveness of prostate cancer	[5], [70], [79]
TNM CLASSIFICATION	TUMOR SIZE: diameter/volume of the tumor, including unit (e.g., 3-4 mm)	[6], [69], [62]
	REGIONAL LYMPH NODES INVOLVED: regional lymph nodes information used to detect staging	[6], [80]
	METASTASIS: whether the tumor has invaded the nearby tissues	[6], [52]
	DISTANT METASTASIS: whether tumor has spread from the original (primary) tumor to distant organs or distant lymph nodes	[6], [80]
	GRADE: appearance of the cancerous cells	[6], [69], [62]
	STAGING FACTORS: factors relevant to staging and used to calculate the TNM stage	[80]
	STAGE: AJCC stage designation (e.g., T1, N1b)	[2], [3], [25], [68], [62], [52], [42], [63]
	TIMING: used to indicate if the staging is clinical or pathological as per the rules of the AJCC manual	[2], [3], [25], [68], [52], [42]
	NEGATION: any negated mention	[2]
	TEMPORALITY: capture historical or future mentions that do not necessarily represent current mentions valid at the point in time when the mention was stated at the patient record	[2]
	SUBJECT INVOLVED: capture TNM mentions related to family relatives or others	[2]
TUMOR DESCRIPTION	ANATOMICAL SITE: anatomic locations (e.g., “segment 5” or “left lobe”) with attributes (Liver, Non Liver), target location (e.g., liver and segment #7) as well as non-target location (e.g., breast)	[13], [8], [54], [81], [44], [25], [9], [23], [26], [43]
	LATERALITY: side of a paired organ	[25]

Frame	Frame Elements	References
	associated with origin of the primary tumor	
	TYPE: primary/metastatic	[25]
	STATUS: benign or malignancy status along with diagnostic information such as 'suggestive of cyst'	[81], [82], [43], [59]
	HISTOLOGY: morphologic type of the tumor	[54], [9]
	STAGE: tumor stage	[8], [70]
	GRADE: appearance of the cancerous cells	[8], [54], [9], [50], [83], [26]
	INVASION: whether or not more than 50% of an organ is invaded	[54], [84]
	SIZE: quantitative size of tumor (e.g., 2.2 x 2.0 cm), diameter/volume of the tumor, including unit (e.g., 1 cm, 0.3 x 0.5 x 0.7 cm)	[13], [54], [84], [81], [44], [25], [50], [43]
	SIZE TYPE: radiological/pathological	[25]
	NEGATION: indicator to some negation of a tumor reference (e.g., no)	[81], [82], [43]
	COUNT: number of tumor/nodule references (e.g., two or multiple)	[13], [84], [81]
	TUMOR REFERENCE: a radiologic artifact that may reference a tumor (e.g., lesion or focal density)	[84], [81]
	MENTION: tumor major object (e.g., tumor, lesion, mass, and nodule)	[13], [44], [49], [51], [43]
	QUANTIFIER: one, two, three, several	[44]
	TEMPORAL INFORMATION: refers to information about time (e.g., year, month, and date, 2007-08-04)	[54], [44], [82]
	NON-TUMOR SIZE ITEMS: LeVeen needle, which is used in RFA treatment	[44]
	STATUS: this indicates the final overall tumor status (e.g., regression, stable, progression, irrelevant)	[85], [43]
	METASTATIC STATUS INDICATORS: phrases denoting a metastatic tumor	[9], [86]
	MAGNITUDE: indicates the qualitative extent of change, if any (e.g., mild, moderate, marked)	[85]
	SIGNIFICANCE: indicates the subjective clinical significance of change, if any	[85]

Frame	Frame Elements	References
	(e.g., uncertain, possible, probable)	
	TRACK ASSIGNMENT: Fleischner Society based surveillance track assignment for patients who received LCS LDCT (guidelines about follow-up procedure depending on the nodule size)	[13]
	PROCEDURE: method for obtaining the tumor, can be a separate frame [CANCER PROCEDURE]	[13], [9], [52]
	TNM STAGE: cancer staging system, a separate frame with further information [TNM CLASSIFICATION]	[54], [83], [52]
	CLOCK-FACE: clock-face location of the tumor	[25]
BLADDER TUMOR DESCRIPTION	MUSCLE INFORMATION: presence or absence of muscle, for those that mentioned muscle, rates of muscle presence in the surgical specimen	[8], [83]
	CARCINOMA IN SITU: statements regarding presence of carcinoma in situ	[83]
ASSESSMENT FOR CANCER CARE	SURGERY RELATED: described in separate frame [ASSESSMENT FOR CANCER SURGERY]	[79]
	PRE-TREATMENT RELATED: pretreatment process quality measures, described in frame [PRE-TREATMENT CANCER ASSESSMENT]	[87]
ASSESSMENT FOR CANCER SURGERY	TNM STAGE: cancer staging system, a separate frame with further information [TNM CLASSIFICATION]	[79]
	MARGIN STATUS: status of surgical margin, described in separate frame [MARGIN]	[79]
ASSESSMENT FOR PROSTATE CANCER SURGERY	GLEASON SCORE: described in separate frame [GLEASON SCORE]	[79]
PRE-TREATMENT CANCER ASSESSMENT	DOCUMENTATION: documentation within 6 months prior to initial treatment (e.g. treatment of prostate-specific antigen)	[87]

Frame	Frame Elements	References
	PROCEDURE: performing diagnostic test, described in frame [CANCER PROCEDURE]	[87]
PRE-TREATMENT PROSTATE CANCER ASSESSMENT	GLEASON SCORE: described in separate frame [GLEASON SCORE]	[87]
MARGIN	TYPE: type (e.g., anatomical, surgical)	[54]
	STATE/STATUS: if the inked margin of resection is found to contain tumor, it indicates that the tumor extended to the tissue that was cut, resulting in a positive surgical margin (e.g. positive surgical margin, negative surgical margin, not applicable or no explicit diagnosis provided)	[54], [79]
	DIMENSION: margin size	[54]
CANCER PROCEDURE	NAME: name of the procedure/test (e.g. chest x-ray)	[6], [88], [89], [78], [73], [90], [15], [1], [9], [91], [92], [62], [52], [45]
	CODE TERMINOLOGY: clinical terminology used for the procedure mention	[54]
	CODE VALUE: value of the terminology code	[54], [1]
	INSTITUTION: institution where the procedure was performed	[54]
	NEGATION: whether the mention is negated	[54], [87]
	MENTION: words related to any procedure term (e.g. flex sig, guaiac card)	[78], [54], [15], [93], [87], [45]
	MARGIN: usually the rim of normal tissue taken removed during or after procedure (surgical margin)	[54]
	ANATOMICAL SITE: part of body procedure targets (e.g., breast)	[88], [54]
	TEMPORAL INFORMATION: time and date descriptors (e.g., "colonoscopy in 2005", "flexible sigmoidoscopy 5 years ago), date of completion	[54], [15], [1], [10]
	STATUS: procedure or treatment status (e.g., refused, declined, scheduled, planned, completed, reported vs not	[15], [44], [1], [91]

Frame	Frame Elements	References
	reported)	
	MODIFIER: negation and other modifiers that change the status of procedure (e.g., "no", "never")	[15]
DIAGNOSTIC PROCEDURE	PROCEDURE RESULT: result of a procedure mentioned in the report (e.g., Calcifications, hyperplasia etc.)	[88]
	TEST RESULT: result of the test (e.g. positive vs negative)	[14], [91], [94], [92], [87]
THERAPEUTIC PROCEDURE	TYPE: treatment type (e.g., Radiofrequency ablation and Transarterial chemoembolization)	[44], [62], [52]
	LINE OF THERAPY: initial treatment is referred to as first-line treatment or first-line therapy, however, a second-line treatment may be suggested later	[52]
	THERAPY DOSE: the total about of treatment (e.g., radiation) the patient is exposed to (e.g. radiation therapy dose)	[10]
	TOXICITIES: various toxicities related to cancer treatment therapy along with negation and certainty	[10]
CANCER TREATMENT	TYPE: drug, procedure, radiation, etc.	[25], [10], [45]
	PATIENT CENTERED OUTCOMES: outcomes as interpreted and documented by clinicians following a patient's cancer treatment along with their semantic context such as negations	[95]
	PATIENT REPORTED OUTCOMES: symptoms experienced by patients during cancer treatment along with negations	[96]
	SITE: metastatic site of cancer where the treatment was targeted	[10]
CANCER FINDING	PROCEDURE: name of the associated cancer procedure (e.g., Breast Core biopsy), can be a separate frame [CANCER PROCEDURE]	[7], [88], [73], [90], [11]
	FINDING TYPE: type of the finding (e.g., negative, normal, positive, possible, probably, history, mild, stable,	[89]

