A Novel Corpus of Annotated Medical Imaging Reports and Information Extraction Results Using BERT-based Language Models

Namu Park[1*], Kevin Lybarger[2*], Giridhar Kaushik Ramachandran[2], Spencer Lewis[3], Aashka Damani[4], Özlem Uzuner[2], Martin Gunn[5], Meliha Yetisgen[1]

[1]Department of Biomedical Informatics & Medical Education, University of Washington
[2]Department of Information Sciences and Technology, George Mason University
[3]Department of Radiology, School of Medicine, Stanford University
[4]School of Medicine, University of Washington
[5]Department of Radiology, School of Medicine, University of Washington
[1,4,5]Seattle, WA, USA, [2]Fairfax, VA, USA, [3]Stanford, CA, USA
{npark95, aashkad, marting, melihay}@uw.edu
{klybarge, gramacha, ouzuner}@gmu.edu
lewispen@stanford.edu
[*]Authors contributed equally to this paper.

Abstract

Medical imaging is critical to the diagnosis, surveillance, and treatment of many health conditions, including oncological, neurological, cardiovascular, and musculoskeletal disorders, among others. Radiologists interpret these complex, unstructured images and articulate their assessments through narrative reports that remain largely unstructured. This unstructured narrative must be converted into a structured semantic representation to facilitate secondary applications such as retrospective analyses or clinical decision support. Here, we introduce the Corpus of Annotated Medical Imaging Reports (CAMIR), which includes 609 annotated radiology reports from three imaging modality types: Computed Tomography, Magnetic Resonance Imaging, and Positron Emission Tomography-Computed Tomography. Reports were annotated using an event-based schema that captures clinical indications, lesions, and medical problems. Each event consists of a trigger and multiple arguments, and a majority of the argument types, including anatomy, normalize the spans to pre-defined concepts to facilitate secondary use. CAMIR uniquely combines a granular event structure and concept normalization. To extract CAMIR events, we explored two BERT (Bi-directional Encoder Representation from Transformers)-based architectures, including an existing architecture (mSpERT) that jointly extracts all event information and a multi-step approach (PL-Marker++) that we augmented for the CAMIR schema.

Keywords: Natural Language Processing, Radiology, Information Extraction, Corpus, Clinical Informatics

1. Introduction

Radiology reports document radiologists' interpretation of medical images through detailed narrative text. Although some studies have explored structured reports that utilize common data elements to express radiologists' interpretations through predefined medical concepts (Rubin and Kahn Jr, 2017), the majority of radiology reports utilize narrative text (Willemink et al., 2020). Information extraction (IE) techniques can automatically convert unstructured reports to structured semantic representations to allow utilization of the textual information in secondary-use applications. Example applications include cohort discovery (Casey et al., 2021), epidemiology (Casey et al., 2021), image retrieval (Gerstmair et al., 2012), automated follow-up tracking (Mabotuwana et al., 2019), computer-vision applications (Zech et al., 2018), decision support (Demner-Fushman et al., 2009), and report summarization (Wiggins et al., 2021).

Although there is a well-established body of radiology IE research, most of this research focuses on specific clinical tasks (Casey et al., 2021; Donnelly et al., 2022) or medical conditions, utilizes a single imaging modality, or implements an annotation schema that does not comprehensively capture the available information. To address these limitations, we introduce a novel annotated corpus, the *Corpus of Annotated Medical Imaging Reports* (CAMIR), that is relevant to a broad set of applications. CAMIR includes Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography-Computed Tomography (PET-CT) reports. The reports are annotated using a granular event schema, where clinical indication, lesion, and medical problem findings are characterized through multiple arguments, including assertion values (present vs. absent), normalized anatomy using a hierarchical ontology of 87 SNOMED-CT concepts, and other clinically important attributes. CAMIR includes

LREC-COLING 2024, pages 1280–1292

609 annotated radiology reports with 1,494 indication events, 5,709 lesion events, and 6,255 medical problem events. CAMIR has a high inter-annotator agreement (>0.805 F1) for event triggers and an overall inter-annotator agreement of 0.762 F1. We present initial IE results using two BERT (Bi-directional Encoder Representation from Transformers)-based language models trained and evaluated on CAMIR, including the Multi-label Span-based Entity and Relation Transformer (mSpERT) (Eberts and Ulges, 2020; Lybarger et al., 2023) and an augmented version of Packed-levitated Markers (Ye et al., 2022) (referred to as PL-Marker++). Both architectures achieve performance comparable to the inter-annotator agreement (IAA), with PL-Marker++ achieving the highest overall performance.

2. Related Work

There is a significant body of research that explores IE within the radiology domain, including the creation of annotated corpora and the development of extraction models (Pons et al., 2016; Casey et al., 2021; López-Úbeda et al., 2022). In this section, we discuss existing research in clinical NLP focusing on radiology corpora and relevant IE techniques.

2.1. Radiology Corpora

Radiology reports present nuanced and complex descriptions of medical findings, which existing annotated corpora capture with varying degrees of granularity. Document-level or sentence-level annotations map relevant text to normalized values, targeting diverse label categories such as metastases characteristics (Do et al., 2021) and incidental findings (Trivedi et al., 2019). Entity annotations identify phrases of interest, capturing concepts like anatomical location (Wang et al., 2019) or tumor attributes (Yim et al., 2016). Relation and event annotations enable more nuanced representations, like the multi-attribute characterization of medical problems (Lau et al., 2022). Selected studies have integrated the normalization of radiological concepts related to anatomy (Lybarger et al., 2022; Datta and Roberts, 2022; Nishigaki et al., 2023) and other radiology terminology (Datta et al., 2020a).