Frame	Frame Elements	References
	improved, or recommendation)	
	FINDING MODIFIER: modifying words from within the report (e.g., type: mild, modifiers: tortuous)	[89]
	LATERALITY: location or sidedness of the finding (e.g., “left”, “right,” "both", or “bilateral”)	[88]
	BODY PARTS: body organs on which the finding is reported	[11]
PATHOLOGY FINDING	POSITIVE LYMPH NODE NUMBER: number of positive lymph nodes mentioned in the finding	[73], [50], [67]
	POSITIVE LYMPH NODE STATUS: presence or absence of positive lymph nodes	[66]
	LYMPH NODES REMOVED: number of lymph nodes removed	[50]
	NUCLEAR GRADE: describes how closely the nuclei of cancer cells look like the nuclei of normal cells	[73]
	PLOIDY: refers to amount of DNA the cancer cells contain	[73]
	QUALITATIVE S-PHASE: indicator of tumor growth rate	[73]
	BIOMARKER: name of the biomarker	[12], [42]
	BIOMARKER TEST RESULTS / MUTATION STATUS: positive or negative for biomarkers such as ALK and EGFR	[12], [42], [57], [67]
	SPECIMEN: specimens on which findings are described (e.g. core biopsies or organ portions)	[68]
	STATUS OF INVASION: presence or absence of invasion such as blood vessel invasion	[66]
IMAGING FINDING	CALCIFICATION: presence or absence of calcification	[7], [97], [98], [24]
	MASS: presence or absence of mass	[7], [98], [24]
	ARCHITECTURAL DISTORTION: whether architectural distortion is present or absent	[7], [98], [24]
	CYSTS: whether cyst is present or absent	[7], [97]
	NME: whether non-mass enhancement	[7]

Frame	Frame Elements	References
	is present or absent	
	FOCUS: whether focus is present or not in imaging such as breast magnetic resonance imaging	[7]
	RECOMMENDATION: follow-up recommendation	[90]
	CALCIFICATION CHARACTERISTICS: characteristics of calcification	[98]
	MUTATION CHARACTERISTICS: imaging descriptors/features to distinguish wild-type vs mutated patients (e.g. distinguish wild-type and KRAS mutations)	[99]
BREAST IMAGING FINDING	IMPLANTS: whether breast implants are present or absent	[7]
	ASYMMETRY: whether asymmetry is present or absent	[7], [98], [24]
	BIRADS CATEGORY: refers to the number stated for each breast and laterality (e.g. 1 for negative category)	[7], [90]
	DENSITY: breast density or breast composition	[98], [24]

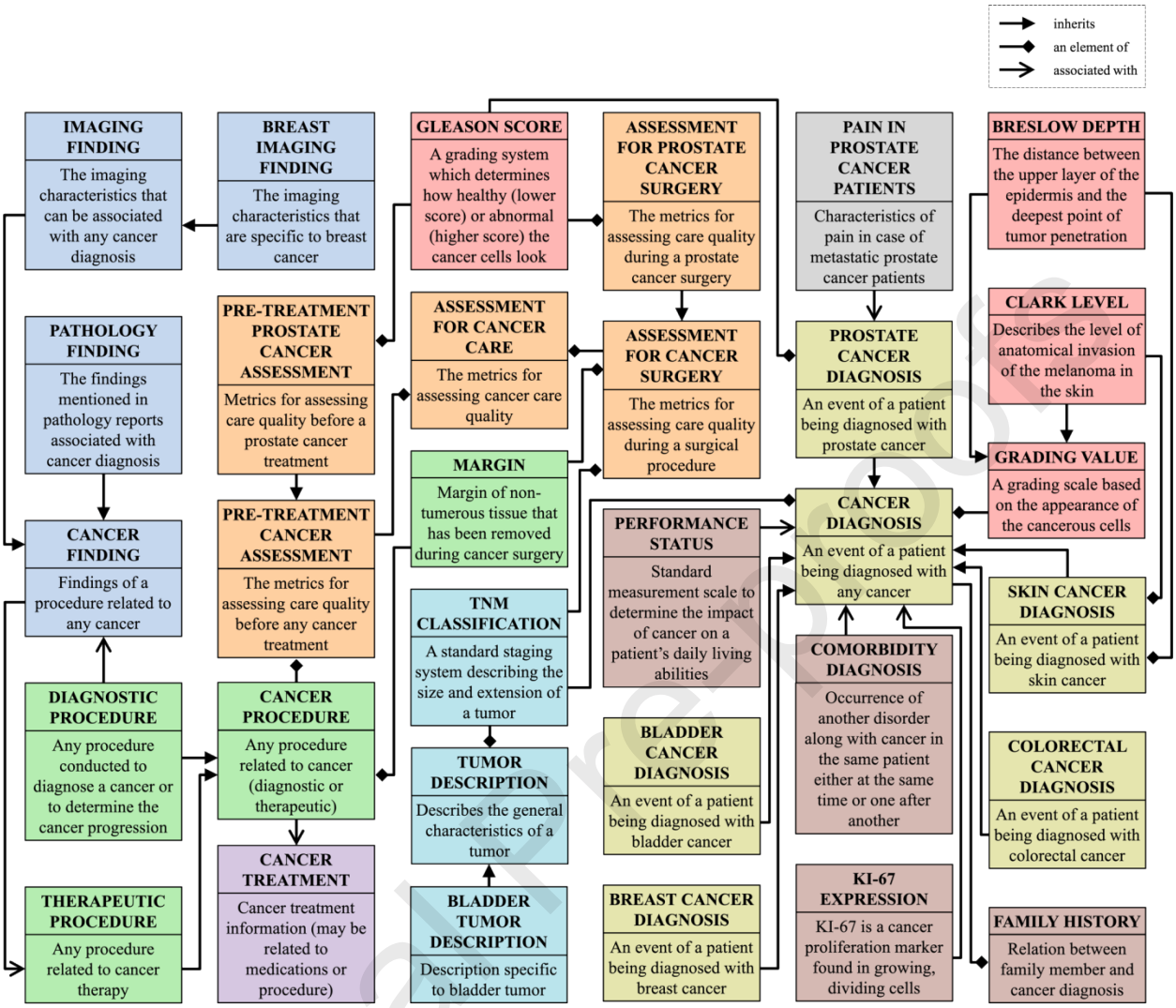


Figure 2: Frames and their relations as expressed in literature. Frames with similar purpose (e.g., diagnosis, imaging, assessment) are assigned the same colors.

Most Extracted Frames:

As demonstrated in Table 1, 'CANCER DIAGNOSIS' is the most referenced frame overall, i.e. it has the highest number of associated articles (n=36). This was expected as most efforts focused on extracting general information related to diagnosis of cancer. However, this does not necessarily mean that a paper included in 'CANCER DIAGNOSIS' frame extracted all the elements present in that frame. We observe that a subset of papers extracting some specific elements of this frame also extracted elements present in one of its child frames. For instance, a breast cancer related article by Breischneider et al. [62] extracted general cancer data elements such as TNM CLASSIFICATION and GRADE present in 'CANCER DIAGNOSIS' frame whereas it

also extracted RECEPTOR STATUS and RECEPTOR PERCENT elements described in the 'BREAST CANCER DIAGNOSIS' frame. Thus, some papers extract only a subset of elements present in one or more frames and not all of the elements related to a single frame. In terms of the total number of articles associated with a frame, the 'CANCER DIAGNOSIS' frame is followed by 'TUMOR DESCRIPTION' (*n*=21) and CANCER PROCEDURE' (*n*=19) frames.

Most Extracted Frame Elements

We also note the distribution of papers at the frame element level. This provides an idea about the important clinical attributes found in EHR text related to cancer that interest researchers. Table 2 highlights the 10 most extracted frame elements along with the number of referenced articles.