Existing corpora often exhibit limitations in various dimensions, such as the diversity of the patient populations represented, the range of imaging modalities included, the scale of the datasets, or the granularity and comprehensiveness of the annotation schemas employed. Some studies concentrate on specific diseases or conditions like hepatocellular carcinoma (Yim et al., 2016) or appendicitis (Rink et al., 2013), limiting the represented patient populations. Others are limited to single imaging modalities (Lau et al., 2022; Sugimoto et al., 2021) or small corpora (n<200) (Hassanpour and Langlotz, 2016). Other relevant relation extraction work does not include the normalization of extracted spans to key concepts (Jain et al., 2021). More recent work (Lybarger et al., 2022) extracts findings with the associated normalized anatomy values; however, the findings are not fully characterized through granular attributes.

To our knowledge, CAMIR is the first annotated corpus to uniquely combine clinical concept normalization with granular event annotations to comprehensively capture important clinical findings. Additionally, CAMIR includes a diverse set of reports from three imaging modalities that were sampled from all patients at the University of Washington (UW). CAMIR's fine-grained annotation schema with concept normalization and heterogeneous set of reports can support a wide range of secondary-use applications.

2.2. IE Methods in Radiology

Early radiology IE research employed discrete machine learning models with engineered features. For instance, Support Vector Machines were used to detect appendicitis findings (Rink et al., 2013) and Conditional Random Fields were utilized to extract anatomy and findings (Hassanpour and Langlotz, 2016). These discrete modeling approaches were supplanted by neural network architectures, such as Convolutional Neural Network and Recurrent Neural Networks. These neural architectures outperform their predecessors in many radiology IE tasks, including but not limited to recommendation extraction (Carrodeguas et al., 2019; Steinkamp et al., 2021), clinical concept identification (Zhu et al., 2018), and spatial information extraction (Datta et al., 2020b). Currently, pre-trained Language Models dominate the IE landscape in radiology, similar to other domains. BERT (Devlin et al., 2019) models have been extensively implemented for tasks ranging from observation detection (Irvin et al., 2019) to anatomy classification (Nishigaki et al., 2023) and relation-based finding extraction (Lybarger et al., 2022). Most recently, Generative Pre-trained Transformers (GPT) models are being leveraged to extract structured information from radiology reports (Fink et al., 2023; Mukherjee et al., 2023; Adams et al., 2023). In this paper, we present the extraction results of two high-performing BERT-based models, which were tailored to reflect the granularity of our annotation schema and serve as a foundation upon which future work can build.

3. Methods

3.1. Corpus Creation

We used an existing clinical database of 1,417,586 CT, 541,388 MRI, and 39,150 PET-CT reports from 2007-2020 which includes the general patient population from four UW Medical System hospitals. We randomly sampled reports from each modality: 203 CT, 202 MRI, and 204 PET-CT. The reports were automatically de-identified using a neural de-identifier (Lee et al., 2021). The study was approved by the UW Institutional Review Board (IRB).

3.1.1. Annotation Schema

In CAMIR's event schema, each event includes a trigger that identifies the event and arguments that characterize the event. Table 1 summarizes the schema, and Figure 1 presents annotation examples from the BRAT rapid annotation tool (Stenetorp et al., 2012), which was used throughout the annotation process. CAMIR includes three event types: (1) *Indication* - reason for the imaging (e.g., "cancer" in line 1 of Figure 1), (2) *Lesion* – mass-occupying pathological structures (e.g., "metastasis" in line 3 of Figure 1); and (3) *Medical Problem* - non-mass-like abnormalities, defined as a finding that is not a potential mass (e.g., "scarring" in line 1 of Figure 1). There are two argument types: (1) *span-only* arguments assign text spans an argument label (e.g., "focal" assigned *Characteristic* argument in line 2 of Figure 1) and (2) *span-with-value* arguments assign text spans both an argument label and argument subtype label (e.g., "New" assigned *Size Trend* argument with subtype value *new* in line 2 of Figure 1). To improve the granularity of our annotation schema, anatomy arguments are normalized to a set of hierarchical anatomical SNOMED-CT concepts, including 16 *Anatomy Parent* and 71 *Anatomy Child* labels listed in Table 2 (e.g., "Bilateral apical lung" assigned *Anatomy Parent - Respiratory* and *Anatomy Child - Lung* in line 1 of Figure 1).

3.1.2. Annotation Process

Four medical students annotated CAMIR. A senior radiology resident and an experienced board-certified attending radiologist provided domain expertise in creating the annotation guidelines and resolving the ambiguities during annotation. The annotation guidelines were designed with the efforts of a medical resident and a board-certified radiologist with 20+ years of experience and profound knowledge of clinical NLP. After a series of meetings, we updated the annotation guidelines multiple times to ensure the guidelines accurately and comprehensively capture the indication, finding, and lesion information relevant to a wide range of clinical research, including our current exploration of cancer and incidental findings. The hierarchical anatomy normalization schema was developed with the help of a board-certified radiologist by reflecting the widely used SNOMED-CT concepts. Two pairs of two medical students doubly annotated 357 reports, and 252 reports were single-annotated by the same annotators. Annotators reached a consistent level of IAA after 5 rounds of double annotation. Disagreements were adjudicated with the help of domain experts who created and revised the annotation guidelines when needed. We then transitioned to single annotation for the next four rounds to expedite the annotation process. CAMIR includes training, validation, and test set splits (70%:10%:20%). The training set is 41% doubly annotated, and the entire validation and test sets are doubly annotated to ensure evaluation reliability.

We singly annotated the training set to create a larger and more diverse training set, while providing the most robust data set for the validation and test sets using double-annotation. The consistency of annotations between singly annotated and doubly annotated reports was evaluated by analyzing the average frequency of labels per report. The doubly annotated reports have an average of 2.65±0.48 Indication, 10.15±1.31 Medical Problem, and 9.77±0.99 Lesion triggers per report, and the singly annotated reports include an average of 2.14±0.26 Indication, 9.91±2.58 Medical Problem, and 8.78±1.06 Lesion triggers per report. The frequency of triggers is slightly lower in the singly annotated reports, suggesting there is some reduced annotation recall for the singly annotated reports; however, the evaluation was performed on the doubly annotated test set, and any annotation noise associated with the singly training examples is captured by this evaluation.