Table 2: Most extracted frame elements along with the number of referenced articles

Frame Element	Number of articles extracting this element
GRADE	19
ANATOMICAL SITE	18
TUMOR SIZE	17
TNM CLASSIFICATION	16
CANCER PROCEDURE NAME	16
STAGE RELATED (Overall + TNM)	15
RECEPTOR/BIOMARKER RELATED	13
HISTOLOGY	13
STATUS (tumor and cancer related)	9
LATERALITY	8

Papers Extracting Multiple Frames

Many eligible papers extracted varied types of cancer information corresponding to different frames. The NLP system developed by Savova et al. [25] extracted various cancer phenotypes from EHRs, which were then used to generate summaries containing both cancer and tumor characteristics. A similar trend was observed for other papers including [44] and [54], where information related to multiple frames were extracted. Notably, some frames (particularly the ones which are not directly associated) contain a few common elements. For example, TUMOR SIZE element is present in both 'CANCER DIAGNOSIS' and 'TUMOR DESCRIPTION' frames as size is one of the characteristics that can be associated with both tumor and cancer diagnosis

(cancer being linked to tumor and tumor characteristics). Thus, assigning any paper to such a common frame element (e.g., TUMOR SIZE) was primarily based upon whether the paper dealt specifically with tumor description or cancer diagnosis. Table 3 shows a detailed overview of some of the noteworthy papers extracting information that could be represented using at least four different cancer frames.

Table 3: Papers extracting varied frames

Paper	Number of frames extracted	Frames extracted	Number of frame elements extracted
Cary et al. [52]	6	CANCER DIAGNOSIS, BLADDER CANCER DIAGNOSIS, TNM CLASSIFICATION, TUMOR DESCRIPTION, CANCER PROCEDURE, THERAPEUTIC PROCEDURE	11
Napolitano et al. [5]	6	PROSTATE CANCER DIAGNOSIS, SKIN CANCER DIAGNOSIS, GLEASON SCORE, GRADING VALUE, CLARK LEVEL, BRESLOW DEPTH	7
Coden et al. [54]	5	CANCER PROCEDURE, CANCER DIAGNOSIS, TUMOR DESCRIPTION, GRADING VALUE, MARGIN	28
Ping et al. [44]	5	CANCER PROCEDURE, THERAPEUTIC PROCEDURE, TUMOR DESCRIPTION,	19

Paper	Number of frames extracted	Frames extracted	Number of frame elements extracted
		CANCER DIAGNOSIS, COMORBIDITY DIAGNOSIS	
Savova et al. [25]	5	TUMOR DESCRIPTION, CANCER TREATMENT, CANCER DIAGNOSIS, BREAST CANCER DIAGNOSIS, TNM CLASSIFICATION	16
Tang et al. [50]	5	CANCER DIAGNOSIS, BREAST CANCER DIAGNOSIS, PATHOLOGY FINDING, KI-67 EXPRESSION, TUMOR DESCRIPTION	11
Breischneider et al. [62]	5	CANCER DIAGNOSIS, BREAST CANCER DIAGNOSIS, TNM CLASSIFICATION, CANCER PROCEDURE, THERAPEUTIC PROCEDURE	10
Xu et al. [73]	4	CANCER FINDING, PATHOLOGY FINDING, CANCER PROCEDURE, BREAST CANCER DIAGNOSIS	7
Thiebaut et al. [67]	4	CANCER DIAGNOSIS, BREAST CANCER DIAGNOSIS, KI-67 EXPRESSION, PATHOLOGY FINDING	7

Paper	Number of frames extracted	Frames extracted	Number of frame elements extracted
Schroeck et al. [63]	4	CANCER DIAGNOSIS, BLADDER CANCER DIAGNOSIS, TNM CLASSIFICATION, GRADING VALUE	6

Significant Cancer Types Based on Elements Extracted

Although a large proportion of the papers reviewed extracted general cancer information from EHRs using NLP, there are some papers which focused on extracting cancer-type specific information. The most notable cancer type was breast cancer, with the highest number of associated articles (*n=14*), followed by bladder cancer (*n=6*) and prostate cancer (*n=4*).

The frame elements for the three most commonly extracted cancer types based on the number of articles are summarized in Table 4. There may be more articles related to each of these cancer types, but the number here includes those that extract only type specific information (e.g., information specific to breast cancer instead of general cancer information extracted only from the notes of breast cancer patients).

Table 4: Most extracted cancer types

Cancer type	Important frame elements specific to this cancer	Number of associated papers
Breast Cancer	STATUS OF CANCER TYPES, RECEPTOR NAME (mainly Estrogen, progesterone, human epidermal growth factor 2), RECEPTOR STATUS, RECEPTOR PERCENT, RECEPTOR STATUS NEGATION, SCARFF-BLOOM-RICHARDSON STAGE, CONTRALATERAL EVENT	14

Cancer type	Important frame elements specific to this cancer	Number of associated papers
Bladder Cancer	INVASION STATUS, DEPTH OF INVASION, MUSCULARIS PROPRIA, CARCINOMA IN SITU	6
Prostate Cancer	GLEASON SCORE, PSA	4

We also note in Table 1 that the 'PROSTATE CANCER DIAGNOSIS' frame has been associated with 'PAIN IN PROSTATE CANCER PATIENTS' frame, all of whose elements were extracted from clinical text in a case of metastatic prostate cancer patients for identifying unknown pain phenotypes [76]. Napolitano et al. [5] extracted 'BRESLOW DEPTH' and 'CLARK LEVEL', both of which are grading values for skin cancer, as well as 'GLEASON SCORE' from pathology reports.

Frames Related to Cancer Quality Measures

A few papers focused on automatically identifying quality measures information for assessing the handling of cancer patients such as grouping patients for clinical trials and making decisions about their treatment plans. The frames created for these quality measures, however, are general-purpose (largely different types of patient assessments) and could be used for other applications as well. D'Avolio et al. [79] specifically worked on extracting such measures like MARGIN STATUS for patients who underwent prostatectomies. Another paper by Hernandez-Boussard et al. [87] extracted information about pre-treatment quality metrics. In context to assessing quality for cancer care, the frame 'ASSESSMENT FOR CANCER CARE' contains two elements, ASSESSMENT FOR SURGERY and PRE-TREATMENT CANCER ASSESSMENT, each of which further describe their individual elements (shown in Table 1). We observe in the table that besides GLEASON SCORE and surgical MARGIN STATUS which is already stated above, 'TNM CLASSIFICATION' is also identified by the College of American Pathologists (CAP) as one of the three Category I measures [79] for assessing quality.

Frames Related to Cancer Screening

Some articles applied NLP techniques to extract colonoscopy testing or colorectal cancer (CRC) screening related information such as test MENTION, TEMPORAL INFORMATION (e.g. 'colonoscopy in 2005'), STATUS (e.g. 'refused', 'scheduled') and NEGATION (e.g., negation

modifiers such as ‘no’ and ‘never’) of the tests [14][15][1]. Since the screening tests facilitate the cancer diagnosis process, we have represented the information using the ‘DIAGNOSTIC PROCEDURE’ frame (Table 1), which inherits from the ‘CANCER PROCEDURE’ frame (Figure 2).

Another study by Ritzwoller et al. [13] extracted general characteristics of lung nodules (e.g. nodule size) from radiology reports that the providers submitted to a Centers for Medicare and Medicaid Services (CMS)-approved registry following any Low-Dose CT Lung cancer screening procedure. As represented in Table 1, all nodule related information are captured in the ‘TUMOR DESCRIPTION’ frame.

Frames Related to Image Findings

While a majority of papers identified information from pathology findings (which usually contain detailed tumor-related information), a few studies attempted to automatically extract data elements from imaging findings (e.g. mammography reports) [7][90][98][24]. Interestingly, these findings were mostly related to breast cancer screening. For example, Lacson et al. [7] extracted elements such as CALCIFICATION, MASS, IMPLANTS, BIRADS CATEGORY, CYSTS, etc. with the aim to use these extracted elements for populating a breast cancer screening registry. He et al. [24] employed deep learning and NLP methods to extract similar information such as BREAST DENSITY, MASS, ARCHITECTURAL DISTORTION etc. from mammographic findings for identifying high risk patients and patients for whom biopsy is recommended. We have represented all the imaging related data elements in the ‘IMAGING FINDING’ frame and its subframe, ‘BREAST IMAGE FINDING’.

Other Important Frame Elements

Apart from the above mentioned frames and their respective elements, some of the other important data elements that multiple researchers (at least 3) were concerned about were: TEST RESULT in the ‘DIAGNOSTIC PROCEDURE’ frame [14][91][94][87]; TEMPORAL INFORMATION [15][44][1][10] and STATUS [15][44][1][91] in the ‘CANCER PROCEDURE’ frame; CALCIFICATION, MASS, ASYMMETRY, ARCHITECTURAL DISTORTION in the ‘IMAGING

FINDING' frame [7][98][24]; TYPE in the 'THERAPEUTIC PROCEDURE' frame [44][62][52]; and KI-67 SCORE in the 'KI-67 EXPRESSION' frame [6][78][50][67].

Two of the studies we reviewed extracted a patient's performance status information. One such performance measure is the Eastern Cooperative Oncology Group (ECOG) Scale of Performance Status, extracted by Herath et al. [57] and Najafabadipour et al. [42], whereas Karnofsky measure was extracted additionally by Najafabadipour et al. [42] from clinical narratives.

Although our main focus is not to examine the NLP methods used for extracting cancer information, we conducted a brief investigation of the general approaches taken in building the NLP systems. We found that 26 papers developed rule-based systems; 10 papers developed traditional machine learning (ML) models; 6 papers developed neural network-based ML ("deep learning") models; and 10 papers developed hybrid systems (combining rule- and ML-based methods). The rule-based approaches included adding various heuristic, linguistic, semantic, and symbolic rules. These mainly involved pattern-matching, developing grammar and tree-based parsing, and relying on dictionary-lookups. Among the traditional machine learning approaches, the most prevalent ones were conditional random field (CRF), support vector machine (SVM), boosting, and Naïve Bayes classifiers. The major deep learning-based frameworks used in the papers were hierarchical attention networks utilizing bi-directional recurrent neural network (RNN), convolutional neural network (CNN), and bi-directional long short-term memory RNN (Bi-LSTM) CRF models. 26 papers did not specify or elaborate the system architecture in detail, and 11 of these 26 papers used already available medical NLP systems such as GATE, Lingumatics I2E, KnowledgeMap Concept Identifier, Clearforest NLP Software, MedTAS/P (Medical Text Analysis System/Pathology), Medtex, IBM Watson for Oncology (WFO), and LifeCode (A-Life Medical, Inc.) system. A detailed overview of the methods used, evaluation process of the NLP systems as well as the metrics used for evaluation in the eligible papers are presented in Supplementary Table 2.

DISCUSSION

Our scoping review provides a detailed review of current research in the cancer information extraction domain from EHR text using NLP. 173 papers were identified as relevant to clinical

NLP for cancer. Of these, 78 were included in the frame structure (Table 1 and Figure 2). Appropriate interpretation of cancer information extracted as frames requires context, specifically tying the frames to the data source (EHR free text notes, in our case). We found that many papers use non-EHR sources for extracting cancer information (e.g., from biomedical literature), and these likely would have had entirely different frame structures (e.g., less patient-focused, more general knowledge-focused, more genomic information). For instance, a CANCER TREATMENT frame in EHR text could be assumed to discuss a specific attempted or potential treatment for that patient, and it would not be out of place to have a DATE element describing when the treatment was/will be performed, or a SITE element describing where the treatment was applied. In the literature, however, elements like SUCCESS RATE or 5-YEAR SURVIVAL—as aggregated over a cohort of a patients in a randomized trial—would be more likely. Even an element they may share in common, SITE, could have different implications (the anatomical location for the patient versus the organizational location of a trial).

Our review demonstrates evidence of the growing interest among NLP researchers in automatically extracting cancer related information from clinical text. The following points synthesize our key findings:

The first key finding is the redundancy of certain information types (e.g., 19 papers extracted cancer grade information). This suggests there has potentially been a tremendous replication of effort in these areas (Table 2), as state-of-the-art NLP systems require medium to large manually-annotated training sets to build high-performance machine learning models. Yet, as far as we can tell, the papers describe systems developed exclusively on local data and do not involve the sharing of data or models, forcing future researchers and clinicians in need of these cancer information types to "reinvent the wheel", replicating a time-consuming process requiring substantial NLP expertise. Pilot projects are underway to meet some of the needs of cancer researchers requiring such systems [25] [59], but the breadth of overlap (Table 1) suggests that without a more widely-available open resource, continued effort duplication will continue. Even more daunting are the cancer information needs that go unfulfilled because the costs of NLP projects are too high. This review demonstrates that there is a tremendous need for a more general-purpose cancer NLP resource.

The second key finding is that, given the fact that so few papers focus on more than a few information types, the decision on which cancer types to extract for a project is clearly an ad hoc process. The genesis of most of these papers was the need to support a given research project or clinical use case. This leads, for example, to cancer-specific diagnostic information being performed only for bladder, breast, colorectal, prostate, and skin cancer. Missing are common cancers such as lung cancer, lymphoma, kidney cancer, and leukemia. Each of these have information types specific to that form of cancer (e.g., certain blood tests for leukemia, or a blood marrow biopsy result). Given, again, the high effort and cost that can be associated with an NLP project, this collection of ad hoc projects—with both sizable overlap and yet key gaps—would be better served by a more general-purpose effort. This effort would identify widely-needed cancer information types and create an NLP system that supports a broad range of use cases. It would be impossible to cover all the use cases required by the 78 papers within this study, but there are a few specific, recurring information needs (Table 1).

Impact of Frame Semantics

This review organized cancer information according to the theory of frame semantics. As one can see from Figure 2, frames are an intuitive means of organizing information, not unlike entity-relationship and class diagrams from computer science. Frame semantics, however, is a fundamentally linguistic theory, as already described. This largely helps to minimize discrepancies between possible frame representations, though we make no claim the one presented here is the only or 'best' method for representing cancer information. There is certainly no shortage of ontologies that already exist to describe cancer information. But given the prominence of frames in general NLP methods, it is nonetheless beneficial to consider what a set of frames for cancer would look like when targeted toward NLP for EHR notes. Further, we did not find evidence, during the course of this review, of any emerging standard for organizing information extracted by NLP for cancer.

We consider this work, then, as the first step in an iterative process to derive an authoritative set of frames for cancer NLP targeted toward EHR notes. More investigation is certainly needed, and it would be critical to link the frames to existing ontologies such as the NCI Thesaurus and

SNOMED-CT where appropriate. Most importantly, frame structures need to stand up to the test of EHR text itself. Frame semantics asserts that the information (the elements) within a frame are often found together within some reasonable context (e.g., a cancer's GRADE and ANATOMICAL SITE can be found within the same sentence, if not the same clause). So for the proposed frames, this needs to be tested for frames to be used as a definitive NLP target representation.

Regardless of the utility in the proposed frames for actual NLP development, the organization of this information into frame structures is nonetheless useful. Table 1 contains over 160 elements, and any attempt to describe all of these information types without some intermediate structure (a frame that contains some set of elements) would be difficult to interpret. For instance, if a future researcher were interested in several specific information types related to cancer therapy procedures, finding the 'THERAPEUTIC PROCEDURE' frame will allow access to papers related to the information types organized under this frame (e.g., THERAPY DOSE and TOXICITIES information).

LIMITATIONS

This work has several limitations. The first limitation of this review is that, given the rapid pace of NLP development and publication, it is likely that papers meeting inclusion criteria were omitted. Specifically, papers published starting in late 2018 were likely missed. Second, although we have tried to accurately interpret and represent all the possible information types extracted in the literature as frames, there might be a few which are not captured or mis-represented in our final frame list. Finally, as already discussed, there may be inconsistencies in assigning the frames and associated elements across all the papers.

CONCLUSIONS

Our scoping review provides a detailed overview of the current research in the cancer information extraction domain from unstructured EHR notes using NLP. We conducted the review from a frame semantic perspective, described various frames along with their elements as well as examined the relations between frames. Since many researchers are trying to extract similar frames or frame elements (though not always using the language of frame semantics),

this review can help developers of a general-purpose cancer frame resource and NLP system that would extract a broad range of important cancer information types.