3.2. Extraction Architectures

To extract the CAMIR events, we explored two state-of-the-art BERT (Devlin et al., 2019)-based Language Models: (1) mSpERT (Eberts and Ulges, 2020; Lybarger et al., 2023) and (2) an augmented version of PL-Marker (Ye et al., 2022) referred to as PL-Marker++. For both systems, we decomposed events into a set of entities and relations, where the relation head is a trigger and the relation tail is an argument.

3.2.1. mSpERT

SpERT (Eberts and Ulges, 2020) jointly extracts entities and relations using BERT (Devlin et al.,

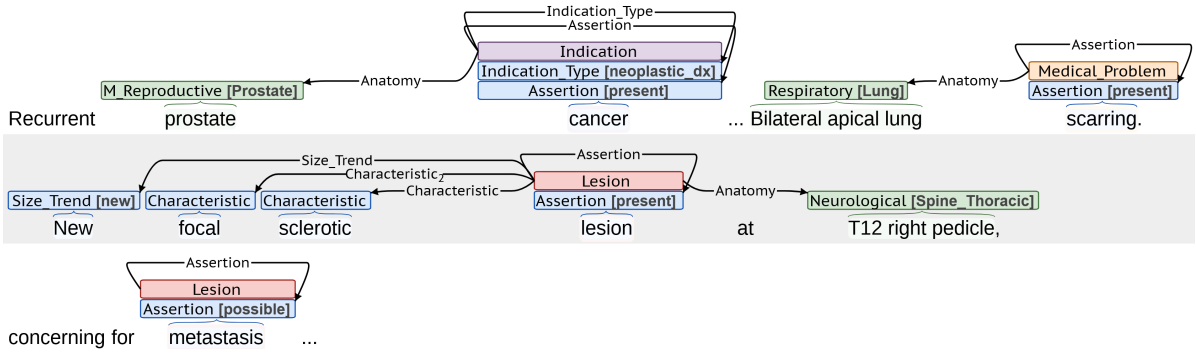

Figure 1: Examples of sentences annotated with event schema

Event	Trigger/ Argument	Argument subtypes	Span examples
Indication	Trigger*	–	"hemorrhage," "sepsis"
	Type*	{trauma, symptom, neoplastic diagnosis, non-neoplastic diagnosis}	"seminoma," "sarcoid"
	Assertion*	{present, absent, possible}	"r/o," "concern"
	Anatomy	Anatomy Parent and Child labels+	"abdominal," "alveolar"
Lesion	Trigger*	–	"'lymhpadenopathy"
	Assertion*	{present, absent, possible}	"most likely,"
	Anatomy	Anatomy Parent and Child labels+	"lower back"
	Size	{current, past}	"up to 5mm"
	Size Trend	{new, disappear, increasing, decreasing, no-change}	"increasing in size"
	Count	–	"multiple," "numerous"
	Characteristic	–	"peripheral," "enlarged"
Medical Problem	Trigger*	–	"dilation," "calcification"
	Assertion*	{present, absent, possible}	"possibly"
	Anatomy	Anatomy Parent and Child labels+	"mucosal," "supraagger"

Table 1: Summary of the event schema. * indicates the argument is required. + *Anatomy Parent* and *Anatomy Child* are list in Table 2. "Dx" refers to diagnosis.

2019) with output layers that classify spans and predict relations. mSpERT (Lybarger et al., 2023) includes additional output layers to allow multi-label span predictions, which we use to predict subtype labels. Figure 2 shows the mSpERT architecture, which includes Entity Type, Entity Subtype, and Relation output layers. The Entity Type and Relation layers of mSpERT are identical to the original SpERT implementation, and the Entity Subtype layer allows multi-label span predictions. The Entity Type classifier (ϕ_e) is a linear layer that operates on the sentence representation (e_{CLS}), max-pooled span hidden states ($e(s_i)$), and learned span width embeddings (w_{k+1}). The Entity Subtype classifiers (ϕ_S) are separate linear layers for each *span-with-value* argument that operate on the same input as the Entity Type classifier but also incorporate the Entity Type logits. The Relation classifier (ψ_r) predicts links between entity spans using a linear layer that operates on max-pooled spans ($e(s_i)$ and ($e(s_l)$), span width embeddings (w_{k+1}), and max-pooled hidden states between the entity spans ($c(s_i, s_j)$). The mSpERT predictions can generate the CAMIR event structure.

3.2.2. PL-Marker++

PL-Marker (Ye et al., 2022) is a multi-stage extraction framework, where the first stage identifies entities and second stage resolves relations. To extract CAMIR events, we introduced an augmented version of PL-Marker, referred to as PL-Marker++, which includes a third classification stage for the *span-with-value* subtype labels. Figure 3 presents

Anatomy Parent	Anatomy Children
Abdomen	Abdominal Wall, Adrenal Gland, Mesentery, Peritoneal Sac, Retroperitoneal, & Spleen
Body Regions	Entire Body, Lower Limb, Pelvis, & Upper Limb
Cardiovascular	Arterial, Coronary Artery, Heart, Pericardial Sac, Pulmonary Artery, & Venous
Digestive	Esophagus, Intestine, Large Intestine, Small Intestine, & Stomach
Female Reproductive & Obstetric	Adnexal, Breast, Extra-embryonic, Female Genital Structure, Fetus, Ovary, Placenta, Umbilical Cord, & Uterus
Head & Neck	Ear, Eye, Laryngeal, Mouth, Nasal Sinus, Neck, Pharynx, & Thyroid
Hepato-Biliary	Bile Duct, Gallblader, Liver, & Pancreas
Lymphatic	–
Male Reproductive	Epididymis, Prostate, & Testis
Musculo-Skeletal	Bone/Joint, & Skeletal and/or Smooth Muscle
Neurological	Brain, Cerebrospinal Fluid Pathway, Cerebrovascular System, Extraaxial, Nerve, Pituitary, & Spine - Cervical, Cord, Lumbar, Sacral, Thoracic, or Unspecified
Respiratory	Lung, Pleural Membrane, & Tracheobronchial
Skin	Skin and or Mucous Membrane, & Subcutaneous
Thoracic	Mediastinal
Urinary	Kidney, Ureter, & Urinary Bladder
Miscellaneous	Adipose Tissue, Biomedical Device, & Connective Tissue

Table 2: Anatomy Parent-Child Hierarchy. All 16 Parent and 71 Child labels map to SNOMED-CT concepts, but label names are shortened for space. All Parent labels include an *Undetermined* child label.