Acknowledgements

The authors would like to thank Funda Meric-Bernstam for reviewing a copy of the manuscript and suggesting clarifications from an oncology perspective.

Funding

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH), under award R00LM012104, and the Cancer Prevention and Research Institute of Texas (CPRIT), under award RP170668.

References

[1] J.C. Denny, J.F. Peterson, N.N. Choma, H. Xu, R.A. Miller, L. Bastarache, N.B. Peterson, Extracting timing and status descriptors for colonoscopy testing from electronic medical records, J. Am. Med. Inform. Assoc. 17 (2010) 383–388. doi:10.1136/jamia.2010.004804.

[2] A.K. AAlAbdulsalam, J.H. Garvin, A. Redd, M.E. Carter, C. Sweeny, S.M. Meystre, Automated Extraction and Classification of Cancer Stage Mentions fromUnstructured Text Fields in a Central Cancer Registry, AMIA Jt. Summits Transl. Sci. Proceedings.AMIA Jt. Summits Transl. Sci. 2017 (2018) 16–25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961766/.

[3] J.R. Gregg, M. Lang, L.L. Wang, M.J. Resnick, S.K. Jain, J.L. Warner, D.A. Barocas, Automating the Determination of Prostate Cancer Risk Strata From Electronic Medical Records, JCO Clin. Cancer Informatics. 2017 (2017) 1–8. doi:10.1200/CCI.16.00045.

[4] F.R. Schroeck, O. V Patterson, P.R. Alba, E.A. Pattison, J.D. Seigne, S.L. DuVall, D.J. Robertson, B. Sirovich, P.P. Goodney, Development of a Natural Language Processing Engine to Generate Bladder Cancer Pathology Data for Health Services Research, Urology. 110 (2017) 84–91. doi:S0090-4295(17)30966-4 [pii].

[5] 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. doi:10.1007/s10552-010-9616-4.

[6] D. Segagni, V. Tibollo, A. Dagliati, A. Zambelli, S.G. Priori, R. Bellazzi, An ICT infrastructure to integrate clinical and molecular data in oncology research, BMC Bioinformatics. 13 Suppl 4 (2012) S5. doi:10.1186/1471-2105-13-S4-S5.

[7] R. Lacson, K. Harris, P. Brawarsky, T.D. Tosteson, T. Onega, A.N.A. Tosteson, A. Kaye, I. Gonzalez, R. Birdwell, J.S. Haas, Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry, J. Digit. Imaging. 28 (2015) 567–575. doi:10.1007/s10278-014-9762-4.

[8] J. Cohen, A. Glaser, L. Okorji, D. Oberlin, J. Meeks, Creation of a quality-improvement database for transurethral resection of bladder tumors, J. Urol. 197 (2017) e115. doi:10.1016/j.juro.2017.02.346.

[9] E. Soysal, J.L. Warner, J.C. Denny, H. Xu, Identifying Metastases-related Information from Pathology Reports of Lung Cancer Patients, AMIA Jt. Summits Transl. Sci. Proceedings.AMIA Jt. Summits Transl. Sci. 2017 (2017) 268–277. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543353/.

[10] S. Zheng, S.K. Jabbour, S.E. O'Reilly, J.J. Lu, L. Dong, L. Ding, Y. Xiao, N. Yue, F. Wang, W. Zou, Automated Information Extraction on Treatment and Prognosis for Non–Small Cell Lung Cancer Radiotherapy Patients: Clinical Study, JMIR Med. Informatics. 6 (2018) e8. doi:10.2196/medinform.8662.

[11] 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. doi:10.1136/jamia.2009.002295.

[12] J. Lee, H.-J. Song, E. Yoon, S.-B. Park, S.-H. Park, J.-W. Seo, P. Park, J. Choi, Automated extraction of Biomarker information from pathology reports, BMC Med. Inform. Decis. Mak. 18 (2018) 29. doi:10.1186/s12911-018-0609-7.

[13] D.P. Ritzwoller, N.M. Carroll, A. Burnett-Hartman, H.S. Feigelson, E.E. Lyons, Lung

Cancer Screening And Nodule Evaluation: Performance Of Natural Language Processing In Identifying Lung Nodule Characteristics After Low-Dose CT Lung Cancer Screening, Am. J. Respir. Crit. Care Med. 193 (2016) 1. https://search.proquest.com/openview/c98409400dc08468521bb0a6c104b2f7/1.

[14] A. Kamineni, S. Halgrim, G. Gundersen, S. Fuller, G. Hart, D. Carrell, C. Rutter, PS2-26: Coordinating Heterogeneous Data and Mixed Collection Methods to Support Population-Based Cancer Screening Research, Clin. Med. Res. 11 (2013) 154. doi:10.3121/cmr.2013.1176.ps2-26.

[15] J.C. Denny, N.N. Choma, J.F. Peterson, R.A. Miller, L. Bastarache, M. Li, N.B. Peterson, Natural Language Processing Improves Identification of Colorectal Cancer Testing in the Electronic Medical Record, Med. Decis. Mak. 32 (2012) 188–197. doi:10.1177/0272989X11400418.

[16] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, L. Li, Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning, Sci. Rep. 6 (2016) 27327. doi:10.1038/srep27327.

[17] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J.A.W.M. van der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M.C. van Dijk, P. Bult, F. Beca, A.H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M.Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M.M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, R. Venâncio, Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer, JAMA. 318 (2017) 2199. doi:10.1001/jama.2017.14585.

[18] H. Li, D. Sheth, K.R. Mendel, L. Lan, M.L. Giger, Deep learning in computer-aided diagnosis incorporating mammographic characteristics of both tumor and parenchyma stroma, in: 2018: pp. 1071801–1071806. doi:10.1117/12.2318282.

[19] J. Jeong, Deep Learning for Cancer Screening in Medical Imaging, Hanyang Med. Rev. 37 (2017) 71–76. doi:10.7599/hmr.2017.37.2.71.

[20] D. Zhang, L. Zou, X. Zhou, F. He, Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer, IEEE Access. (2018). doi:10.1109/ACCESS.2018.2837654.

[21] J. Choi, I. Oh, S. Seo, J. Ahn, G2Vec: Distributed gene representations for identification of cancer prognostic genes, Sci. Rep. 8 (2018) 13729. doi:10.1038/s41598-018-32180-0.

[22] J. Liu, X. Wang, Y. Cheng, L. Zhang, Tumor gene expression data classification via sample expansion-based deep learning, Oncotarget. 8 (2017) 109646. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5752549/.

[23] J.X. Qiu, H.-J. Yoon, P.A. Fearn, G.D. Tourassi, Deep learning for automated extraction of primary sites from cancer pathology reports, IEEE J. Biomed. Heal. Informatics. 22 (2018) 244–251. doi:10.1109/JBHI.2017.2700722.

[24] T. He, M. Puppala, R. Ogunti, J.J. Mancuso, X. Yu, S. Chen, J.C. Chang, T.A. Patel, S.T.C. Wong, Deep learning analytics for diagnostic support of breast cancer disease