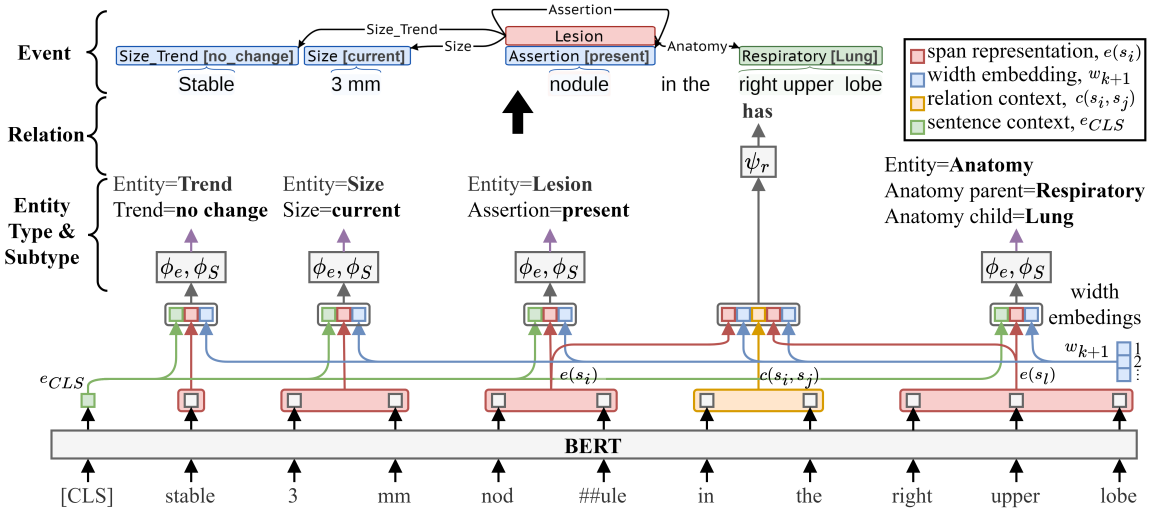

Figure 2: mSpERT framework

the PL-Marker++ architecture, where the Entity Type and Relation stages are identical to the original PL-Marker model. The Entity Type stage uses a group packing approach to process many spans concurrently while considering their interdependencies. The Relation stage uses a subject-oriented packing strategy to pack each relation head and all associated relation tails into an instance, allowing the dependencies between span pairs to be modeled. The Entity Subtype classification generates a new input sentence for each extracted entity, where typed markers identify the target entity. This entity-specific version of the sentence feeds into BERT, and the CLS token hidden state feeds into a multi-label classifier consisting of separate linear layers for each *span-with-value* argument.