management, in: 2017: pp. 365–368. doi:10.1109/BHI.2017.7897281.
[25] G.K. Savova, E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris, H. Hochheiser, C. Lin, G. Chavan, R.S. Jacobson, DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records, Cancer Res. 77 (2017) e115–e118. doi:10.1158/0008-5472.CAN-17-0615.
[26] S. Gao, M.T. Young, J.X. Qiu, H.-J. Yoon, J.B. Christian, P.A. Fearn, G.D. Tourassi, A. Ramanthan, Hierarchical attention networks for information extraction from cancer pathology reports, J. Am. Med. Informatics Assoc. 25 (2017) 321–330. doi:10.1093/jamia/ocx131.
[27] G. Tourassi, Deep learning enabled national cancer surveillance, in: 2017: pp. 3982–3983. doi:10.1109/BigData.2017.8258411.
[28] C.F. Baker, FrameNet : A Knowledge Base for Natural Language Processing, Proc. OfFrame Semant. NLP A Work. Honor OfChuck Fill. (2014) 1–5. doi:10.1016/0093-691X(78)90114-0.
[29] K. Roberts, Y. Si, A. Gandhi, E. V. Bernstam, A FrameNet for Cancer Information in Clinical Narratives: Schema and Annotation, 11th Int. Conf. Lang. Resour. Eval. (2018). https://www.aclweb.org/anthology/L18-1041.
[30] G. Marzinotto, J. Auguste, F. Bechet, G. Damnati, A. Nasr, Semantic Frame Parsing for Information Extraction : the CALOR corpus, 11th Int. Conf. Lang. Resour. Eval. (2018). https://www.aclweb.org/anthology/L18-1159.
[31] W.W. Yim, M. Yetisgen, W.P. Harris, W.K. Sharon, Natural Language Processing in Oncology Review, JAMA Oncol. 2 (2016) 797–804. doi:10.1001/jamaoncol.2016.0213.
[32] I. Spasić, J. Livsey, J.A. Keane, G. Nenadić, Text mining of cancer-related information: Review of current status and future directions, Int. J. Med. Inform. 83 (2014) 605–623. doi:10.1016/j.ijmedinf.2014.06.009.
[33] X. Sun, X. Xu, J. Wang, J. Feng, S. Chen, Classifying Lung Cancer Knowledge in PubMed According to GO Terms Using Extreme Learning Machine, Int. J. Intell. Syst. 29 (2014) 1047–1059. doi:10.1002/int.21675.
[34] Z. Yin, W. Xie, B.A. Malin, Talking About My Care: Detecting Mentions of Hormonal Therapy Adherence Behavior in an Online Breast Cancer Community, in: 2017: p. 1868. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977653/.
[35] R. Thackeray, S.H. Burton, C. Giraud-Carrier, S. Rollins, C.R. Draper, Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month, BMC Cancer. 13 (2013) 508. doi:10.1186/1471-2407-13-508.
[36] V. Garla, C. Taylor, C. Brandt, Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management, J. Biomed. Inform. 46 (2013) 869–875. doi:10.1016/j.jbi.2013.06.014.
[37] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., Proceedings. AMIA Symp. (2001) 17–21. http://www.ncbi.nlm.nih.gov/pubmed/11825149 (accessed December 13, 2018).
[38] G.K. Savova, J.J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications., J. Am. Med. Inform. Assoc. 17 (2010) 507–13. doi:10.1136/jamia.2009.001560.
[39] J.D. Osborne, M. Wyatt, A.O. Westfall, J. Willig, S. Bethard, G. Gordon, Efficient identification of nationally mandated reportable cancer cases using natural language

processing and machine learning, J. Am. Med. Informatics Assoc. 23 (2016) 1077–1084. doi:10.1093/jamia/ocw006.
[40] H. Xu, Z. Fu, A. Shah, Y. Chen, N.B. Peterson, Q. Chen, S. Mani, M.A. Levy, Q. Dai, J.C. Denny, Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases, AMIA Annu. Symp. Proc. 2011 (2011) 1564. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243156/.
[41] F. Xie, J. Lee, C. Munoz-Plaza, E. Hahn, W. Chen, Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization, J. Pathol. Inform. 8 (2017) 48. doi:10.4103/jpi.jpi_55_17.
[42] M. Najafabadipour, J.M. Tuñas, A. Rodríguez-González, E. Menasalvas, Lung Cancer Concept Annotation from Spanish Clinical Narratives, ArXiv Prepr. ArXiv1809.06639. (2018). https://arxiv.org/abs/1809.06639.
[43] B. Karunakaran, D. Misra, K. Marshall, D. Mathrawala, S. Kethireddy, Closing the Loop - Finding Lung Cancer Patients using NLP, in: 2017: pp. 2452–2461. doi:10.1109/BigData.2017.8258203.
[44] X.-O. Ping, Y.-J. Tseng, Y. Chung, Y.-L. Wu, C.-W. Hsu, P.-M. Yang, G.-T. Huang, F. Lai, J.-D. Liang, Information Extraction for Tracking Liver Cancer Patients' Statuses: From Mixture of Clinical Narrative Report Types, Telemed. e-Health. 19 (2013) 704–710. doi:10.1089/tmj.2012.0241.
[45] A. Giri, R.T. Levinson, S. Keene, G. Holman, S.D. Smith, L. Clayton, W. Lovett, S.P. Stansel, M.-R.B. Snyder, J.T. Fromal, Preliminary results from the Pharmacogenetics Ovarian Cancer Knowledge to Individualize Treatment (POCKIT) study, (2018). doi:10.1158/1538-7445.AM2018-4229.
[46] E. Pattison, D. Denhalter, O. Patterson, S. DuVall, J. Seigne, B. Sirovich, P. Goodney, D. Robertson, F. Schroeck, Leveraging bladder cancer pathology reports for research: Gleaning meaning despite widely variable language, J. Urol. 195 (2016) e425. doi:10.1016/j.juro.2016.02.1267.
[47] R.K. Taira, R. Taira, A. Bui, A.A. Bui, W. Hsu, V. Bashyam, S. Dube, E. Watt, L. Andrada, S. El-Saden, T. Cloughesy, H. Kangarloo, A tool for improving the longitudinal imaging characterization for neuro-oncology cases, AMIA Annu. Symp. Proc. (2008) 712. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656085/.
[48] J.M. Buckley, S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A.K. Belli, E.M.H. Kim, J.E. Garber, B.L. Smith, M.A. Gadd, M.C. Specht, C.A. Roche, T.M. Gudewicz, K.S. Hughes, The feasibility of using natural language processing to extract clinical information from breast pathology reports, J. Pathol. Inform. 3 (2012) 23. doi:10.4103/2153-3539.97788.
[49] F. Acevedo, V.D. Armengol, Z. Deng, R. Tang, S.B. Coopey, D. Braun, A. Yala, R. Barzilay, C. Li, A. Colwell, A. Guidi, C.L. Cetrulo, J. Garber, B.L. Smith, T. King, K.S. Hughes, Pathologic findings in reduction mammoplasty specimens: a surrogate for the population prevalence of breast cancer and high-risk lesions, Breast Cancer Res. Treat. 173 (2019) 201–207. doi:10.1007/s10549-018-4962-0.
[50] R. Tang, L. Ouyang, C. Li, Y. He, M. Griffin, A. Taghian, B. Smith, A. Yala, R. Barzilay, K. Hughes, Machine learning to parse breast pathology reports in Chinese, Breast Cancer Res. Treat. 169 (2018) 243–250. doi:10.1007/s10549-018-4668-3.
[51] F. Acevedo, R. Tang, S. Coopey, A. Yala, R. Barzilay, C. Li, A. Colwell, A. Guidi, C. Cetrulo, J.E. Garber, Pathologic findings in reduction mammoplasty procedures identified

by natural language processing of breast pathology reports: A surrogate for the population incidence of cancer and high risk lesions, (2018). doi:10.1200/JCO.2018.36.15_suppl.e13569.

[52] C. Cary, A. Roberts, A.K. Church, G. Eckert, F. Ouyang, J. He, D.A. Haggstrom, Development of a novel algorithm to identify staging and lines of therapy for bladder cancer, (2017). doi:10.1200/JCO.2017.35.15_suppl.e18235.

[53] Z. Zeng, X. Li, S. Espino, A. Roy, K. Kitsch, S. Clare, S. Khan, Y. Luo, Contralateral Breast Cancer Event Detection Using Nature Language Processing, in: 2017: p. 1885. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977664/.

[54] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P.C. de Groen, Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, J. Biomed. Inform. 42 (2009) 937–949. doi:10.1016/j.jbi.2008.12.005.

[55] A. Nguyen, M. Lawley, D. Hansen, S. Colquist, Structured pathology reporting for cancer from free text: Lung cancer case study, Electron. J. Heal. Informatics. 7 (2012). doi:10.1109/TELSKS.2001.955803.

[56] A.N. Nguyen, J. Moore, J. O'Dwyer, S. Philpot, Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository, AMIA ...Annual Symp. Proceedings.AMIA Symp. 2016 (2017) 964–973. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333242/.

[57] D.H. Herath, D. Wilson-Ing, E. Ramos, G. Morstyn, Assessing the natural language processing capabilities of IBM Watson for oncology using real Australian lung cancer cases., J. Clin. Oncol. (2016). doi:10.1200/JCO.2016.34.15_suppl.e18229.

[58] H.-J. Yoon, S. Robinson, J.B. Christian, J.X. Qiu, G.D. Tourassi, Filter pruning of Convolutional Neural Networks for text classification: A case study of cancer pathology report comprehension, in: 2018: pp. 345–348. doi:10.1109/BHI.2018.8333439.

[59] Y. Si, K. Roberts, A Frame-Based NLP System for Cancer-Related Information Extraction The University of Texas Health Science Center at Houston, AMIA Annu. Symp. Proc. (2018). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371330/.

[60] F. Schroeck, O. Patterson, P. Alba, S. DuVall, B. Sirovich, D. Robertson, J. Seigne, P. Goodney, Harnessing full text pathology data from the electronic health record to advance bladder cancer care – Development of a natural language processing system to generate longitudinal pathology data, J. Urol. 197 (2017) e413. doi:10.1016/j.juro.2017.02.987.

[61] A. Burnett-Hartman, P.A. Newcomb, C.X. Zeng, Y. Zheng, J.M. Inadomi, C. Fong, M.P. Upton, W.M. Grady, Using medical informatics to evaluate the risk of colorectal cancer in patients with clinically diagnosed sessile serrated polyps, (2017). doi:10.1158/1538-7445.CRC16-PR05.

[62] C. Breischneider, S. Zillner, M. Hammon, P. Gass, D. Sonntag, Automatic extraction of breast cancer information from clinical reports, in: 2017: pp. 213–218. doi:10.1109/CBMS.2017.138.

[63] F.R. Schroeck, K.E. Lynch, J. won Chang, T.A. MacKenzie, J.D. Seigne, D.J. Robertson, P.P. Goodney, B. Sirovich, Extent of Risk-Aligned Surveillance for Cancer Recurrence Among Patients With Early-Stage Bladder Cancer, JAMA Netw. Open. 1 (2018) e183442–e183442. doi:10.1001/jamanetworkopen.2018.3442.

[64] R. Weegar, H. Dalianis, Creating a rule-based system for text mining of Norwegian breast

cancer pathology reports, Sixth Int. Work. Heal. Text Min. Inf. Anal. (2015) 73–78. https://www.aclweb.org/anthology/W15-2609.

[65] Z. Zexian, R. Ankita, L. Xiaoyu, E. Sasa, C. Susan, K. Seema, L. Yuan, Using Clinical Narratives and Structured Data to Identify Distant Recurrences in Breast Cancer, in: 2018: pp. 44–52. doi:10.1109/ICHI.2018.00013.

[66] A. Yala, R. Barzilay, L. Salama, M. Griffin, G. Sollender, A. Bardia, C. Lehman, J.M. Buckley, S.B. Coopey, F. Polubriaginof, J.E. Garber, B.L. Smith, M.A. Gadd, M.C. Specht, T.M. Gudewicz, A.J. Guidi, A. Taghian, K.S. Hughes, Using machine learning to parse breast pathology reports, Breast Cancer Res. Treat. 161 (2017) 203–211. doi:10.1007/s10549-016-4035-1.

[67] N. Thiebaut, A. Simoulin, K. Neuberger, I. Ibnouhsein, N. Bousquet, N. Reix, S. Molière, C. Mathelin, An innovative solution for breast cancer textual big data analysis, ArXiv Prepr. ArXiv1712.02259. (2017). http://arxiv.org/abs/1712.02259.

[68] N. Viani, L. Chiudinelli, C. Tasca, A. Zambelli, M. Bucalo, A. Ghirardi, N. Barbarini, E. Sfreddo, L. Sacchi, C. Tondini, R. Bellazzi, Automatic Processing of Anatomic Pathology Reports in the Italian Language to Enhance the Reuse of Clinical Data, Stud. Health Technol. Inform. 247 (2018) 715–719. doi:10.3233/978-1-61499-852-5-715.

[69] F. Schroeck, K. Lynch, J.W. Chang, D. Robertson, J. Seigne, P. Goodney, B. Sirovich, A national study of risk-aligned surveillance practice for non-muscle invasive bladder cancer, J. Urol. 199 (2018) e587. doi:10.1016/j.juro.2018.02.1420.

[70] 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. doi:10.1136/amiajnl-2012-000928.

[71] A.P. Nunes, E. Green, T. Dalvi, J. Lewis, N. Jones, J.D. Seeger, A real-world evidence study to define the prevalence of endocrine therapy-naïve hormone receptor-positive locally advanced or metastatic breast cancer in the US, (2017). doi:10.1158/1538-7445.SABCS16-P5-08-20.

[72] J.L. Warner, M.A. Levy, M.N. Neuss, ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data, J. Oncol. Pract. 12 (2016) 157–158. doi:10.1200/JOP.2015.004622.

[73] 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. doi:10.3233/978-1-60750-949-3-565.

[74] M.K. Breitenstein, H. Liu, K.N. Maxwell, J. Pathak, R. Zhang, Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats From Treatment of Breast Cancer at a Single Institution, Clin. Transl. Sci. 11 (2018) 85–92. doi:10.1111/cts.12514.

[75] K.M. Christopherson, X. Lei, C.H. Barcenas, T.A. Buchholz, K. Hoffman, H.M. Kuerer, S.F. Shaitelman, G.H. Perkins, G.L. Smith, M.C. Stauder, (S003) Curative-Intent Treatment for Newly Diagnosed Breast Cancer With Limited Metastatic Disease to the Sternum or Mediastinum, Int. J. Radiat. Oncol. Biol. Phys. 98 (2017) E1–E2. doi:10.1016/j.ijrobp.2017.02.039.

[76] N.H. Heintzelman, R.J. Taylor, L. Simonsen, R. Lustig, D. Anderko, J.A. Haythornthwaite, L.C. Childs, G.S. Bova, 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. doi:10.1136/amiajnl-2012-001076.

[77] S. Mehrabi, A. Krishnan, A.M. Roch, H. Schmidt, D. Li, J. Kesterson, C. Beesley, P. Dexter, M. Schmidt, M. Palakal, H. Liu, Identification of Patients with Family History of Pancreatic Cancer--Investigation of an NLP System Portability, Stud. Health Technol. Inform. 216 (2015) 604–608. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863760/.

[78] D. Segagni, V. Tibollo, A. Dagliati, L. Perinati, A. Zambelli, S. Priori, R. Bellazzi, The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology, Stud. Health Technol. Inform. 169 (2011) 887. doi:10.3233/978-1-60750-806-9-887.

[79] L.W. D'Avolio, M.S. Litwin, S.O. Rogers, A.A.T. Bui, JAMIA Facilitating Clinical Outcomes Assessment through the Automated Identification of Quality Measures for Prostate Cancer Surgery, J. Am. Med. Informatics Assoc. 15 (2008). doi:10.1197/jamia.M2649.

[80] A.N. Nguyen, M.J. Lawley, D.P. Hansen, R. V Bowman, B.E. Clarke, E.E. Duhig, S. Colquist, Symbolic rule-based classification of lung cancer stages from free-text pathology reports, J. Am. Med. Informatics Assoc. 17 (2010) 440–445. doi:10.1136/jamia.2010.003707.

[81] W.-W. Yim, T. Denman, S.W. Kwan, M. Yetisgen, Tumor information extraction in radiology reports for hepatocellular carcinoma patients., AMIA Jt. Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci. (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001784/.

[82] W. Yim, S.W. Kwan, M. Yetisgen, Classifying tumor event attributes in radiology reports, J. Assoc. Inf. Sci. Technol. 68 (2017) 2662–2674. doi:10.1002/asi.23937.

[83] A.P. Glaser, B.J. Jordan, J. Cohen, A. Desai, P. Silberman, J.J. Meeks, Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing, JCO Clin. Cancer Informatics. 2 (2018) 1–8. doi:10.1200/CCI.17.00128.

[84] W. Yim, S.W. Kwan, M. Yetisgen, Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction, J. Biomed. Inform. 64 (2016) 179–191. doi:10.1016/j.jbi.2016.10.005.

[85] L.T.E. Cheng, J. Zheng, G.K. Savova, B.J. Erickson, Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing, J. Digit. Imaging. 23 (2010) 119–132. doi:10.1007/s10278-009-9215-7.

[86] A. Halwani, K.M. Rasmussen, V. Patil, Z. Burningham, S. Narayanan, S.-W. Lin, S. Carroll, L.-I. Hsu, J.N. Graff, R. Dreicer, S. Gupta, C. Low, B.C. Sauer, Racial disparities in metastatic castrate-resistant prostate cancer (mCRPC): Evidence from the Veterans Health Administration (VHA), in: Cancer Res., 2018: pp. A055–A055. doi:10.1158/1538-7445.prca2017-a055.