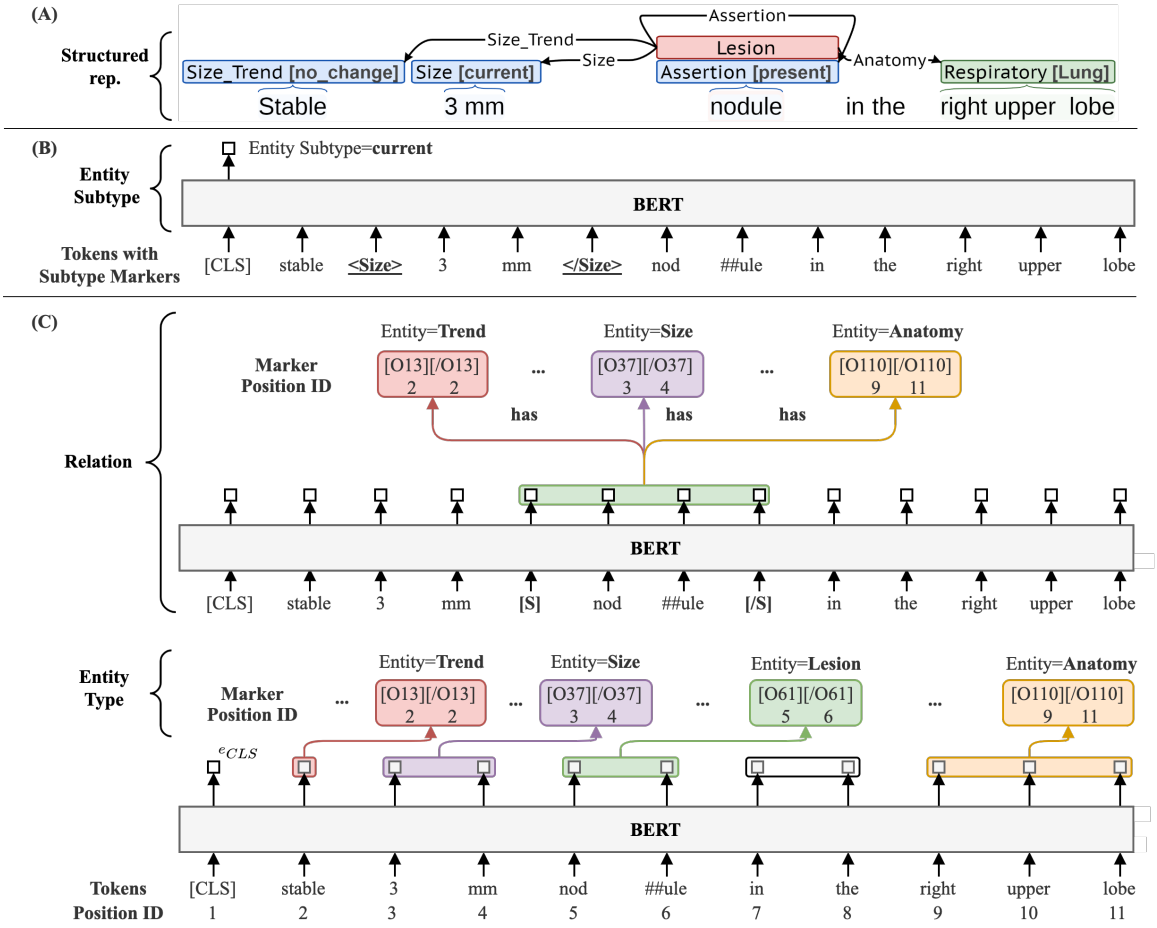

Figure 3: PL-Marker++ framework

3.3. Evaluation

Model hyperparameters were tuned using the CAMIR training and validation sets, and the final performance is reported for the withheld CAMIR test set. Performance is presented using the overlap span equivalence criterion, where two spans are considered equivalent if they overlap. For instance, when extracting anatomy spans in line 2 of Figure 1, the anatomy span, "right pedicle" would be considered equivalent to "T12 right pedicle" since there is an overlap between the spans. Triggers are considered equivalent if the event types are identical and spans overlap. *Span-only* arguments are considered equivalent if the argument types match, argument spans overlap, and connected triggers are equivalent. *Span-with-value* argument equivalence is similar to the equivalence of *span-only* arguments, except that the subtype labels must also match. Overlap span equivalence is relevant to the CAMIR annotation schema and extraction task as most arguments are normalized to predefined concepts. This overlap criterion is also suited for downstream secondary-use applications, and we performed extensive error analyses to validate this criterion (see Section 5.3). Performance is evaluated using precision, recall, and F1, and statistical significance is calculated using a non-parametric (bootstrap) test (Berg-Kirkpatrick et al., 2012).

4. Results

4.1. Corpus

This section summarizes CAMIR, including the IAA and distribution of annotations. Table 3 presents IAA for the doubly annotated reports. The overall IAA for all the triggers and arguments in the doubly-annotated reports was 0.762 F1 using the criteria defined in Section 3.3. Consensus regarding the trigger annotation was higher at 0.856, 0.805, and 0.854 F1 for *Indication*, *Lesion*, and *Medical Problem* triggers, respectively. *Size*, *Size Trend*, and *Count* occur much less frequently than the other arguments, contributing to the lower IAA for these arguments. *Characteristic* spans are very linguistically diverse, resulting in frequent false negatives. The double annotation and adjudication of the validation and test sets

mitigates the impact of this lower IAA on the evaluation.

Event type	Argument type	F1
Indication	Trigger	0.856
	Assertion	0.820
	Anatomy	0.797
	Indication Type	0.804
Lesion	Trigger	0.805
	Assertion	0.762
	Anatomy	0.710
	Size	0.715
	Size Trend	0.560
	Count	0.564
	Characteristic	0.481
Medical Problem	Trigger	0.854
	Assertion	0.815
	Anatomy	0.751
Overall		0.762

Table 3: Inter-annotator agreement (IAA) for doubly annotated radiology reports (n=357)

Table 4 summarizes the distribution of the annotated phenomena in CAMIR. While the focus of the imaging modality may differ, the distribution of annotations is similar across modalities for most argument types. In each report, 2.4-2.5 *Indication* triggers were identified and the reason for the imaging test was mostly *neoplastic diagnosis*, which refers to the abnormal growth of certain tumors. The number of *Lesion* and *Medical Problem* was similar in all three modality types, where most triggers for both events were assigned *Assertion* value *present*. Approximately 9.2-9.7 *Lesion* and 10.2-10.4 *Medical Problem* events were identified on average in each radiology report.

Lesion-specific attributes such as *Characteristic*, *Size*, *Size Trend*, *Count* add supplementary clinical information that might be crucial for interpreting the result of the imaging tests. In addition, we provide *Assertion* values to each event to clearly indicate the absence, possibility, or presence of each finding. These *Assertion* labels are very important for creating accurate and comprehensive representations and are relevant to wide range of secondary use cases. The granularity of CAMIR also provides the opportunity for more advanced multi-modal research by combining text and relevant images.

4.2. Information Extraction

Table 5 summarizes the extraction performance on the held-out CAMIR test set. PL-Marker++ achieved significantly higher overall performance than mSpERT (0.759 F1 vs 0.736 F1). While the performance of mSpERT and PL-Marker++ models were similar for extracting *Indication* and *Medical Problem* triggers and arguments, PL-Marker++ performed significantly better in extracting *Lesion* triggers and all but one argument type. The PL-Marker++ model achieved gains of $+\Delta 0.05$ F1 in extracting *Characteristic*, *Size*, and *Size Trend* arguments for *Lesion* events. The overall improved performance of PL-Marker++ can be attributed to the infusion of the trigger and argument location information through all layers of the BERT model.

5. Discussion

5.1. Annotation Quality

The IAA for CAMIR exceeds 0.70 F1 for most arguments. Exceptions are *Size Trend*, *Count*, and *Characteristic*. We observed that *Size Trend* and *Count* are relatively infrequent in our data set and are therefore easy to overlook during annotation. *Characteristic* was introduced as an inclusive catchall category, resulting in diverse lexical phrasing and semantics, consequently yielding a comparatively low IAA.

The IAA for *Lesion* and *Medical Problem* triggers was above 0.80 F1. Majority of the remaining disagreements resulted from ambiguity between event types. For example, generic words such as “disease” can refer to a *Lesion* trigger in the context of “residual disease,” indicating a small number of cancer cells. At the same time, “disease” can refer to a *Medical Problem* in the context of “small vessel disease”. Similarly, “recurrence” can be either *Lesion* or *Medical Problem* depending on the finding that is recurring.

5.2. Model Performance

Table 5 shows the BERT models achieved the highest performance for the extraction of triggers and some of the more regularly-expressed arguments such as *Count*. *Anatomy* is a crucial argument for capturing the meaning of the radiology reports and has an extraction performance of 0.628-0.718 F1, indicating further study is needed to improve extraction performance.

5.3. Strict vs Overlap Evaluation

To validate the span overlap criterion, we evaluated the performance of PL-Marker++ on event triggers using a strict, exact match span criterion. This evaluation resulted in test set performance of 0.749 F1 for *Indication*, 0.681 F1 for *Lesion*, and 0.765 F1 for *Medical Problem* triggers. There were 279 triggers that were equivalent using the overlap criterion but not equivalent using exact match. We manually reviewed these trigger predictions to assess their clinical meaning relative to the reference

Event Type	Argument Type	Argument Subtype	Frequency (avg. per report)		
			CT	MR	PET-CT
Indication	Trigger	-	507 (2.5)	496 (2.4)	491 (2.4)
	Assertion	present	449	435	436
		absent	11	1	5
		possible	47	60	50
	Type	neoplastic dx	184	181	193
		non-neoplastic dx	112	102	91
		symptom	149	150	134
		trauma	23	32	21
	Anatomy	all	276	263	278
Lesion	Trigger	-	1855 (9.2)	1967 (9.7)	1887 (9.3)
	Assertion	present	1190	1302	1222
		absent	547	531	539
		possible	118	134	126
	Anatomy	all	2321	2536	2378
	Size	current	303	364	349
		past	46	63	36
	Size Trend	decreasing	26	38	36
		disappear	22	18	26
		increasing	35	61	32
		new	64	58	46
		no change	109	142	130
	Count	-	119	112	132
	Characteristic	-	762	841	921
Medical Problem	Trigger	-	2063 (10.2)	2111 (10.4)	2080 (10.2)
	Assertion	present	1217	1294	1189
		absent	607	592	631
		possible	239	225	260
	Anatomy	all	2197	2316	2083
		Total number of reports (N)	203	202	204

Table 4: Distribution of the annotated event types and arguments in CAMIR by modality. Numbers in parentheses indicate the average number of triggers per report.

triggers. For all 279 of these discrepancies between the overlap and strict criterion, the predicted triggers still captured all information important to identifying clinical findings. We found that 203 of these trigger predictions were shorter than the reference, often omitting modifiers (e.g. reference - "Mild FDG activity" vs. predicted - "FDG activity" or reference - "hypodense lesions" vs. predicted - "lesions") and 76 trigger predictions were longer than the reference, often including modifiers (e.g. reference - "lesion" vs. predicted "mass lesion" or reference "carcinoma" vs. predicted "renal cell carcinoma").

5.4. Generalizability of the Annotation and Extraction Performance

Our annotation guidelines are designed to be comprehensive and foundational to derive overall clinical findings from radiology reports. The guidelines do not rely on specific templates or formats used in our institution. Even though the structure of the medical imaging reports may differ across modalities or institutions, we expect the description of the main clinical findings in the reports to be compatible with our annotation guidelines. Moreover, although our annotation focused on three imaging modalities, the annotation schema is not specific to particular modalities. Therefore, we anticipate that minimal modifications will be required to the annotation schema to create annotated datasets at different institutions or for other modality types, if any. However, since the content and linguistics may vary among institutions and modality types, directly using information extraction models trained on CAMIR may achieve lower performance on the reports at other institutions or for other modalities. Domain adaptation of the CAMIR-trained models may be required, to maintain high performance.

Event	Argument	Count	mSpERT			PL-Marker++		
			P	R	F1	P	R	F1
Indication	Trigger	285	0.818	0.758	**0.787**	0.878	0.705	0.782
	Assertion	285	0.816	0.730	**0.770**	0.852	0.684	0.759
	Anatomy Parent	157	0.696	0.554	0.617	0.711	0.580	**0.639**
	Anatomy Child	157	0.675	0.529	0.593	0.711	0.580	**0.639**
	Type	262	0.783	0.687	**0.732**	0.782	0.683	0.729
Lesion	Trigger	1169	0.859	0.846	0.853	0.880	0.888	**0.884**†
	Assertion	1169	0.840	0.810	0.825	0.863	0.870	**0.866**†
	Anatomy Parent	1448	0.753	0.620	0.680	0.769	0.673	**0.718**†
	Anatomy Child	1448	0.720	0.586	0.646	0.733	0.642	**0.684**†
	Characteristic	652	0.654	0.420	0.512	0.776	0.477	**0.591**†
	Count	75	0.833	0.800	**0.816**	0.902	0.733	0.809
	Size	294	0.761	0.670	0.713	0.890	0.691	**0.778**†
	Size Trend	206	0.720	0.587	0.647	0.795	0.714	**0.752**†
Medical Problem	Trigger	1271	0.897	0.832	0.863	0.886	0.866	**0.875**
	Assertion	1271	0.878	0.802	0.839	0.854	0.834	**0.844**
	Anatomy Parent	1349	0.792	0.623	**0.697**	0.752	0.633	0.688
	Anatomy Child	1349	0.725	0.563	**0.633**	0.687	0.578	0.628
OVERALL		12847	0.798	0.684	0.736	0.805	0.718	**0.759**†

Table 5: Event extraction performance for mSpERT and PL-Marker++ evaluated using overlap criteria on the held-out test set. Higher F1-scores are bolded. † indicates statistical significance ($p < 0.05$)

6. Conclusion

We introduce a novel annotated corpus, CAMIR, consisting of CT, MRI, and PET-CT reports from a large hospital system. CAMIR has been annotated using a granular event schema, where clinical indication, lesion, and medical problem findings are captured through multiple arguments and most arguments are normalized to predefined radiological concepts. Using CAMIR, we explored two BERT-based architectures (1) mSpERT, an existing system which jointly extracts all event information, and (2) PL-Marker++, a system that extracts the event information through multiple stages, which we augmented before applying to CAMIR. These systems performed comparable to IAA. Our PL-Marker++ achieved significantly higher overall performance than mSpERT (0.759 F1 vs 0.736 F1). These systems show that the fine-grained information in CAMIR can be reliably extracted by automatic methods. While these systems perform well overall, triggers and their assertion arguments are more reliably extracted than other arguments such as anatomy. The annotation guidelines for CAMIR and the source code for the IE models presented in this paper are available on our GitHub repository*. CAMIR is unique in that it combines clinical concept normalization with the granularity of relation/event annotations to produce comprehensive semantic representations that can easily be incorporated into secondary-use applications, including clinical decision support (Demner-Fushman et al., 2009), surveillance (Haas et al., 2005), follow-up tracking (Mabotuwana et al., 2019), report simplification (Qenam et al., 2017), cross-specialty diagnosis correlation (Filice, 2019), and automated impression generation (Wiggins et al., 2021).

*https://github.com/uw-bionlp/CAMIR

7. Limitations

This study is limited to data from a single urban hospital system and focuses on three imaging modalities. While CAMIR includes more than 13,000 clinical events, it only consists of 609 reports. Therefore, the generalizability of the annotated corpus and extraction architectures to other hospital systems and other imaging modalities needs further exploration. In future work, we will incorporate additional modalities such as radiographs, ultrasound, and mammography. Additionally, we will evaluate the performance of larger generative Large Language Models (e.g. GPT4) in fine-tuning and in-context learning settings.

8. Acknowledgements

This work was supported by the National Institutes of Health (NIH) - National Cancer Institute (Grant Nr. 1R01CA248422-01A1) and National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (Grant Nr. T15LM007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

9. Ethics

We obtained the necessary approvals from our institution's IRB, with a waiver of patient consent to use their clinical notes. Radiology reports may contain patient Protected Health Information (PHI), like names, contact information, and other identifiers. Each report was automatically de-identified using a neural de-identification model and then subsequently manually de-identified by medical student annotators, to ensure no remaining PHI. All radiology reports, including the original and de-identified versions, were stored on a Health Insurance Portability and Accountability Act (HIPAA)-compliant server, to ensure patient privacy. All researchers and annotators received the necessary human subjects training to interact with patient data, including PHI.

The annotated reports in our corpus were randomly sampled from the general population of patients with medical imaging from a single institution. The demographics of the patients were not considered during data collection, and the patient populations in our corpus may not be representative of populations at other institutions or the broader population, which may inadvertently bias the distribution of annotated medical conditions. Additionally, radiology reports of other institutions may differ in format and language. These factors may impact the generalizability of the extraction models developed using the corpus.

Bibliography

Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. 2023. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Emmanuel Carrodeguas, Ronilda Lacson, Whitney Swanson, and Ramin Khorasani. 2019. Use of machine learning to identify follow-up recommendations in radiology reports. *Journal of the American College of Radiology*, 16(3):336–343.

Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu, and Beatrice Alex. 2021. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21(1):179.

Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts. 2020a. Radlex normalization in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2020, page 338. American Medical Informatics Association.

Surabhi Datta and Kirk Roberts. 2022. Fine-grained spatial information extraction in radiology as two-turn question answering. *International Journal of Medical Informatics*, 158:104628.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. 2020b. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of Biomedical Informatics*, 108:103473.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Richard K. G. Do, Kaelan Lupton, Pamela I. Causa Andrieu, Anisha Luthra, Michio Taya, Karen Batch, Huy Nguyen, Prachi Rahurkar, Lior Gazit, Kevin Nicholas, Christopher J. Fong, Natalie Gangai, Nikolaus Schultz, Farhana Zulkernine, Varadan Sevilimedu, Krishna Juluru, Amber Simpson, and Hedvig Hricak. 2021. Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT radiology reports over a 10-year period. *Radiology*, 301(1):115–122.

Lane F. Donnelly, Robert Grzeszczuk, and Carolina V. Guimaraes. 2022. Use of natural language processing (NLP) in evaluation of radiology reports: An update on applications and technology advances. *Seminars in Ultrasound, CT and MRI*, 43(2):176–181.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with

transformer pre-training. In *European Conference on Artifical Intelligence*, pages 2006–2013.

Ross W Filice. 2019. Deep-learning language-modeling approach for automated, personalized, and iterative radiology-pathology correlation. *J Am Coll Radiol*, 16(9):1286–1291.

Matthias A Fink, Arved Bischoff, Christoph A Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F Weber. 2023. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*, 308(3):e231362.

Axel Gerstmair, Philipp Daumke, Kai Simon, Mathias Langer, and Elmar Kotter. 2012. Intelligent image retrieval based on radiology reports. *European Radiology*, 22(12):2750–2758.

Janet P Haas, Eneida A Mendonça, Barbara Ross, Carol Friedman, and Elaine Larson. 2005. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *Am J Infect Control*, 33(8):439–443.

Saeed Hassanpour and Curtis P. Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66:29–39.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting clinical entities and relations from radiology reports. In *Neural Information Processing Systems*.

Wilson Lau, Kevin Lybarger, Martin L Gunn, and Meliha Yetisgen. 2022. Event-based clinical finding extraction from radiology reports with pre-trained language model. *Journal of Digital Imaging*, pages 1–14.

Kahyun Lee, Nicholas J Dobbins, Bridget McInnes, Meliha Yetisgen, and Özlem Uzuner. 2021. Transferability of neural network clinical deidentification systems. *Journal of the American Medical Informatics Association*, 28(12):2661–2669.

Pilar López-Úbeda, Teodoro Martín-Noguerol, Krishna Juluru, and Antonio Luna. 2022. Natural language processing in radiology: update on clinical applications. *Journal of the American College of Radiology*.

Kevin Lybarger, Aashka Damani, Martin Gunn, Özlem Uzuner, and Meliha Yetisgen. 2022. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. In *AMIA Informatics Summit*.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Ozuner, and Meliha Yetisgen. 2023. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*.

Thusitha Mabotuwana, Christopher S Hall, Vadiraj Hombal, et al. 2019. Automated tracking of follow-up imaging recommendations. *American Journal of Roentgenology*, 212(6):1287–1294.

Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. 2023. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*, 309(1):e231147.

Daiki Nishigaki, Yuki Suzuki, Tomohiro Wataya, Kosuke Kita, Kazuki Yamagata, Junya Sato, Shoji Kido, and Noriyuki Tomiyama. 2023. Bert-based transfer learning in sentence-level anatomic classification of free-text radiology reports. *Radiology: Artificial Intelligence*, 5(2):e220097.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: Translation and evaluation. *J Med Internet Res*, 19(12):e417.

Bryan Rink, Kirk Roberts, Sanda Harabagiu, Richard H Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. 2013. Extracting actionable findings of appendicitis from radiology reports using natural language processing. *AMIA Summits on Translational Science*, 2013:221.

Daniel L Rubin and Charles E Kahn Jr. 2017. Common data elements in radiology. *Radiology*, 283(3):837–844.

Jackson Steinkamp, Charles Chambers, Darco Lalevic, and Tessa Cook. 2021. Automatic fully-contextualized recommendation extraction from radiology reports. *Journal of Digital Imaging*, 34:374–384.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, et al. 2012. BRAT: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata, Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, and Yasushi Matsumura. 2021. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.

Gaurav Trivedi, Esmaeel R. Dadashzadeh, Robert M. Handzel, Wendy W. Chapman, Shyam Visweswaran, and Harry Hochheiser. 2019. Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports. *Applied Clinical Informatics*, 10(4):655–669.

Yanshan Wang, Saeed Mehrabi, Sunghwan Sohn, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. 2019. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Medical Informatics and Decision Making*, 19:23–29.

Walter F Wiggins, Felipe Kitamura, Igor Santos, and Luciano M Prevedello. 2021. Natural language processing of radiology text reports: Interactive text classification. *Radiology: Artificial Intelligence*, page e210035.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, et al. 2020. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4904–4917, Dublin, Ireland.

Wen-wai Yim, Tyler Denman, Sharon W Kwan, and Meliha Yetisgen. 2016. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science*, 2016:455.

John Zech, Margaret Pain, Joseph Titano, et al. 2018. Natural language–based machine learning models for the annotation of clinical radiology reports. *Radiology*, 287(2):570–580.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*.

Appendix A. SNOMED-CT Concepts for Anatomy Normalization

Anatomy Parent	Anatomy Children	Count
Abdomen (113345001)	Abdominal Wall (83908009), Adrenal Gland (23451007), Mesentery (89679009), Peritoneal Sac (118762006), Retroperitoneal (699600004), Spleen (78961009), Undetermined	512
Cardiovascular System (59820001)	Arterial (51114001), Coronary Artery (41801008), Heart (80891009), Pericardial Sac (76848001), Pulmonary Artery (81040000), Venous (119553000), Undetermined	770
Digestive System (49596003)	Esophagus (32849002), Intestine (113276009), Large Intestine (14742008), Small Intestine (30315005), Stomach (69695003), Undetermined	425
Female Reproductive System (27436002) & Obstetric (308762002)	Adnexal (23043003), Breast (76752008), Extra-embryonic (314908006), Female Genital Structure (53065001), Fetus (55460000), Ovary (15497006), Placenta (78067005), Umbilical Cord (29870000), Uterus (35039007), Undetermined	272
Head & Neck (774007)	Ear (117590005), Eye (371398005), Laryngeal (4596009), Mouth (385294005), Nasal Sinus (2095001), Neck (45048000), Pharynx (54066008), Thyroid (69748006), Undetermined	1096
Hepato-Biliary System (34707002, 122489005)	Bile Duct (28273000), Gallbladder (28231008), Liver (10200004), Pancreas (15776009), Undetermined	609
Lymphatic (91688001)	Undetermined	559
Male Reproductive System (90264002)	Epididymis (87644002), Prostate (119231001), Testis (40689003), Undetermined	49
Miscellaneous	Adipose Tissue (55603005), Biomedical Device (63653004), Connective Tissue (21793004), Undetermined	59
Musculoskeletal (312717002)	Bone/Joint, Skeletal and or Muscle (71616004), Undetermined	1811
Neurological System (25087005)	Brain (12738006), Cerebrospinal Fluid Pathway (280371009), Cerebrovascular System (28661005), Extraaxial (1231004), Nerve (3057000), Pituitary (56329008), Spine Cervical (122494005), Spine Cord (2748008), Spine Lumbar (122496007), Spine Sacral (699698002), Spine Thoracic (122495006), Spine Unspecified (421060004), Undetermined	3235
Other Body Regions (272625005)	Entire Body (38266002), Lower Limb (61685007), Pelvis (12921003), Upper Limb (53120007), Undetermined	887
Respiratory System (714323000)	Lung (39607008), Pleural Membrane (3120008), Tracheobronchial (91724006), Undetermined	1200
Skin (400199006)	Skin and or Mucous Membrane (707861009), Subcutaneous (71966008), Undetermined	58
Thoracic (51185008)	Mediastinum (72410000), Undetermined	772
Urinary System (122489005)	Kidney (64033007), Ureter (119220009), Urinary Bladder (89837001), Undetermined	378

Table A1: Anatomy Parent-Child SNOMED Hierarchy. SNOMED concept names are shortened due to lack of space. There are 16 Parent and 71 Child labels. Undetermined Child labels are catch-all categories. Count represents the number of annotations for Parent labels.