[87] T. Hernandez-Boussard, P. Kourdis, R. Dulal, M. Ferrari, S. Henry, T. Seto, K. McDonald, D.W. Blayney, J.D. Brooks, A natural language processing algorithm to measure quality prostate cancer care, (2017). doi:10.1200/JCO.2017.35.8_suppl.232.

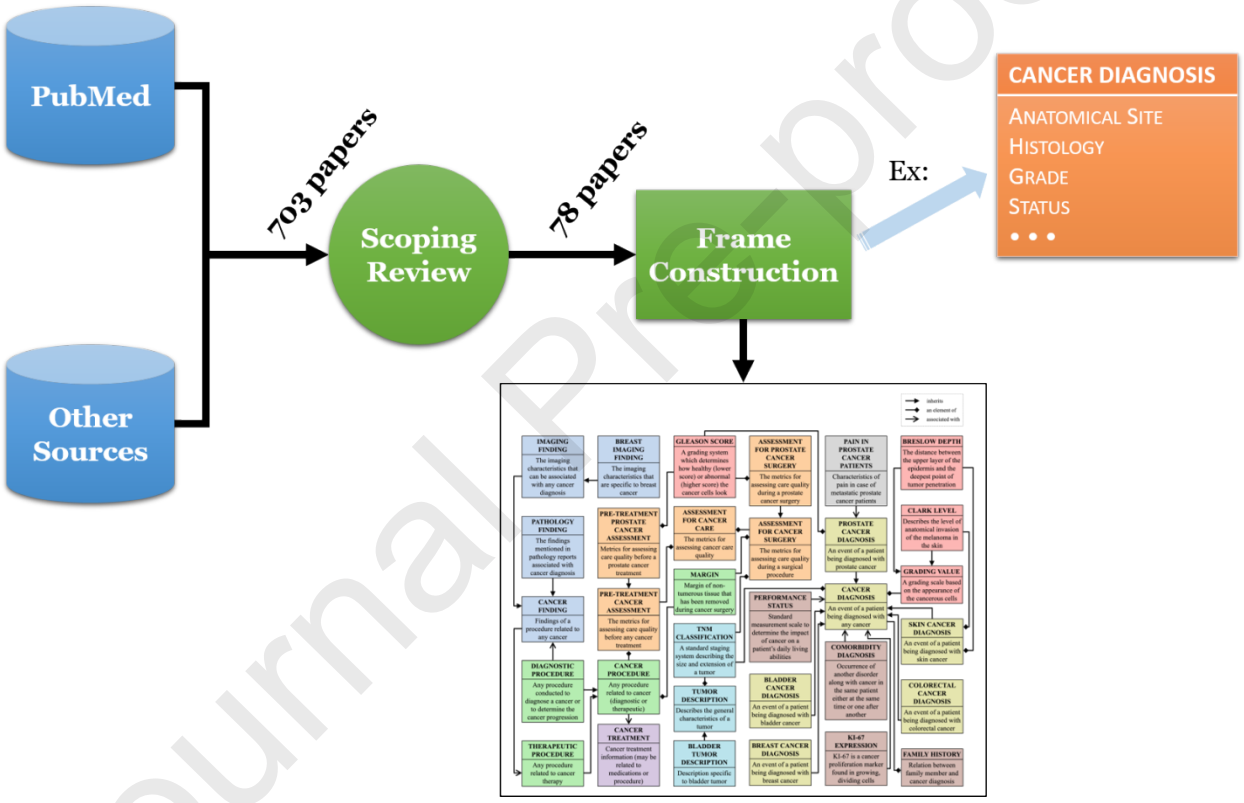
[88] A. Wieneke, E. Bowles, D. Cronkite, K. Wernli, H. Gao, D. Carrell, D. Buist, Validation of natural language processing to extract breast cancer pathology procedures and results, J. Pathol. Inform. 6 (2015) 38. doi:10.4103/2153-3539.159215.

[89] B.W. Mamlin, D.T. Heinze, C.J. McDonald, Automated extraction and normalization of findings from cancer-related free-text radiology reports, AMIA Annu. Symp. Proc. (2003)

420. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479955/.
[90] C.R. Moore, A. Farrag, E. Ashkin, Using Natural Language Processing to Extract Abnormal Results From Cancer Screening Reports, J. Patient Saf. 13 (2017) 138–143. doi:10.1097/PTS.0000000000000127.
[91] B.H.L. Goulart, E. Silgard, C.S. Baik, A. Bansal, M. Greenwood-Hickman, A. Hanson, S.D. Ramsey, S. Schwartz, Validation of natural language processing (NLP) for automated ascertainment of EGFR and ALK tests in SEER cases of non-small cell lung cancer (NSCLC), (2017). doi:10.1200/JCO.2017.35.15_suppl.6528.
[92] B. Goulart, E. Silgard, C. Baik, A. Bansal, M. Greenwood-Hickman, A. Hanson, S. Ramsey, S. Schwartz, P3. 07-013 Determining EGFR and ALK Status in a Population-Based Cancer Registry: A Natural Language Processing Validation Study: Topic: Other – Geographical Differences, J. Thorac. Oncol. 12 (2017) S1438. doi:10.1016/j.jtho.2016.11.2204.
[93] H.-J. Tan, R. Clarke, K. Chamie, A.L. Kaplan, A.I. Chin, M.S. Litwin, C.S. Saigal, A.S. Hackbarth, Development and Validation of an Automated Method to Identify Patients Undergoing Radical Cystectomy for Bladder Cancer Using Natural Language Processing, Urol. Pract. 4 (2017) 365–372. doi:10.1016/j.urpr.2016.09.011.
[94] M. Ananda-Rajah, C. Bergmeir, F. Petitjean, M.A. Slavin, K.A. Thursky, G.I. Webb, Toward Electronic Surveillance of Invasive Mold Diseases in Hematology-Oncology Patients: An Expert System Combining Natural Language Processing of Chest Computed Tomography Reports, Microbiology, and Antifungal Drug Data, JCO Clin. Cancer Informatics. 1 (2017) 1–10. doi:10.1200/CCI.17.00011.
[95] T. Hernandez-Boussard, P.D. Kourdis, T. Seto, M. Ferrari, D.W. Blayney, D. Rubin, J.D. Brooks, Mining Electronic Health Records to Extract Patient-Centered Outcomes Following Prostate Cancer Treatment, AMIA ...Annual Symp. Proceedings.AMIA Symp. 2017 (2018) 876–882. https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/29854154/.
[96] A.W. Forsyth, R. Barzilay, K.S. Hughes, D. Lui, K.A. Lorenz, A. Enzinger, J.A. Tulsky, C. Lindvall, Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records, J. Pain Symptom Manage. 55 (2018) 1492–1499. doi:S0885-3924(18)30082-4 [pii].
[97] B.U. Wu, J.W. Chung, W. Yu, D.L. Conwell, D. Yadav, S.J. Pandol, Mo1250-Risk of Pancreatic Cancer in Patients with Newly Diagnosed Chronic Pancreatitis, Gastroenterology. 154 (2018) S-720. doi:10.1016/S0016-5085(18)32520-4.
[98] M. Puppala, T.C. He, R. Ogunti, S.T.C. Wong, Use of natural language processing on mammography and pathology findings to supplement BI-RADS to improve clinical decision making in breast cancer care, (2017). doi:10.1158/1538-7445.SABCS16-P5-03-08.
[99] Y. Pershad, S. Govindan, A.K. Hara, M.J. Borad, T. Bekaii-Saab, A. Wallace, H. Albadawi, R. Oklu, Using Naive Bayesian Analysis to Determine Imaging Characteristics of KRAS Mutations in Metastatic Colon cancer, Diagnostics (Basel, Switzerland). 7 (2017) 10.3390/diagnostics7030050. doi:10.3390/diagnostics7030050.

Highlights

- We reviewed studies extracting cancer information from EHR notes using NLP method
- We organized information following frame semantics to identify areas of overlap, gaps
- We constructed frames and frame relations resulting in 32 frames, 163 frame elements
- This review is useful to develop general-purpose cancer frame resource and NLP system



Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: