Journal of Imaging Informatics in Medicine

Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset

--Manuscript Draft--

Manuscript Number:	JDIM-D-26-00614R1
Full Title:	Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset
Article Type:	Original Paper
Section/Category:	Editorial
Funding Information:	National Center for Advancing Translational Sciences (2UL1TR001425-05A1) — Dr Elanchezhian Somasundaram
Abstract:	Purpose: General-purpose clinical NLP tools are increasingly used to automatically label clinical reports for research and quality improvement. However, independent evaluations for specific tasks like labeling pediatric CXR reports remain limited. This study compares four commercial clinical NLP models: AWS, AZ, GC, SP, for entity extraction and assertion detection of clinically relevant findings in pediatric CXR reports. Two dedicated CXR report labelers, CheXpert and CheXbert, were also evaluated for comparison. Materials and Methods: A total of 95,008 pediatric CXR reports from a large academic hospital were analyzed. Entities and their assertion statuses (positive, negative, uncertain) were extracted from findings and impression sections using the four general-purpose models. Entities from impressions were mapped to 12 disease categories plus No Findings using regular expressions. CheXpert and CheXbert processed same reports for the same 13 labels. Outputs from all six systems were compared across assertion categories using Fleiss' Kappa to quantify inter-model agreement. A manual review was done by a board-certified pediatric radiologist on a subset of 360 exams stratified across the 12 CheXpert labels and across the 3 assertion statuses (positive, negative, and uncertain). Precision, recall and F1-scores were calculated for each NLP service against the ground truth. Results: Significant differences were observed in the mean number of extracted entities among the NLP models ($p < 0.001$). SP extracted the most unique entities (49,688), followed by AZ (31,543), AWS (27,216), and GC (16,477). Assertion distributions also differed significantly ($p < 0.001$). For the CheXpert labels comparison, the mean Fleiss' Kappa across all assertion categories, was 0.68 ± 0.17 including all reports and 0.35 ± 0.14, excluding reports where all six models predicted the disease as absent. For manual validation on the subset, CheXbert and CheXpert achieved the highest macro F1-scores (0.565 and 0.561), followed by Google (0.453), Azure (0.421), Spark NLP (0.420), and AWS (0.266). Overall, positive assertions were best detected (F1 0.782), uncertain the worst (0.468). Performance varied by label, with highest F1 for pleural effusion, pneumothorax, and pneumonia, and lowest for lung opacity, enlarged cardiomediastinum, and lung lesion. Conclusions: Marked variability exists among NLP models in entity extraction and uncertainty handling, underscoring the need for further validation before clinical deployment.
Corresponding Author:	Shruti Hegde, MS Cincinnati Children's Hospital Medical Center UNITED STATES OF AMERICA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Cincinnati Children's Hospital Medical Center
Corresponding Author's Secondary Institution:	

First Author:	Shruti Hegde, MS
First Author Secondary Information:	
Order of Authors:	Shruti Hegde, MS
	Mabon Manoj Ninan, BS
	Jonathan R. Dillman, MD, MSc
	Shireen Hayatghaibi, PhD
	Elanchezhian Somasundaram, PhD
Order of Authors Secondary Information:	
Author Comments:	

Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset

Abstract

Purpose:

General-purpose clinical NLP tools are increasingly used to automatically label clinical reports for research and quality improvement. However, independent evaluations for specific tasks like labeling pediatric CXR reports remain limited. This study compares four commercial clinical NLP models: AWS, AZ, GC, SP, for entity extraction and assertion detection of clinically relevant findings in pediatric CXR reports. Two dedicated CXR report labelers, CheXpert and CheXbert, were also evaluated for comparison.

Materials and Methods:

A total of 95,008 pediatric CXR reports from a large academic hospital were analyzed. Entities and their assertion statuses (*positive, negative, uncertain*) were extracted from findings and impression sections using the four general-purpose models. Entities from impressions were mapped to 12 disease categories plus *No Finding*s using regular expressions. CheXpert and CheXbert processed same reports for the same 13 labels. Outputs from all six systems were compared across assertion categories using Fleiss' Kappa to quantify inter-model agreement. A manual review was done by a board-certified pediatric radiologist on a subset of 360 exams stratified across the 12 CheXpert labels and across the 3 assertion statuses (*positive, negative,* and *uncertain*). Precision, recall and F1-scores were calculated for each NLP service against the ground truth.

Results:

Significant differences were observed in the mean number of extracted entities among the NLP models ($p < 0.001$). SP extracted the most unique entities (49,688), followed by AZ (31,543), AWS (27,216), and GC (16,477). Assertion distributions also differed significantly ($p < 0.001$). For the CheXpert labels comparison, the mean Fleiss' Kappa across all assertion categories, was 0.68 ± 0.17 including all reports and 0.35 ± 0.14, excluding reports where all six models predicted the disease as absent. For manual validation on the subset, CheXbert and CheXpert achieved the highest macro F1-scores (0.565 and 0.561), followed by Google (0.453), Azure (0.421), Spark NLP (0.420), and AWS (0.266). Overall, positive assertions were best detected (F1 0.782), uncertain the worst (0.468). Performance varied by label, with highest F1 for pleural effusion, pneumothorax, and pneumonia, and lowest for lung opacity, enlarged cardiomediastinum, and lung lesion.

Conclusions:

Marked variability exists among NLP models in entity extraction and uncertainty handling, underscoring the need for further validation before clinical deployment.

Keywords and Abbreviations

Natural Language Processing (NLP), Named Entity Recognition (NER), Assertion, Amazon Comprehend Medical (AWS), Google Healthcare NLP (GC), Azure Clinical NLP (AZ), SparkNLP (SP), CheXpert, CheXbert, Chest X-Ray (CXR)

Introduction:

Patient Electronic Health Records (EHRs) have become central to discovery and innovation in data-driven healthcare [1]. EHRs encompass structured data such as diagnostic codes, physiological measurements, and medication records, as well as unstructured free-text clinical health records (e.g., office visit, inpatient, and surgical notes and imaging reports), which provide valuable insights to enhance care provision and inform clinical decision-making [2]. Natural Language Processing (NLP) has become integral to extracting insights from these unstructured clinical reports [3], a task that is highly tedious when performed manually. Advancements in NLP have facilitated various clinical and research applications, including automated clinical coding [4], clinical decision support systems [5], predictive analytics for patient outcomes, and population health management [7]. Additionally, NLP enables the extraction of diagnostic information from radiology reports [8], enhances disease surveillance [9], supports large-scale clinical research [6], improves patient care through personalized treatment recommendations [10] and enables quality improvement efforts.

NLP models for processing clinical text can be categorized into rule-based and statistical learning-based approaches. Rule-based systems utilize medical ontology libraries that host expert-curated knowledge bases encompassing medical concepts, diagnostic codes, and medication categories [11]. In contrast, machine learning (ML) and hybrid (ML + rule-based) tools have been increasingly adopted for general-purpose clinical NLP applications [12]. Notable among these are recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which have significantly advanced the handling of sequential data [13]. However, the introduction of attention mechanisms through the Transformer architecture has surpassed previous benchmarks, establishing Transformers as the industry standard in NLP [14]. Transformer-based methods have led to the development of two primary model families: the GPT (Generative Pre-trained Transformer) family [15], excelling in applications such as question answering and summarization through autoregressive generation, and the BERT (Bidirectional Encoder Representations from Transformers) family [16], optimized for tasks like named entity recognition, assertion detection, entity linking, and relationship extraction. Named Entity Recognition (NER) is the task of identifying and classifying specific pieces of information (called entities) in unstructured text into predefined categories (e.g., disease, procedure entities). Assertion detection determines the status or context of a recognized entity in clinical text (e.g., positive/negative/uncertain). Relationship Extraction is the task of identifying and classifying semantic relationships between two or more entities mentioned in unstructured text. In the clinical context, this means detecting how entities are connected, such as recognizing that a medication is prescribed to treat a specific condition.

Chest radiography (CXR) is pivotal for diagnosing and monitoring respiratory and cardiovascular conditions such as pneumonia [17], tuberculosis [18], lung cancer [19], and heart failure [20]. CXR reports offer detailed diagnostic information that aids in identifying and tracking these conditions, thereby supporting informed treatment and patient management decisions. However, interpreting CXR images is inherently complex, with studies revealing significant variability among physicians and radiologists in their assessments [21]. This complexity makes CXR reports an ideal use case for evaluating NLP models, as they encapsulate nuanced diagnostic information that challenges automated extraction and analysis.

The availability of several large public CXR datasets with images, reports, and disease labels has spurred the development and assessment of advanced computer vision, NLP, and multimodal AI models tailored

specifically to CXR applications [22]. Specifically, the CheXpert framework [23], which includes a comprehensive dataset of CXR images, reports, and associated disease labels, provides multiple report labeling models that are used to generate labels for other public datasets, such as the MIMICS CXR dataset [24]. Numerous research studies rely on these datasets to develop and evaluate their model performance [25]. Concurrently, general-purpose clinical NLP models, ranging from open-source rule-based and Machine Learning models like MetaMap [26], cTAKES [27] and Stanza [28] to recent commercial systems employing proprietary algorithms, are widely utilized for various clinical note processing tasks [29]. However, the performance of these more general systems on specific note types and tasks, such as pediatric CXR report labeling, is not well documented. This lack of standardized, independent comparison leaves users unaware of inherent errors in these systems and how such inaccuracies may propagate into downstream applications. A rigorous, standardized comparison of competing NLP models on independent datasets is crucial to understanding their limitations, assessing their uncertainties, and ensuring more reliable integration into clinical workflows.

The primary objective of this study is to compare four commercial clinical NLP models: Amazon Comprehend Medical (AWS) [30], Google Healthcare NLP (GC) [31], Azure Clinical NLP (AZ) [32], and SparkNLP (SP) from John Snow Labs [33], for extracting clinically relevant entities (commonly known as named entity recognition) and determining their assertion status (also known as assertion detection) from an independent database of pediatric CXR reports. The secondary objective is to evaluate the agreement of two widely used benchmark CXR report labeling models, CheXpert [23] and CheXbert [34], against an extraction pipeline constructed using general-purpose clinical NLP models, assessing how each performs when applied to an independent pediatric dataset.

Methods:

Figure 1 shows the overall methodology of this analysis. Patient consent for this retrospective study was waived by the Institutional Review Board (IRB), which approved this study. All reports were de-identified of Protected Health Information (PHI) by the Radiology Informatics team.

a) Dataset:

Pediatric CXR reports were extracted from the radiology information system at a large pediatric hospital for the study period spanning June 2015 to June 2020. A total of 95,008 reports constituted the data sample for this study's evaluation. The study size was determined by including all eligible pediatric CXR reports available during the study period. No a-priori sample size calculation was performed. Reports were deemed eligible if they were finalized radiologist interpretations of chest radiographs. Examinations with incomplete text, missing impression sections, or non-diagnostic content were excluded. No sampling or case–control matching was applied. The mean age of the participants was 7.5 ± 7.5 years, with a balanced male-to-female ratio of 1:1 (50503 males, 44286 females, 219 unknown). The CXR reports were typically in a semi-structured format, comprising the following sections: 1. Clinical History, 2. Findings, and 3. Impression. Each CXR radiology report was stored as an individual text file, linked via anonymized identifiers to ensure patient confidentiality. On average, the reports were approximately 80 words in length (range: 21-422 words, standard deviation: 28 words).

b) General purpose clinical NLP models:

This study compares four commercial clinical NLP models: AWS [30], GC [31], AZ [32], and SP's NLP models for radiology reports [33]. The clinical NLP models for AWS, AZ, and GC are accessed via their respective cloud-based Application Programming Interfaces (APIs), all of which are compliant with the

Health Insurance Portability and Accountability Act (HIPAA, USA). Deidentified radiology reports were submitted to these APIs, and the output results were returned in JSON format. Each of the three cloud-based systems performed four major tasks: entity extraction, assertion detection, entity linking, and relationship extraction.

In contrast, the Spark NLP (SP) radiology models were accessed through an academic license and operated on local hardware. The extraction pipeline for SP consisted of BERT-based radiology models specifically designed for entity extraction and assertion detection. Unlike the cloud-based systems, the SP models are tailored to handle the unique nuances of radiology reports.

c) Disease entity recognition and quantification:

The output data structure and the number of entity categories extracted were unique to each NLP model. A custom postprocessing pipeline was developed to handle the JSON outputs from each system. The pipeline retained entities assigned to disease-related categories by each system (e.g., 'Medical Condition', 'Symptom', or 'Problem'), while entities categorized as anatomy, procedures, or devices were excluded. Table 1 summarizes the total number of named entity categories extracted by each system and identifies those retained for disease-focused analysis. Entities recognized by each NLP model were normalized using a lemmatization library called Natural Language Toolkit (NLTK) version 3.14 to standardize related terms and lexical variants.

To quantify inter-system variability in disease-entity recognition, we calculated the total number of disease related entities recognized within each report section, the mean number of entities recognized per report and the number of unique entities identified by each system across the dataset. These metrics were used to assess system-specific entity recognition density and semantic breadth. Differences in the number of recognized entities per report between systems were statistically compared using paired t-tests across all model combinations, with Bonferroni correction applied for multiple comparisons. We also manually reviewed the reports with the most disagreement between models to understand why these differences occurred and to verify if the models strictly captured only disease-related terms.

d) Standardization and analysis of assertion status:

Assertion status refers to the contextual classification of recognized entities (e.g., positive, negative, uncertain), which is essential for interpreting clinical findings. Each of the four clinical NLP models identified assertions for extracted entities using proprietary schemas. While AZ, GC, and SP provided categorical assertion statuses, the AWS model provided only a numerical confidence score (CS) for the negation attribute at the time of the study. For a standardized comparative analysis, assertions were consolidated into three categories: positive, negative, or uncertain (Table 2). To standardize the AWS output, a threshold-based calibration approach was developed using the negation CS distribution. Entities with a null negation CS were automatically classified as positive. To establish the mathematical cutoff for the uncertain category, an optimization analysis was conducted. Candidate thresholds were tested across a range of absolute values (0.75 to 0.90) as well as quantile distribution markers (the 25% (0.945) and 50% (0.958) quantiles of the entire dataset's negation score distribution). The resulting AWS assertion distributions at these various cutoffs were then evaluated against the baseline average proportion of negation and uncertainty exhibited by the other three commercial models. The threshold that yielded the highest distribution alignment with the multi-model average was selected as the final CS cutoff for determining AWS uncertainty. For the remaining models, categorical outputs were mapped to

the standardized statuses based on their clinical intent; for instance, SP's "planned" and "someone else" were mapped to the negative category, as they do not represent current findings in the patient.

For the Findings and Impression sections, the distribution of extracted disease entities across the three assertion categories was summarized for each NLP system. Differences in assertion distributions between systems were assessed separately for each section using exam-level generalized estimating equations (GEE) with Poisson errors, clustering by exam to account for within-report correlation. Models included system, assertion category, and their interaction; a significant interaction indicated that assertion distributions differed across systems. Pairwise post hoc comparisons were adjusted using the Holm method. Finally, the frequency of uncertain assertions was analyzed as a function of patient age groups, to assess diagnostic certainty across different pediatric developmental stages.

e) CheXpert disease labels and models:

Two open-source models, CheXpert and CheXbert, were used to extract disease labels from the impression section of the chest radiograph (CXR) reports based on a predefined set of 13 categories relevant to chest radiography. These categories include enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, and a No Findings category indicating reports without known disease findings. CheXpert is a rule-based model that employs heuristic rules and expert-curated mappings to assign disease labels from the textual content of CXR reports. In contrast, CheXbert is built on a BERT-based architecture that leverages deep contextual embeddings for a more nuanced interpretation of CXR reports. Both models produce labels accompanied by three assertion statuses: positive, negative, and uncertain.

f) Mapping clinical entities to CheXpert disease categories and comparison:

Recognized disease entities from the commercial NLP systems, even after lemmatization, exhibited multiple variations for equivalent clinical concepts. To map these varied entities to the disease labels identified by the CheXpert models, a comprehensive regular expression (RegEx) algorithm was developed. RegEx was applied to the extracted entity text in standardized (lemmatized) form, checking whether specific entity and attribute patterns matched the target CheXpert disease labels. This algorithm was adapted directly from the validated, open-source phrase patterns of the Stanford CheXpert labeler codebase. To optimize the framework for our cohort, the underlying dictionary was iteratively expanded via clinical consultation to incorporate specialized pediatric descriptors, relevant abbreviations, and localized institutional reporting nuances unique to our hospital system's dictation templates. By filtering and grouping relevant entities under the appropriate disease labels, the RegEx algorithm provided a consistent, reproducible, and unified categorization across the dataset.

As an example, figure 2 illustrates the mapping for *"enlarged cardiomediastinum",* a condition not often directly mentioned in CXR reports is labeled by identifying related phrases and patterns. The complete details of the RegEx patterns used for each disease label are provided in Appendix A. This standardization was critical for enabling reliable aggregation, comparison, and subsequent analysis of disease labels detected by the commercial NLP models against the CheXpert and CheXbert models.

Inter-model agreement in assertion classification for each disease category was assessed using Fleiss' Kappa across six NLP systems: four commercial models (AWS, AZ, GC, SP) and two open-source models (CheXpert, CheXbert). To facilitate this analysis, a fourth state, "Absent," was assigned to a disease category if a model did not recognize the corresponding label. Fleiss' Kappa was calculated under two

conditions: (i) an inclusive analysis of all reports, and (ii) a sensitivity analysis excluding reports where all six models predicted the disease as “Absent.” The latter condition was used to mitigate the potential inflation of agreement scores driven by uniform non-detection. Finally, the frequency of concordant assertion predictions was visualized for each disease category to characterize inter-model consistency across positive, negative, and uncertain classes.

Additionally, a qualitative review of a subset of reports with conflicting assertion assignments in the Impression section was performed to gain further insight into specific model classification behaviors and diagnostic discrepancies. Given the large volume of disagreement cases, we identified and grouped representative example cases into three major recurring linguistic categories: Negation Scope, Hedging and Probability, and Complex Syntactic Structures. These categories collectively illustrate that clinical NLP performance often depends on subtle linguistic context that current models inconsistently resolve.

g) Manual Validation and LLM analysis:

To further validate model performance, a board-certified pediatric radiologist (blinded) performed a manual review of a subset of 360 reports via a custom web user interface. The sampling strategy targeted 10 reports for each of the three assertion categories across all 12 CheXpert disease labels. For three categories with low clinical prevalence: *enlarged cardiomediastinum, atelectasis, and fracture,* the negative assertion cohort contained fewer than 10 reports; these remaining slots were filled via random sampling from the other two assertion classes (positive and uncertain) for those specific labels. The radiologist evaluated the Impression section of each report and assigned the appropriate assertion category (positive, negative, uncertain), with "absent" serving as the default selection when a disease entity was not mentioned. Using radiologist annotations as the reference standard, we calculated precision, recall, and F1-score for each NLP service after matching reports by exam ID and excluding “absent” manual labels. Metrics were reported as macro-averaged multiclass scores across the 3 assertion classes at the overall service level and for each CheXpert label, with additional class-specific and pooled aggregate analyses.

Further, to analyze the clinical utilities of the raw entities detected by the four clinical NLP services, all the entities extracted from the 360-report manual validation subset were also analyzed by a Large Language Model (medgemma-27b) [36]. The model was prompted to perform a binary classification on each extracted entity to determine whether the term constituted clinically relevant medical terminology or represented non-clinical noise. This automated audit allowed for a systematic, objective assessment of the clinical relevance of the extractions across the different proprietary NLP schemas.

Results:

a) Named entity recognition variability across models:

The clinical NLP model evaluation focused on the total disease-related entity extraction counts from the Findings and Impression sections of the CXR reports. Across these two specific sections, the AZ model demonstrated the highest extraction yield, identifying a combined total of 886,685 entities (689,774 from Findings and 196,911 from Impression). The SP model followed, yielding a total of 628,268 entities (457,540 from Findings and 170,728 from Impression). The AWS model extracted 572,027 entities (417,630 from Findings and 154,397 from Impression). Conversely, the GC model identified 473,717 entities (333,516 from Findings and 140,201 from Impression) across the two report sections. These results highlight a significant differential in extraction density, with the

AZ model identifying approximately 1.87 times the number of entities extracted by the GC model from the Findings and Impression text.

The average number of entities extracted per exam by the four NLP systems is shown in Figure 3a. AZ demonstrated the highest average extraction density, yielding 12.4 ± 3.7 entities. SP followed, extracting an average of 9.5 ± 3.4 entities per exam. AWS had a mean of 8.9 ± 3.0 entities, exhibiting the narrowest observed variability. GC recorded on average 7.8 ± 3.3 entities per exam. All pairwise comparisons in the number of extracted entities per report were statistically significant (Bonferroni-adjusted p-value < 0.001). Figure 3b shows the unique number of disease-related entities extracted by each system across the study dataset. SP extracted considerably more unique entities (49,688) when compared to the other three systems. AZ and AWS extracted 31,543 and 27,216 unique entities, respectively, while GC identified 16,477 unique entities.

The qualitative review of reports with the highest variance in entities recognized revealed that high density was often associated with the inclusion of non-pathological terms. In several instances, the AZ model categorized qualitative modifiers (e.g., "stable," "somewhat") and medical hardware (e.g., "catheter tip," "radiopaque") as clinical symptoms or signs. Similarly, the high unique entity count for SP was driven by the inclusion of complex surgical procedures (e.g., "bidirectional Glenn," "MPA oversewing") and positional descriptors (e.g., "normally aligned"). Conversely, the lower density observed in GC was associated with a more conservative identification of clinical problems, occasionally missing specific components of rare congenital syndromes (e.g., VACTERL association) recognized by other systems.

Figure 4 illustrates the count of the top five most frequently identified findings from the Impression section of the reports, highlighting both frequency and inter-system variability. Pneumonia was the most common entity across all models, with counts ranging from 28,037 (SP) to 32,392 (AZ), reflecting a spread of 4,355 entities (≈14.4% variation). Viral or reactive airway disease showed the smallest variation, with counts tightly clustered between 16,460 (AWS) and 16,391 (SP) - a 0.4% variation, indicating strong inter-system concordance. Atelectasis, however, showed the largest disparity; AZ identified 15,774 instances while SP reported 10,056, a spread of 5,718 entities (≈44.3% variation). Pleural effusion counts were consistent between 3,986 (GC) and 4,793 (AZ), with a difference of 807 (≈18.4% variation). Pneumothorax was the least common in the top five, ranging between 1,885 (GC) and 3,065 (AZ) - 1,180 difference (≈47.7% variation). These statistics suggest that while models demonstrate high concordance on certain entities, like viral or reactive airway, pneumonia and pleural effusion, others such as atelectasis and pneumothorax show greater variability in recognition density.
Assertion distribution in commercial clinical NLP models:

b) Assertion distribution in commercial clinical NLP models:

The threshold calibration for the AWS negation confidence score revealed that the 25th percentile (0.945) provided the closest alignment to the baseline, yielding a negation rate of 21.7% (average: 22.9%) and an uncertainty rate of 7.2% (average: 6.7%). Table 3 presents the distribution of assertion classifications for disease-related entities identified by the four models. In the Findings section, the majority of entities were classified as either positive or negative. Uncertain classifications were infrequent, ranging from 1.4% (GC) to 6.9% (AWS). SP identified 70.5% of Findings as positive, whereas GC identified 36.8%.

In the Impression section, categorical distributions shifted across all models. AWS classified 79.4% of entities as positive, while SP classified 58.2%. Negative assertions were highest for GC (21.8%) and lowest for AWS (2.6%). Notably, the proportion of uncertain classifications was highest for SP (32.6%), compared to AZ (27.9%), AWS (17.9%) and GC (17.2%). Exam-level generalized estimating equations confirmed that assertion distributions differed significantly between NLP systems in both the Findings and Impression sections (overall interaction $p < 0.001$ for both; all pairwise post hoc comparisons remained significant after Holm correction).

As shown in Table 4, an analysis across pediatric age groups revealed that diagnostic uncertainty was highest in the youngest cohort (0–5 years), with SP and AZ reaching 11.5% and 7.7%, respectively. This uncertainty trended downward as patient age increased, reaching its lowest point in the 15–18 year cohort across all models. AWS maintained the lowest frequency of uncertain assertions throughout all age groups (2.3%–3.9%). SP exhibited the most pronounced age-dependent variation, with uncertainty frequency dropping from 11.5% in the 0–5 year cohort to 6.0% in the 15–18 year cohort.

c) Comparison with open-source CXR report labelers on CheXpert labels:

Figure 5 shows the Fleiss' Kappa values for each of 12 disease categories and the *No Findings* category, based on the assertion statuses (positive, negative, uncertain and absent) assigned by all six NLP models. The highest agreement was observed for *pleural effusion* (Kappa: 0.89 in *All*; 0.57 in *Excluding Absent*), while the lowest agreement occurred for e*nlarged cardiomediastinum* (Kappa: 0.24 in *All*; 0.05 in *Excluding Absent*). The mean Fleiss' Kappa across all categories was 0.67 ± 0.17 for the inclusive analysis (All) and 0.33 ± 0.14 for the analysis excluding non-detection (*Excluding Absent)*.

Figure 6 summarizes inter-model agreements across the six NLP systems for each CheXpert disease label and assertion category. Bars indicate the relative frequency of agreement levels (1-6 models), with μ denoting the mean number of models that agreed on the same assertion status and *n* representing the number of reports in which at least one model predicted that disease–assertion combination. Across all labels, agreement patterns varied substantially by assertion category.

Positive assertions showed moderate (μ=3), to high concordance (> 4), with *cardiomegaly* (μ = 4.83) and *pleural effusion* (μ = 4.81) demonstrating the greatest consistency, while *pneumonia* (μ = 2.10) and *enlarged cardiomediastinum* (μ = 1.78) exhibited the weakest alignment. Negative assertions were generally more uniform, peaking for *pleural effusion* (μ = 4.34) and *pneumothorax* (μ = 3.73), but dropping sharply for *enlarged cardiomediastinum* (μ = 1.33) and *atelectasis* (μ = 1.32). In contrast, uncertain assertions showed low consistency, with only *atelectasis* (μ = 2.86) and *edema* (μ = 2.60) approaching moderate agreement while most others remained between 2.30 (*pneumonia)* and 1.20 *(lung opacity)*. The *No Findings* label exhibited the strongest consensus overall, with mean agreements of μ = 4.96 for positive and μ = 4.58 for negative assertions, indicating consistent identification of normal studies across models.

d) Manual verification and LLM classification:

CheXbert and CheXpert had the strongest overall manual-validation performance, with the highest macro F1-scores at 0.565 and 0.561, respectively, and corresponding macro precision/recall of 0.689/0.501 for CheXbert and 0.683/0.500 for CheXpert. Among the cloud-based systems, Google performed best (precision 0.617, recall 0.363, F1 0.453), followed by Azure (0.559, 0.362, 0.421) and Spark NLP (0.551, 0.355, 0.420), while AWS had the weakest performance (0.421, 0.245, 0.266). In the

pooled analysis across all labels and services, macro precision, recall, and F1-score were 0.586, 0.388, and 0.451, respectively. Class-specific results showed that positive assertions were identified most reliably overall (precision 0.786, recall 0.779, F1 0.782), whereas uncertain assertions were substantially more difficult (0.598, 0.384, 0.468), and negative assertions showed high precision but lower recall (0.961, 0.388, 0.553). Across services, the best-performing CheXpert labels overall were *pleural effusion, pneumothorax*, and *pneumonia*, with mean F1-scores of 0.504, 0.494, and 0.476, respectively, whereas *lung opacity*, *enlarged cardiomediastinum*, and *lung lesion* showed the weakest performance, with mean F1-scores of 0.224, 0.268, and 0.293.

Across the validation subset, a total of 5,034 unique free-text entities were extracted: AWS (n = 1,148), GC (n = 938), AZ (n = 1,357), and SP (n = 1,591). Binary classification of these terms segregated them into clinically relevant medical terminology (e.g., specific findings such as pneumonia, pneumothorax, or atelectasis) or irrelevant, non-clinical noise (e.g., generic or ambiguous terms such as feeling fine, dull, fall, or areas).The proportion of entities classified as clinically relevant varied by service: GC achieved the highest relevance rate at 95.3% (894/938), followed closely by AWS at 93.2% (1,070/1,148), while AZ and SP demonstrated lower relevance rates at 87.2% (1,183/1,357) and 85.2% (1,356/1,591), respectively.

e) Qualitative analysis of discrepant assertion statuses on CheXpert labels:

Table 5 provides selected examples of variability in assertion among commercial and open-source NLP models. Given the large volume of disagreements, we identified and grouped representative cases into three recurring linguistic categories: Negation Scope, Hedging and Probability, and Complex Syntactic Structures. Cases 1, 2, and 3 illustrate failures in Negation Scope, where models struggle to correctly associate a negative modifier (e.g., "no," "without") with the specific target entity, often resulting in false positives when "stable" or "no acute" findings are present. Cases 4, 6, and 8 highlight the challenge of Hedging and Probability markers; here, clinical phrases like "may represent," "concerning for," or "is suspected" lead to inconsistent interpretations, with models diverging on whether such language implies a positive diagnosis or mere clinical uncertainty. Finally, cases 5 and 7 demonstrate the impact of Complex Syntactic Structures, where non-standard punctuation, such as slash-delimited differential diagnoses and nested conditional statements, exceed the model's parsing capabilities. These categories collectively illustrate that clinical NLP performance is significantly hindered by the nuanced, probabilistic nature of radiological reporting, where the distinction between a confirmed pathology, a ruled-out condition, and an equivocal finding often depends on subtle linguistic context that current models inconsistently resolve.

In case 1 for *pneumothorax*, the impression "No appreciable pneumothorax on the left" led to mixed interpretations, with SP classifying it as positive while most others marked it negative. For *pneumonia* in case 2, the phrase "without focal pneumonia" resulted in most models classifying it as negative, with only AZ marking it positive and SP classifying it as uncertain. The *cardiomegaly* example (case 3) with "No acute cardiopulmonary abnormality with stable cardiomegaly" term led four models to classify it as positive, while AWS remained uncertain, and SP classified it as negative. In case 4, the hedging phrase "may represent developing consolidation" resulted in three models classifying *consolidation* as uncertain and three as negative.

Similarly, in case 5 for *atelectasis*, the complex description of "airspace disease such as atelectasis/pneumonia" divided models, with three positive, two uncertain, and one negative. In case 6

(edema), the impression “No focal airspace disease or overt pulmonary edema is suspected” was mostly classified as negative, except GC (positive) and AZ (uncertain), reflecting sensitivity to the speculative phrase “is suspected.” For lung lesion in case 7, the uncertain wording “It is not clear whether this represents… or a true finding” led to varied outputs- three models (CheXpert, AWS, GC) marked it positive, while the others (CheXbert, AZ, SP) marked it uncertain, revealing difficulty in interpreting equivocal descriptions. Notably, in case 8, the phrase "opacity concerning for developing pneumonia" led to a split between positive and uncertain classifications of *lung opacity*. These findings demonstrate how radiological language, especially uncertainty markers, qualifiers, and alternative explanations consistently challenge commercial NLP models, with each interpreting probabilistic medical language and hedging expressions differently.

Discussion:

This study is the first to quantify and compare the agreement of commercial clinical NLP models on a large independent pediatric dataset of chest radiograph reports. Commercial clinical NLP models from Amazon, Google, Azure and John Snow Labs were analyzed on the tasks of named entity recognition and assertion detection. A standardization algorithm was developed to map the entities extracted by these models to disease labels defined by the CheXpert framework. The CheXpert and CheXbert models provided a set of disease categories that served as a reference for further comparison. We identified significant inter-system variability that has critical implications for automated clinical phenotyping and large-scale data mining.

a) Implications of entity recognition variability:

Comparison of the disease-related entities extracted by the four commercial NLP models revealed considerable variability in the mean number of entities per report and the number of unique entities detected. AZ recorded the highest mean number of entities per report. However, a qualitative review suggested that this density was frequently driven by "overreaching." Non-pathological modifiers (e.g., "stable") and medical hardware (e.g., "catheter") were categorized as clinical findings. In contrast, SP produced the greatest number of unique entities. This was likely attributable to the usage of the domain-specific *ner_radiology* model for entity detection. Unlike general-purpose systems, this architecture is optimized for the specialized nomenclature and granular anatomical descriptors inherent to radiological reporting. However, it also demonstrated a tendency to incorporate procedural history and positional status within its disease-related results.

In addition, the top five most frequent disease or diagnosis entities reported had variable agreement between the models. Viral or reactive airway disease had near perfect agreement (< 1% difference in counts) between the four models. In contrast, pneumothorax and atelectasis had large differences (> 40% difference in counts). These results illustrate the inconsistencies in named entity recognition between models. They also highlight that higher extraction density does not always correlate with clinical precision.

b) Variability in assertion detection:

Significant variability was observed in how the four models classified assertion status, largely because each system uses a unique and non-standardized schema. After grouping these outputs into three categories, we found major differences in their distributions. This was especially evident in the Impression section where uncertainty ranged from 17.2% to 32.6%, compared to a much narrower 1.4%

to 6.9% in the Findings. These assertion profiles are also influenced by each model's entity recognition logic. Systems that capture medical hardware or surgical history as "positive" disease entities can artificially inflate the positive category, which may obscure or dilute the true frequency of uncertain clinical findings. The analysis revealed a consistent trend across Azure, Google, and SparkNLP where diagnostic uncertainty was highest in the youngest patients (0–5 years) and generally declined as they matured, reaching its lowest point in the 15–18 year cohort. This pattern aligns with the more cautious, speculative reporting style often used for infants and toddlers, where anatomical variability necessitates increased clinical hedging. Importantly, the ability of these three models to capture this trend is a positive finding. It suggests that they are effectively reflecting the inherent ambiguity of the source reports. In contrast, the AWS model maintained a nearly flat frequency of uncertain assertions across all age groups. This relative lack of sensitivity likely stems from the use of negation confidence scores as a proxy for uncertainty, which proved less responsive to the nuanced linguistic patterns identified by models with dedicated assertion schemas.

c) Agreement on CheXpert labels:

Inter-model agreement for assertion classification was substantial when cases with no detected labels were included (mean Fleiss $\kappa = 0.67$). However, agreement declined to fair when restricted to reports where at least one model identified the label ($\kappa = 0.33$). This significant drop indicates that while commercial systems generally agree on which disease entities are mentioned, they differ fundamentally in how they interpret clinical status. The quantitative data shows that while models align well on common positive findings like pleural effusion. However, agreement was nearly random (average $\boldsymbol{\mu} = 1.8$) when classifying uncertainty. Our qualitative review revealed that this "uncertainty gap" is driven by inconsistent handling of speculative phrases (e.g., "is suspected") and complex syntax, such as differential diagnoses separated by slashes.

These inconsistences particularly concerning for the medical AI community. Benchmark chest X-ray datasets such as CheXpert [23] and MIMIC-CXR [24] rely on automatically generated labels that may contain systematic noise. Differences in how NLP systems handle assertion terms can therefore introduce systematic label noise. This noise may propagate through model training and may affect reported generalizability of diagnostic AI models. Previous analyses have raised similar concerns about the reliability of automatically derived labels and their influence on downstream diagnostic model accuracy [35].

d) Observations from manual validation and LLM analysis:

Manual validation on a subset of the data showed that domain-specific models performed best, with CheXbert and CheXpert outperforming the cloud NLP services on precision, recall, and F1-score. Performance varied substantially by label and assertion class: positive findings were identified most reliably, while uncertain assertions were harder to classify, and negative assertions often had lower recall despite high precision. At the label level, pleural effusion and pneumothorax showed the strongest overall performance, whereas lung opacity and enlarged cardiomediastinum were the most difficult labels across services.

Complementing these manual metrics, the automated LLM audit confirmed that variations in entity extraction density were tightly linked to linguistic noise rather than clinical precision. While Google and

AWS maintained high clinical relevance rates (>93%) by remaining conservative, systems with higher extraction densities like Azure and SparkNLP showed a higher propensity for incorporating non-clinical terminology, dropping their clinical relevance rates to 87.2% and 85.2%, respectively.

e) Limitations and future directions:

Investigation of cases with discrepant assertion statuses revealed that linguistic nuances influence each NLP model differently. No single model demonstrated uniform superiority. Instead, output varied by disease category and the contextual phrasing within report sections as shown by the examples in Table 5. Each model shows strengths and limitations depending on the linguistic structure, diagnostic ambiguity, and type of condition described.

These observations underscore the critical need for institutional validation and a standardized framework for defining uncertainty before these models are deployed in clinical workflows. A technical limitation of this study is that the AWS model evaluated did not provide a categorical assertion status. While we employed a threshold-based proxy for uncertainty, newer iterations of the service have introduced explicit confidence attributes that may offer a more refined estimation of diagnostic hedging. Furthermore, this study focused exclusively on pediatric reports from a single institution. As a result, the generalizability of the findings may be limited by institutional variations in reporting styles and the specific characteristics of the pediatric cohort. Such efforts are necessary to ensure that inherent errors do not propagate into patient care or research outcomes.

Conclusion:

Significant variability exists in named entity recognition and assertion assessment on CXR radiology reports across different NLP models. These differences arise from variations in the clinical concept definitions inherent to each model and the absence of a uniform approach to quantifying uncertainty expressions. In addition, the diversity in how CXR reports describe radiological findings and impressions results in models performing optimally under different conditions. Consequently, the use of automated NLP models for labeling imaging reports and for downstream applications such as outcome predictions must be accompanied by careful task specific evaluation and oversight. Ensuring the reliability of these tools is essential for the integrity of large-scale medical datasets and the diagnostic AI models derived from them.

Acknowledgements:

The authors used ChatGPT (OpenAI, GPT-4o) to assist with language refinement and grammar checking.

Statements and Declarations:

1. **Consent to participate: Institutional Review Board Waived Informed Consent for this Retrospective study.**
2. **Consent to publish: Institutional Review Board Waived Informed Consent for this Retrospective study.**
3. **Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.**
4. **Funding: The project described was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health, under Award Number 2UL1TR001425-05A1.**

b) **Author Contributions: All authors contributed to the study conception and design. Material preparation, data cleaning, and analysis were performed by Shruti Hegde, Mabon Manoj Ninan and Elanchezhian Somasundaram. Material review was done by Jonathan R. Dillman and Shireen Hayatghaibi. All authors read and approved the final manuscript.**

5. **Human Ethics and Consent to Participate: Human Ethics and Consent to Participate declarations: not applicable.**
6. **Data Availability declaration: Data are however available from the authors upon reasonable request.**

References:

1. Soumya Upadhyay, H.-f.H., A Qualitative Analysis of the Impact of Electronic Health Records (EHR) on Healthcare Quality and Safety: Clinicians' Lived Experiences. 22, Mar 3.
2. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, Specchia ML. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health. 2016 Feb;26(1):60-4. doi: 10.1093/eurpub/ckv122. Epub 2015 Jun 30. PMID: 26136462.
3. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov 21. PMID: 29162496; PMCID: PMC5771858.
4. Hang Dong, M.F., William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, Honghan Wu, Automated clinical coding: what, why, and where we are? npj Digital Medicine, 2022.
5. Ahmed, Usman & Iqbal, Khurshed & Aoun, Muhammad & Khan, Ghazi. (2023). Natural Language Processing for Clinical Decision Support Systems: A Review of Recent Advances in Healthcare. J Intell Connect Emerg Technol, 2023. 8(2): p. 1-17.
6. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, Osborn D, Hayes J, Stewart R, Downs J, Chapman W, Dutta R. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform. 2018 Dec;88:11-19. doi: 10.1016/j.jbi.2018.10.005. Epub 2018 Oct 24. PMID: 30368002; PMCID: PMC6986921.

7. Jerfy, A., O. Selden, and R. Balkrishnan, The Growing Impact of Natural Language Processing in Healthcare and Public Health. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 2024. 61: p. 00469580241290095.
8. Ewoud Pons, L.M.M.B., M G Myriam Hunink, Jan A Kors Natural Language Processing in Radiology: A Systematic Review. RSNA, 2016.
9. Mahmud Omar, D.B., Benjamin Glicksberg, Eyal Klang, Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. American Journal of Infection Control, 2024. 52(9).
10. Kevin B Johnson, W.Q.W., Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, Jane L Snowdon, Precision Medicine, AI, and the Future of Personalized Health Care. CTS, 2020.
11. Papadopoulos, Petros & Soflano, Mario & Chaudy, Yaelle & Adejo, Olugbenga & Connolly, Thomas. (2022). A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. Health and Technology. 12. 1-15. 10.1007/s12553-022-00672-9.
12. Slawomir Kierner, J.K., Zofia Kierner, Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review. 2023.
13. Benyamin Ghojogh, A.G., Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and Survey. 2023.
14. Ashish Vaswani, N.S., Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need. 2017.
15. Radford, Alec and Karthik Narasimhan. “Improving Language Understanding by Generative Pre-Training.” (2018).
16. Jacob Devlin, M.-W.C., Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
17. Makhnevich A, Sinvani L, Cohen SL, Feldhamer KH, Zhang M, Lesser ML, McGinn TG. The Clinical Utility of Chest Radiography for Identifying Pneumonia: Accounting for Diagnostic Uncertainty in Radiology Reports. AJR Am J Roentgenol. 2019 Dec;213(6):1207-1212. doi: 10.2214/AJR.19.21521. Epub 2019 Sep 11. PMID: 31509449.
18. Piccazzo, R., F. Paparo, and G. Garlaschi, Diagnostic accuracy of chest radiography for the diagnosis of tuberculosis (TB) and its role in the detection of latent TB infection: a systematic review. The Journal of Rheumatology Supplement, 2014. 91: p. 32-40.
19. Kim, J. and K.H. Kim, Role of chest radiographs in early lung cancer detection. Translational lung cancer research, 2020. 9(3): p. 522.
20. Cardinale L, Priola AM, Moretti F, Volpicelli G. Effectiveness of chest radiography, lung ultrasound and thoracic computed tomography in the diagnosis of congestive heart failure. World J Radiol. 2014 Jun 28;6(6):230-7. doi: 10.4329/wjr.v6.i6.230. PMID: 24976926; PMCID: PMC4072810.
21. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary? Postgrad Med J. 2003 Apr;79(930):214-7. doi: 10.1136/pmj.79.930.214. PMID: 12743338; PMCID: PMC1742668.
22. Anis, Shazia & Lai, Khin Wee & Chuah, Joon Huang & Ali, Muhammad & Mohafez, Hamidreza & Hadizadeh, Maryam & Ding, Yan & Chao, Ong. (2020). An Overview of Deep Learning Approaches in Chest Radiograph. IEEE Access. 8. 182347 - 182354. 10.1109/ACCESS.2020.3028390.
23. Jeremy Irvin, P.R., Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, Andrew Y. Ng, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 2019.

24. Alistair E. W. Johnson, T.J.P., Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark & Steven Horng MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. 2019.
25. Chambon, Pierre & Blüthgen, Christian & Delbrouck, Jean-Benoit & Sluijs, Rogier & Połacin, Małgorzata & Zambrano Chaves, Juan Manuel & Abraham, Tanishq & Purohit, Shivanshu & Langlotz, Curtis & Chaudhari, Akshay. (2022). RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. 10.48550/arXiv.2211.12737.
26. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. in Proceedings of the AMIA Symposium. 2001.
27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13. doi: 10.1136/jamia.2009.001560. PMID: 20819853; PMCID: PMC2995668.
28. Peng Qi, Y.Z., Yuhui Zhang, Jason Bolton and Christopher D. Manning., Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020., 2020.
29. Bai, Lu & Mulvenna, Maurice & Wang, Zhibao & Bond, Raymond. (2021). Clinical Entity Extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical. 1-6. 10.1109/ISSC52156.2021.9467856.
30. AWS. *https://aws.amazon.com/comprehend/medical/*. Accessed 26 Jun, 2026.
31. GC. *Google Cloud*. Available from: https://cloud.google.com/healthcare-api/. Accessed 26 Jun, 2026.
32. AZ. *Microsoft Azure*. Available from: https://azure.microsoft.com/en-us/. Accessed 26 Jun, 2026.
33. JSL. *John Snow Labs*. Available from: https://www.johnsnowlabs.com/. Accessed 26 Jun, 2026.
34. Akshay Smit, S.J., Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, Matthew P. Lungren, *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT.* 2020.
35. Oakden-Rayner, L., *Exploring large-scale public medical image datasets.* Academic radiology, 2020. **27**(1): p. 106-112.
36. MedGemma. https://huggingface.co/google/medgemma-27b-it. Accessed 26 Jun, 2026.

Tables:

Table 1: Details of the commercial NLP models used in this study, the number of named entity categories detected by each system, and the selected categories that relate to disease symptoms or findings.

NLP model	Named Entity Categories	Categories Selected for Analysis
AWS – Amazon Medical Comprehend [v1.0.0, DetectEntities]	7	'MEDICAL_CONDITION'
AZ – Azure Text Analytics for Health	36	'SYMPTOM_OR_SIGN', 'DIAGNOSIS'
GC – Google Healthcare NLP model	28	'PROBLEM'
SP – John Snow Labs Spark NLP [ner_radiology model]	13	'Disease_Syndrome_Disorder', 'Symptom', 'ImagingFindings'

Table 2: Standardization criteria for each NLP model, mapping assertion outputs to three standardized assertion statuses: Positive, Negative, and Uncertain for analysis. CS is the negation confidence score for AWS.

NLP model	Positive	Negative	Uncertain
AWS – Amazon Medical Comprehend [v1.0.0, DetectEntities]	CS = 0	*CS > 0.945	0 >= *CS <= 0.945
AZ – Azure Text Analytics for Health	'positive'	'negative'	'positivepossible', 'negativepossible', 'neutralpossible'
GC – Google Healthcare NLP model	'likely'	'unlikely'	'somewhat_likely', 'somewhat_unlikely', 'uncertain', 'conditional'
SP – John Snow Labs Spark NLP [ner_radiology model]	'present', 'none', 'past'	'absent', 'family', 'someone else', 'planned'	'hypothetical', 'possible'

*CS stands for confidence score

Table 3: Distribution of assertion detection across different models (AWS, AZ, GC, SP) for various sections in CXR reports. Percentages are rounded to one decimal place.

Assertion	System	FINDINGS Percentage (Assertion count / Total entities)	IMPRESSIONS Percentage (Assertion count / Total entities)
Positive	AWS	51.2% (213,654 / 417,630)	79.5% (122,652 / 154,397)
	AZ	72.3% (499,003 / 689,774)	67.5% (132,857 / 196,911)
	GC	36.8% (122,789 / 333,516)	61.0% (85,551 / 140,201)
	SP	70.5% (322,788 / 457,540)	58.2% (99,342 / 170,728)
Negative	AWS	41.9% (174,987/417,630)	2.6% (4,014/ 154,397)
	AZ	26.1% (180,147 / 689,774)	4.6% (9,125 / 196,911)
	GC	61.8% (206,052 / 333,516)	21.8% (30,498 / 140,201)
	SP	26.9% (123,421 / 457,540)	9.2% (15,777 / 170,728)
Uncertain	AWS	6.9% (28,816/417,630)	17.9% (27,637/154,397)
	% AZ	1.5% (10,624 / 689,774)	27.9% (54,929 / 196,911)
	GC	1.4% (4,675 / 333,516)	17.2% (24,152 / 140,201)
	SP	2.5% (11,331 / 457,540)	32.6% (55,609 / 170,728)

Table 4: Distribution of diagnostic uncertainty across AWS, Azure, Google Cloud, and SP models stratified by pediatric age cohorts. Uncertainty percentages are rounded to one decimal place. (Count of exams with unknown ages = 169)

Age range (years)	Number of exams n	Uncertainty Percentages			
		AWS	AZ	GC	SP
0 – 5	43,246	3.9%	7.7%	4.7%	11.5%
5 – 10	18,728	2.7%	5.3%	4.7%	9.1%
10 – 15	14,419	2.2%	3.9%	4.2%	6.8%
15 – 18	10,420	2.1%	3.3%	3.8%	6.0%
18+	8,026	2.3%	4.4%	4.7%	7.2%

Table 5: Variability in Assertion Among Commercial NLP Models - Divergent NLP Interpretations of Radiology Reports Impressions.

Sample	Disease Label	Report Impression	Positive	Uncertain	Negative
1	pneumothorax	1. Pectus bars in place. 2. No appreciable pneumothorax on the left.	SP	CheXpert, CheXbert, AWS, GC	AZ
2	pneumonia	Findings consistent with viral or reactive airways disease without focal pneumonia.	AZ	SP	CheXpert, CheXbert, AWS, GC
3	cardiomegaly	No acute cardiopulmonary abnormality with stable cardiomegaly and fracture of one of the pacemakers leads.	CheXpert, CheXbert, AZ, GC	AWS	SP
4	consolidation	Right suprahilar opacity may represent developing consolidation, superimposed on	AWS, AZ, SP	CheXpert, CheXbert, GC	

		findings of viral or reactive airways disease.			
5	atelectasis	Viral/reactive airway disease, superimposed with right upper and left lower lobe airspace disease such as atelectasis/pneumonia.	AWS, AZ, GC	CheXpert, CheXbert	SP
6	edema	Mildly prominent pulmonary vasculature. Cardiac size appears mildly enlarged. No focal airspace disease or overt pulmonary edema is suspected.	GC	AZ	CheXpert, CheXbert, AWS, SP
7	lung lesion	No consolidation. Small oval lucency in the left upper lobe. It is not clear whether this represents superimposed shadows (mock effect) or a true finding. A short-term follow-up two-view chest x-ray is suggested to evaluate the persistence of this finding. If it does persist, it may represent a tiny bleb, bulla, or pneumatocele. The surrounding lung appears normal making a cavitary lesion less likely.	CheXpert, AWS, GC	CheXbert, AZ, SP	
8	lung opacity	Poorly defined left lower lobe opacity concerning for developing pneumonia.	CheXpert, CheXbert, AWS	AZ, GC, SP	

Figure Legends:

Figure 1: This flowchart illustrates the methodology starting from pediatric CXR report extraction, followed by processing using the six NLP models. Postprocessing steps include 1) entity standardization, 2) assertion categorization, and 3) Regular Expression (RegEx) conversion to CheXpert labels. Agreement of the models in entity extraction and assertion detection on all entities extracted by the commercial NLP models and on CheXpert specific disease labels were analyzed.

Figure 2: This image illustrates the regex (regular expression) used to search for phrases and patterns related to the "enlarged cardiomediastinum" condition. The Chest Anatomy Structures list the base keywords that must be present, followed by Attributes that describe the anatomical structures. This combination forms a valid pattern for identifying "enlarged cardiomediastinum". However, if Chest Anatomy Structures are followed by certain Mediastinal Conditions, the pattern is considered invalid and is excluded from the "enlarged cardiomediastinum" search.

Figure 3: Unique number of disease related entities extracted by each clinical NLP model on 95,008 pediatric CXR reports. AWS – Amazon Medical Comprehend, AZ – Azure Text Analytics for Health, GC – Google Healthcare Natural Language API, SP – John Snow Labs' Spark NLP models for radiology.

Figure 4: Count of top 5 most frequently extracted disease entities by the 4 commercial NLP models on the study dataset from the Impression section of the reports. AWS – Amazon Medical Comprehend, AZ – Azure Text Analytics for Health, GC – Google Healthcare Natural Language API, SP – John Snow Labs' Spark NLP models for radiology.

Figure 5: Fleiss' Kappa values for each CheXpert label calculated across the six NLP models under two conditions: (All) includes all reports regardless of whether the disease was detected; (Excluding Absent) excludes reports where the disease was not detected by any of the six NLP models.

Figure 6: Inter-model agreement of assertion statuses. Distribution of agreement levels (1-6 services) across (a) positive, (b) negative, and (c) uncertain assertions for the CheXpert labels. Bars show percentage of cases where specific numbers of services predicted the same assertion category. n indicates total cases where at least 1 model made that specific assertion; μ indicates the mean agreement level for each disease, representing the average number of models that agreed on the assertion for that label.

Figures:

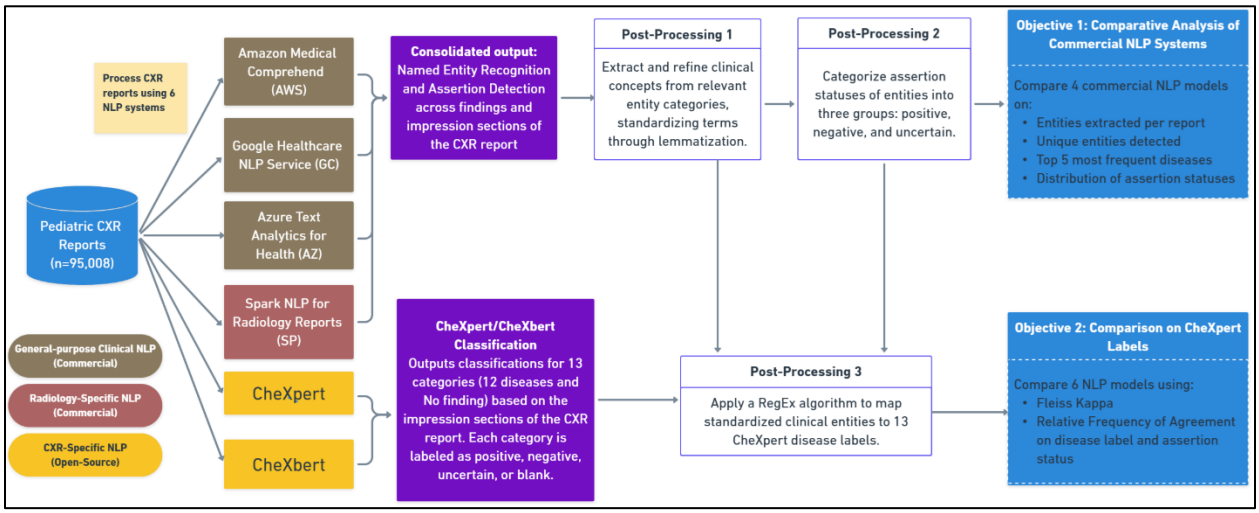


Figure 1: This flowchart illustrates the methodology starting from pediatric CXR report extraction, followed by processing using the six NLP models. Postprocessing steps include 1) entity standardization, 2) assertion categorization, and 3) Regular Expression (RegEx) conversion to CheXpert labels. Agreement of the models in entity extraction and assertion detection on all entities extracted by the commercial NLP models and on CheXpert specific disease labels were analyzed.

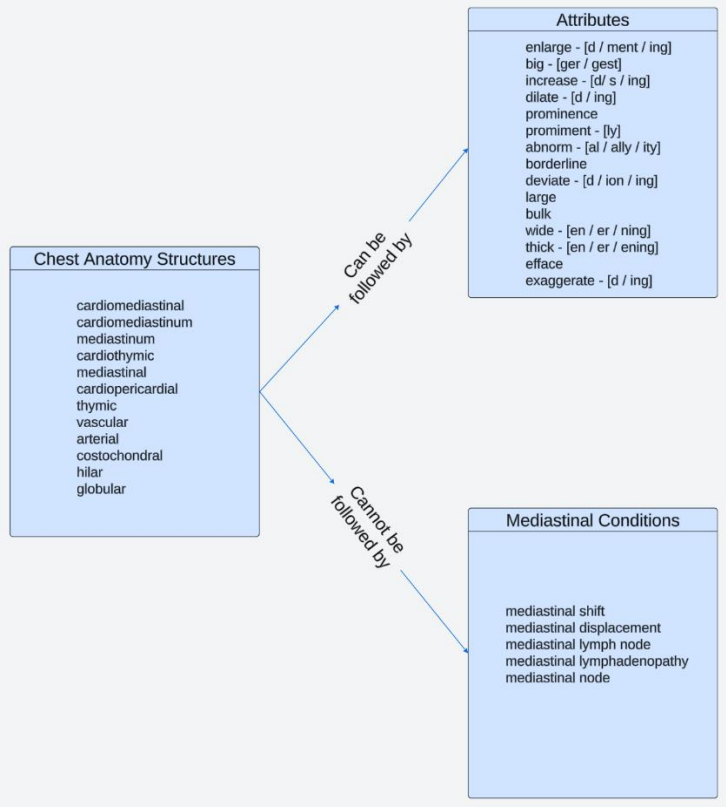


Figure 2: This image illustrates the regex (regular expression) used to search for phrases and patterns related to the "enlarged cardiomediastinum" condition. The Chest Anatomy Structures list the base keywords that must be present, followed by Attributes that describe the anatomical structures. This combination forms a valid pattern for identifying "enlarged

cardiomediastinum". However, if Chest Anatomy Structures are followed by certain Mediastinal Conditions, the pattern is considered invalid and is excluded from the "enlarged cardiomediastinum" search.

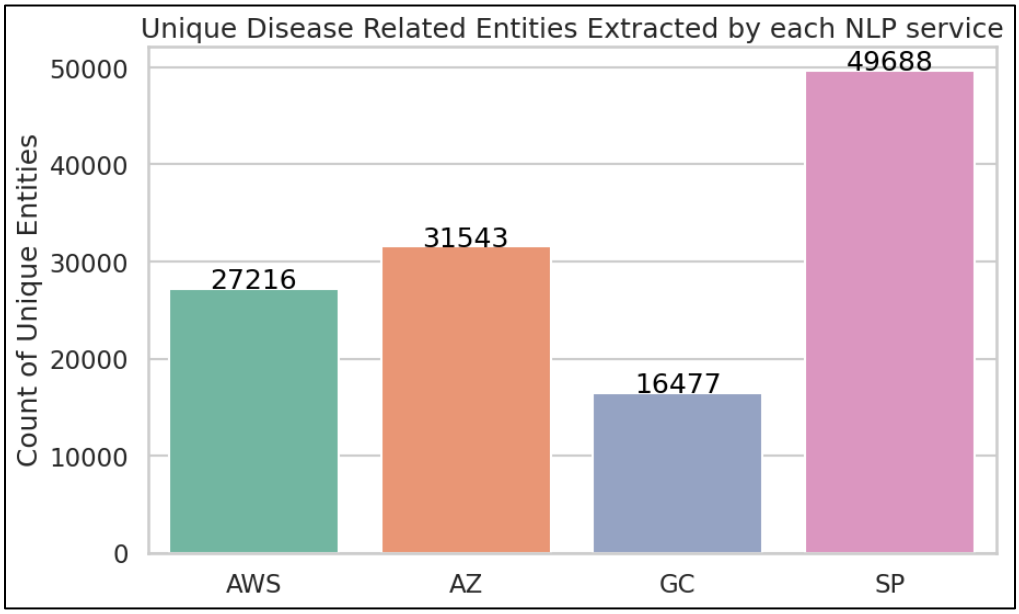


Figure 3: Unique number of disease related entities extracted by each clinical NLP model on 95,008 pediatric CXR reports. AWS – Amazon Medical Comprehend, AZ – Azure Text Analytics for Health, GC – Google Healthcare Natural Language API, SP – John Snow Labs' Spark NLP models for radiology.

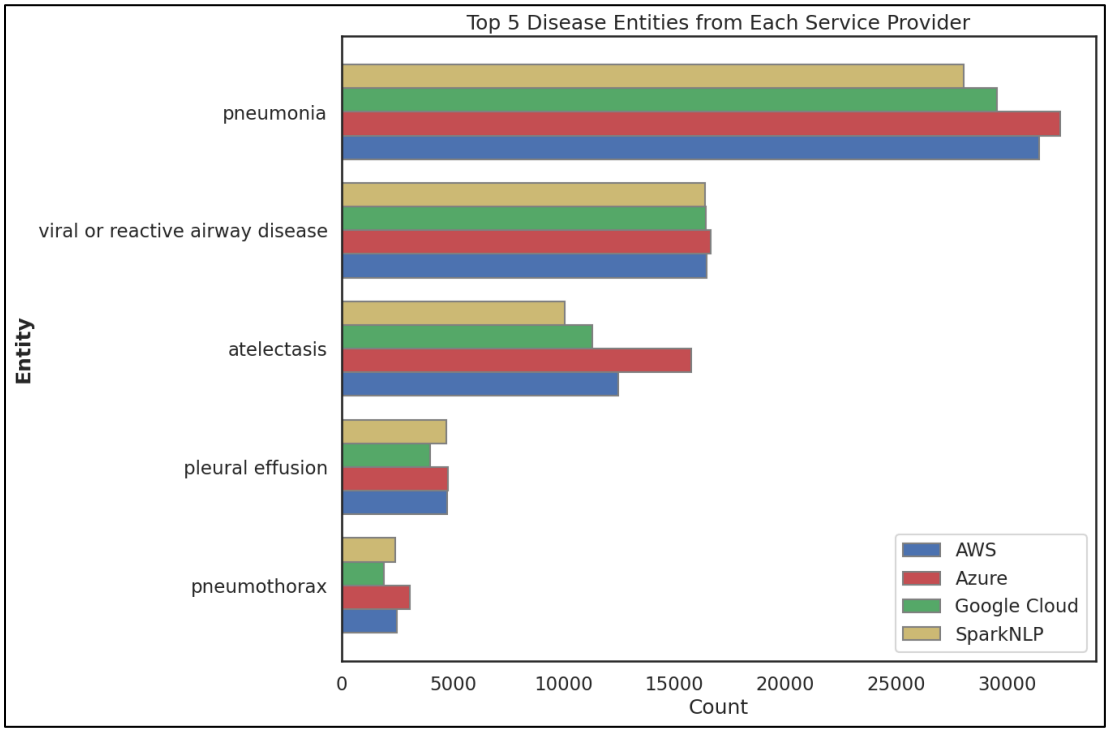


Figure 4: Count of top 5 most frequently extracted disease entities by the 4 commercial NLP models on the study dataset from the Impression section of the reports. AWS – Amazon Medical Comprehend, AZ – Azure Text Analytics for Health, GC – Google Healthcare Natural Language API, SP – John Snow Labs' Spark NLP models for radiology.

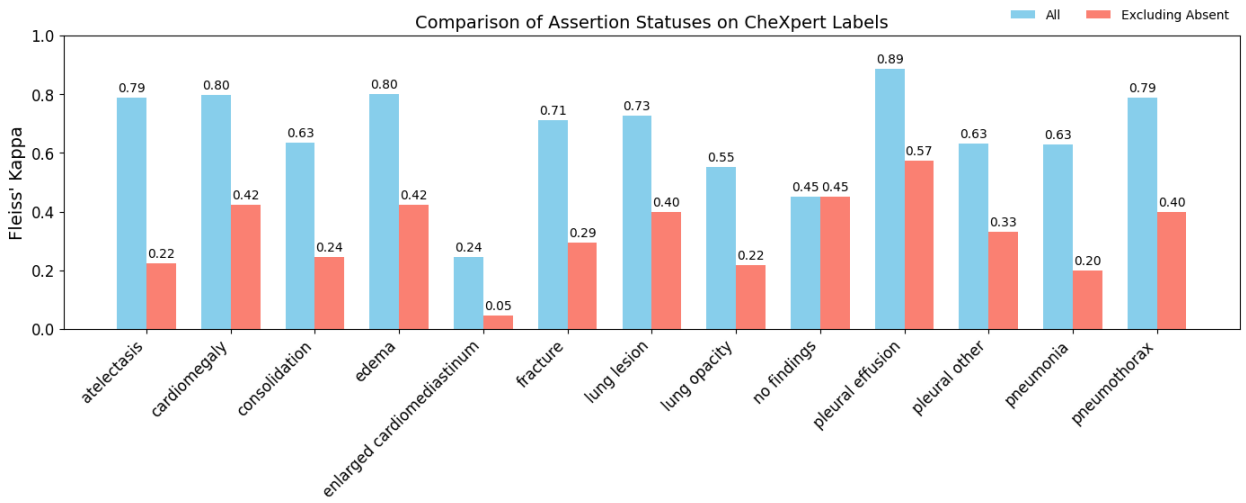


Figure 5: Fleiss' Kappa values for each CheXpert label calculated across the six NLP models under two conditions: (All) includes all reports regardless of whether the disease was detected; (Excluding Absent) excludes reports where the disease was not detected by any of the six NLP models.

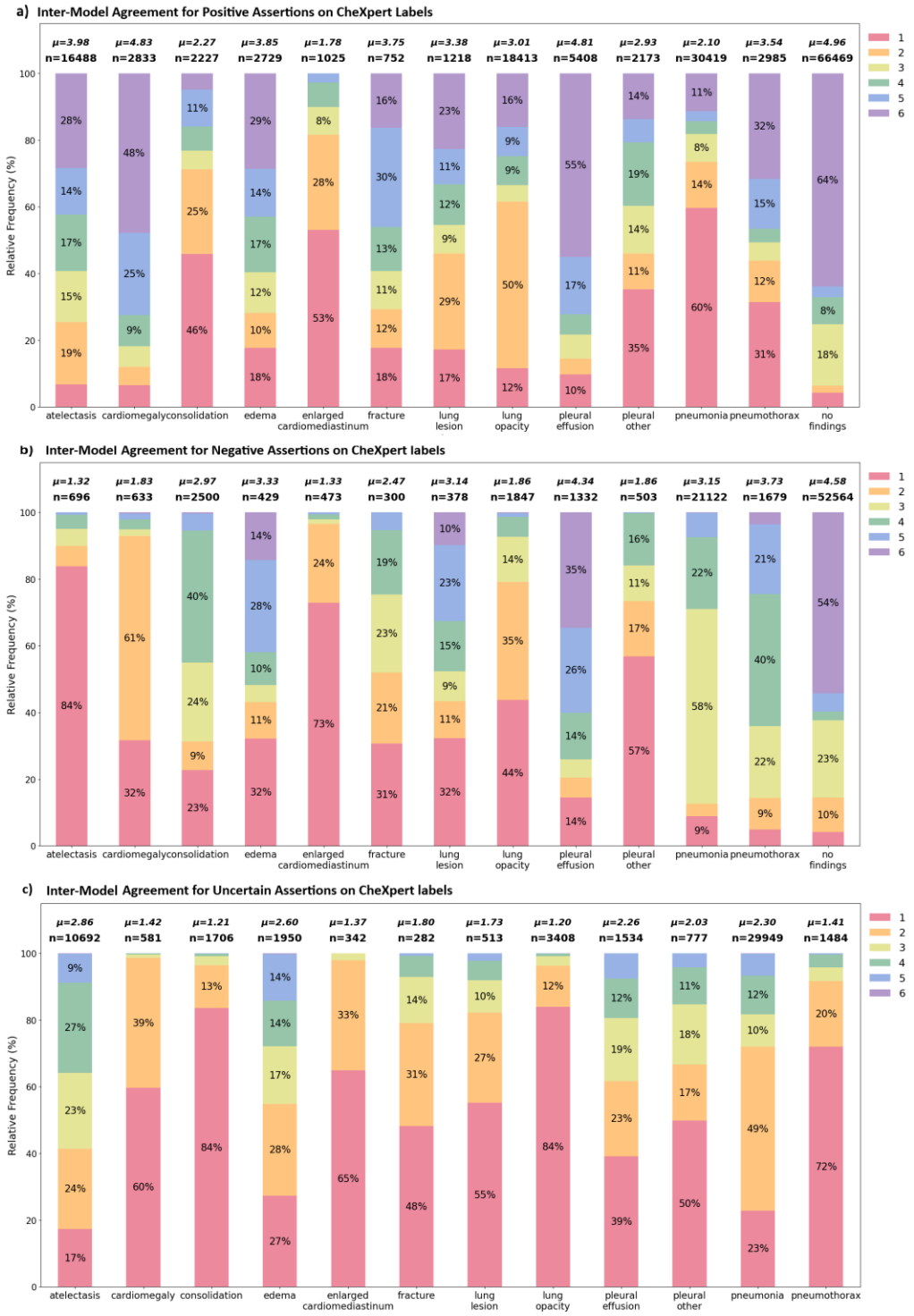


Figure 6: Inter-model agreement of assertion statuses. Distribution of agreement levels (1-6 services) across (a) positive, (b) negative, and (c) uncertain assertions for the CheXpert labels. Bars show percentage of cases where specific numbers of services predicted the same assertion category. n indicates total cases where at least 1 model made that specific assertion; μ indicates the mean agreement level for each disease, representing the average number of models that agreed on the assertion for that label.

Appendix A

Cover Letter

Manuscript with tracked changes

Dear Editor,

Thank you for providing us with the opportunity to revise our manuscript titled "Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset" for publication in the Journal of Digital Imaging (or Journal of Imaging Informatics in Medicine). We sincerely appreciate the editor and reviewers for their constructive feedback and valuable insights.

In this revised version, we have thoroughly addressed all the comments raised during the peer-review process. Specifically, we have performed the following major updates to strengthen the manuscript:

- Manual Validation: Completed a manual review by a board-certified pediatric radiologist on a subset of 360 exams to rigorously validate the assertion performance of the NLP services against a reference standard.
- Clinical Relevance Analysis: Integrated a Large Language Model (LLM)-based evaluation to systematically analyze and categorize the clinical relevance of the most frequent entities extracted by each commercial service.
- AWS Threshold Calibration: Conducted an optimization analysis to establish a systematic, data-driven confidence score threshold for mapping uncertainty within the AWS Comprehend Medical framework, as suggested by the reviewers.

We believe these substantial additions have significantly improved the technical rigor, clinical depth, and overall quality of our work. We remain fully open to any further questions or suggestions the reviewers may have.

Our point-by-point responses to the individual reviewer comments are detailed below.

Comments to the author (if any):
Additional points to address:
1) Include IRB statement in methods.

We verified that the IRB statement is included in methods.

2) Verify references conform to journal format requirements.

We verified that the references conform to journal format requirements.

Reviewer #1: This study compares four commercial clinical NLP systems (AWS, Azure, Google, SparkNLP) and two open-source CXR report labeling models (CheXpert and CheXbert). The

dataset is large and independent (95,008 pediatric CXR reports). The study evaluates named entity recognition, assertion classification, and inter-model agreement using Fleiss' Kappa. The results show substantial variability across NLP systems, especially in uncertainty detection and entity extraction. The study addresses an important problem in clinical NLP and automated report labeling. Overall, this is a well-designed and important study. However, this paper requires minor revision before resubmission. The summary and detailed revision suggestions are provided below.

1. The study measures agreement, not accuracy. There is no manually annotated gold standard dataset. Therefore, the true performance of each model is unknown. The authors should include a manually annotated subset. This would allow calculation of precision, recall, and F1-score.

Thank you for this recommendation. We sampled a subset of 360 exams stratified across the 12 CheXpert labels and across the 3 assertion statuses (Positive, Negative, and Uncertain). We created a custom UI and a pediatric radiologist (author on this paper) reviewed the impression section of the reports and assigned the appropriate rating for each label. We then calculated precision, recall and F1-scores for each NLP service against this ground truth.

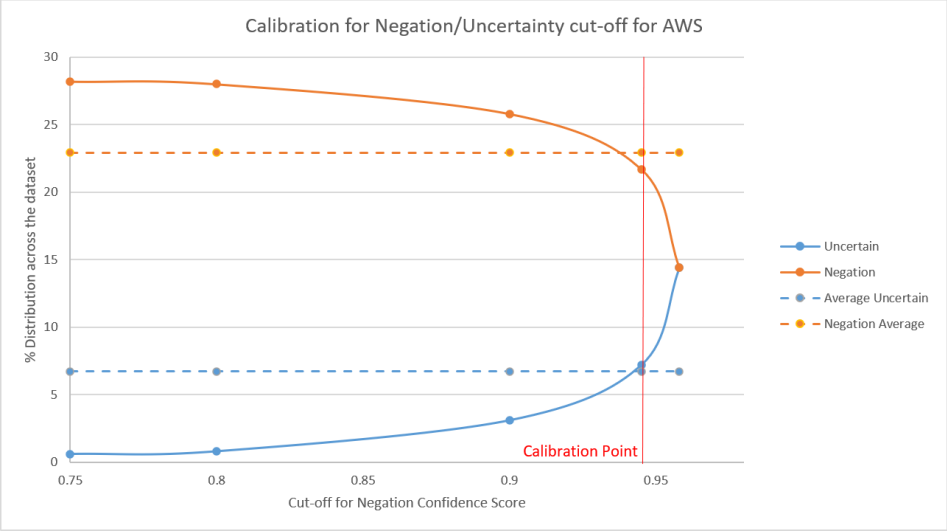
Changes:

1. In Methods, section g) “Manual Validation and LLM analysis” was added to describe the manual review process and metrics calculation.
2. In Results, section d) ”Manual Verification and LLM classification” an additional paragraph discussing the results comparing the NLP service performance to the manual ratings is added.
3. Discussion (section d) and abstract are also appropriately updated with the results from manual validation.

2. AWS did not provide categorical assertion labels. The study used threshold values to classify assertions. These thresholds appear heuristic. They may affect the uncertainty results. The authors should justify the thresholds. A sensitivity analysis is recommended. Newer AWS versions could also be considered.

We appreciate the reviewer’s insightful suggestion to evaluate the latest version of Amazon Comprehend Medical. Because our data processing pipeline was locked for this retrospective cohort study and additional API access was restricted at the time of evaluation, re-running the extensive text corpus through an updated service iteration was not feasible.

However, to fully address the core of the reviewer's comment regarding the determination of the AWS uncertainty threshold, we have replaced the previous threshold strategy with a highly robust calibration optimization. Rather than relying on arbitrary cut-offs, we computed the global, baseline average distribution of negated and uncertain assertions across the other three commercial NLP models (Azure, Google, and SparkNLP) over the entire dataset. We then evaluated candidate AWS confidence score thresholds across absolute values and dataset quantiles to identify the exact point where the AWS assertion distribution mathematically aligned with this multi-model baseline.



Changes:

a) In Methods, section d) “Standardization and analysis of assertion status” was modified to reflect this calibration process for AWS.
b) The resulting changes were applied in Results, sections b) “Assertion distribution in commercial clinical NLP models” and c) “Comparison with open-source CXR report labelers on CheXpert labels”. Figures 5, 6 were also and tables 2, 3 and 4 were also updated.

3. The study compares the number of extracted entities. However, more entities does not mean better performance. Some models may extract non-disease terms. The authors should evaluate the clinical relevance of extracted entities. Manual review of a subset is recommended.

We completely agree with the reviewer that a higher volume of extracted entities does not inherently translate to superior performance, as it can introduce non-clinical noise or irrelevant modifiers. Our intention was not to equate extraction density with accuracy, but rather to highlight the profound systemic inconsistency in how these commercial vendors define, categorize, and capture clinical concepts on an identical dataset. This architectural ambiguity is

precisely why we originally mapped these highly disparate entities to the 12 standardized CheXpert disease labels to ensure a fair comparison.

To fully address the reviewer's concern regarding performance and the clinical relevance of these extractions, we have introduced two major components to this revised manuscript using our validation subset:

- **Radiologist Assertion Validation:** A board-certified pediatric radiologist manually reviewed the subset to assign true CheXpert labels and assertion classes, allowing us to report standard validation metrics (precision, recall, F1-score) for each service against a true reference standard.
- **LLM Clinical Relevance Audit:** We leveraged a clinically fine-tuned Large Language Model (MedGemma) to perform a binary classification on the raw entities extracted by each service within this subset, explicitly auditing them for clinically relevant medical terminology versus non-clinical noise.

Changes:

1. In Methods, section g) Manual Validation and LLM analysis, a paragraph was added to describe the process of evaluation of clinical relevance using MedGemma.
2. In Results, section d) “Manual verification and LLM classification:”, a paragraph was added to summarize LLM findings.
3. In Discussion, section d) “Observations from Manual Validation and LLM analysis”, a discussion on the quantitative support to the claims about variability between systems in section a) of the discussion is added.

4. The RegEx mapping was based on CheXpert definitions. This may introduce bias toward CheXpert and CheXbert. The authors should validate the RegEx mapping. Manual verification on a subset would improve reliability.

We thank the reviewer for the opportunity to clarify this. The CheXpert taxonomy was selected because it represents the standard benchmarking framework for chest radiograph NLP. While CheXpert is rule-based, CheXbert is a transformer-based model trained on a combination of rule-derived labels and human radiologist ground truth to predict this exact same clinical taxonomy.

By utilizing an expanded version of the CheXpert phrase definitions to map the commercial NLP services' extractions to these identical categories, we are establishing a uniform "common ground." This ensures a completely fair baseline comparison across all systems against the same underlying clinical definitions. Far from introducing bias, applying a standardized mapping rule-set to the commercial vendors is the only way to evaluate them objectively against the

established CheXpert/CheXbert standard without artificially disadvantaging their unique proprietary schemas.

5. The manuscript is well written. However, minor grammar edits are needed. Some sentences in the Discussion are too long and should be shortened.

We appreciate the thoughtful feedback on the manuscript. We have addressed the grammar points and streamlined the Discussion section, ensuring the sentences are now more concise and easier to follow.

1. In Discussion, we adjusted the sentence structure to improve clarity.

Reviewer #2: Reviewer Report

Manuscript ID: JDIM-D-26-00614
Title: Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset

The authors present a large-scale comparative study involving 95,008 pediatric chest radiograph reports, evaluating four commercial (AWS, Azure, Google Cloud, SparkNLP) and two open-source (CheXpert, CheXbert) NLP models. The study highlights significant variability in entity extraction and assertion detection, particularly regarding diagnostic uncertainty in pediatric cohorts. While the dataset size is impressive, the lack of a ground truth reference and certain methodological assumptions necessitate a Major Revision before the manuscript can be considered for publication.

1. The current manuscript often implies clinical validation; however, without a gold standard (manual annotation by radiologists), the study can only measure consistency, not accuracy. If expert annotation for the entire dataset is not feasible, the authors must reframe the study as an Inter-Model Agreement and Consistency Analysis. Claims regarding performance or accuracy should be moderated to reflect "agreement" throughout the text.

Thank you for the insightful feedback. We sampled a subset of 360 exams stratified across the 12 CheXpert labels and across the 3 assertion statuses (Positive, Negative, and Uncertain). We created a custom UI and a pediatric radiologist (author on this paper) reviewed the impression

section of the reports and assigned the appropriate rating for each label. We then calculated precision, recall and F1-scores for each NLP service against this ground truth.

Changes:

1. In Introduction, edits were made to reflect "agreement" when comparing models to each other, and "performance/accuracy" only when comparing to the expert-labeled gold standard.
2. In Methods, section g) “Manual Validation and LLM analysis” was added to describe the manual review process and metrics calculation. Additionally, revised the terminology to distinguish model-to-model agreement from performance/accuracy relative to the expert-labeled gold standard.
4. In Results, section d) ”Manual Verification and LLM classification” an additional paragraph discussing the results comparing the NLP service performance to the manual ratings is added. The text was also revised to maintain consistent use of the updated terminology when describing model comparisons and comparisons with the expert-labeled gold standard.
5. Discussion (section d) and abstract are also appropriately updated with the results from manual validation.

2. The methodology relies on a custom RegEx algorithm to map extracted entities to 13 CheXpert disease categories. Any systematic error in this mapping could invalidate the comparative results between commercial services and the CheXpert/CheXbert benchmarks. Perform a technical validation on a random subset of at least 200 reports to ensure the RegEx algorithm correctly maps identified entities to the intended labels. Report the precision and recall of this mapping process.

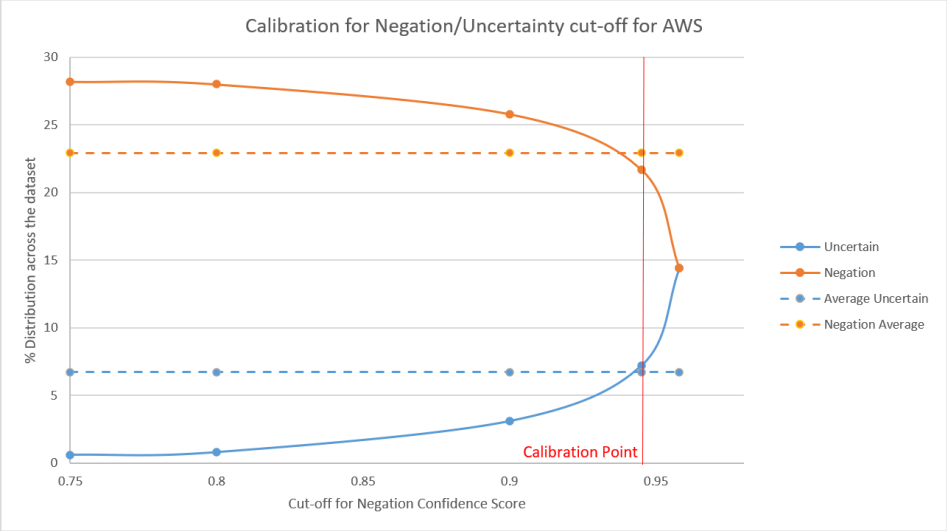
We appreciate the opportunity to clarify the design of our mapping pipeline. As noted in the Methods section, this deterministic algorithm was not built from scratch; rather, it was adapted directly from the validated, open-source phrase patterns of the Stanford CheXpert labeler codebase.Because the commercial systems extract highly disparate, raw text entities, this rule-set serves as a necessary "common ground" to map their extractions to a standardized taxonomy for a fair, downstream comparison. To ensure this mapping didn't artificially penalize the commercial systems due to localized vernacular, the dictionary was expanded to incorporate specialized pediatric descriptors and institutional reporting styles.

Any potential mapping errors are already fully captured and accounted for in the modest validation metrics (precision, recall, F1) we reported against the radiologist's manual annotations. We have refined the text in the Methods section (**section f**) to make this

architecture perfectly transparent and have added the complete phrase mapping dictionary to Appendix A for full reproducibility.

3. The study uses arbitrary thresholds (CS < 0.25 for positive and CS > 0.75 for negative) to categorize AWS negation confidence scores into assertion statuses. This likely explains the significantly lower uncertainty rate (1.4%) observed for AWS compared to other models like SparkNLP (32.6%). Provide a clinical or empirical justification for these specific thresholds. Conduct a sensitivity analysis showing how varying these thresholds (e.g., 0.20/0.80 or 0.30/0.70) affects the distribution of assertion categories for AWS.

To standardize the AWS output, a threshold-based calibration approach was developed using the negation confidence score (CS) distribution. Entities with a null negation CS were automatically classified as positive. To define the cutoff for the uncertain category, an optimization procedure was performed in which candidate thresholds were evaluated across a range of absolute values (0.75-0.90) as well as quantile-based markers (the 25th percentile [0.945] and 50th percentile [0.958] of the overall negation score distribution). AWS assertion distributions under each candidate threshold were then compared against the mean proportions of negation and uncertainty observed across the other three commercial models. The threshold that achieved the closest alignment with the multi-model average distribution was selected as the final CS cutoff for identifying uncertainty in AWS.



Changes:

1. In Methods, section d), an additional paragraph was added describing the threshold-based calibration approach for standardizing AWS output.
2. In Results, section b), a paragraph was added reflecting assertion distribution using threshold-based calibration for standardizing AWS output.

3. Figures 5 and 6 and tables 2, 3 and 4 were updated after standardizing the AWS output with the new threshold value.

4. Table 5 provides only 8 examples of discrepant cases. Given the vast dataset, a more systematic approach to qualitative failure analysis is required. Categorize the main drivers of disagreement (e.g., complex syntax, hedging, negation handling, or specific anatomical modifiers). Discuss how these patterns of disagreement might influence model selection for specific clinical research tasks.

Thank you for the suggestion. Given the large volume of disagreements, we identified and grouped representative example cases into three major recurring linguistic categories: Negation Scope, Hedging and Probability, and Complex Syntactic Structures. These categories collectively illustrate that clinical NLP performance often depends on subtle linguistic context that current models inconsistently resolve.

Changes:

1. In Methods, section f) "Mapping clinical entities to CheXpert disease categories and comparison", a qualitative review of discrepant cases was described.
2. In Results, section e) "Qualitative analysis of discrepant assertion statuses on CheXpert labels", included a paragraph to expand on the analysis of discrepant cases by grouping representative examples into the three major recurring linguistic categories. Additionally, we provide illustrative examples provided for each category.

5. The authors mention using a "lemmatization algorithm" for entity normalization. Specify the exact library and version used (e.g., spaCy, NLTK) to ensure the findings are reproducible by other researchers.

Thank you for this comment. We used Natural Language Toolkit (NLTK) version 3.14 for entity normalization.

Changes:

1. In Methods, section c) "Disease entity recognition and quantification", the lemmatization tool/library and version used for entity normalization, NLTK, was specified.

6. With a sample size of n = 95,008 almost all differences will achieve $p < 0.001$. Supplement the p-values with effect size measurements to clarify the clinical and practical significance of the reported differences.

Thank you for this important comment. We agree that, with a sample size of this magnitude, p-values alone are insufficient to convey practical significance. In response, we replaced the

original chi-square tests with a more appropriate exam-level generalized estimating equations (GEE) analysis, which accounts for within-exam correlation when comparing assertion distributions across NLP systems. We also supplemented statistical significance with effect-size summaries based on the observed assertion-status distributions, reported as section-specific percentages and absolute percentage-point differences for positive, negative, and uncertain assertions. In addition, pairwise post hoc comparisons between systems were performed within each section using the same GEE framework and adjusted using the Holm correction for multiple testing. Together, these revisions provide a more appropriate inferential framework and a clearer interpretation of the magnitude, not just the significance, of inter-system differences.

Changes:

1. In Methods, section d) “Standardization and analysis of assertion status”, a paragraph to describe methods to assess differences in assertion distributions between systems were described.
2. In Results, section b) “Assertion distribution in commercial clinical NLP models”, a paragraph to discuss the differences in assertion distribution.

Evaluating Clinical NLP Services for Chest Radiograph Report Labeling: A Comparative Study on an Independent Pediatric Dataset

Shruti Hegde MS[1*] (Shruti.Hegde@cchmc.org)

Mabon Manoj Ninan BS[2] (ninanmm@tamu.edu)

Jonathan R. Dillman MD, MSc[1] (Jonathan.Dillman@cchmc.org)

Shireen Hayatghaibi PhD[1] (Shireen.Hayatghaibi@cchmc.org)

Elanchezhian Somasundaram PhD[1] (Elanchezhian.Somasundaram@cchmc.org)

[1] Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States

[2] Texas A&M University, College Station, Texas, United States

* Corresponding author: Shruti Hegde MS (Shruti.Hegde@cchmc.org)

Statements and Declarations:

1. Consent to participate: Institutional Review Board Waived Informed Consent for this Retrospective study.
2. Consent to publish: Institutional Review Board Waived Informed Consent for this Retrospective study.
3. Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.
4. Funding: The project described was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health, under Award Number 2UL1TR001425-05A1.
5. Author Contributions: All authors contributed to the study conception and design. Material preparation, data cleaning and analysis were performed by Shruti Hegde, Mabon Manoj Ninan and Elanchezhian Somasundaram. Material review was done by Jonathan R. Dillman and Shireen Hayatghaibi. All authors read and approved the final manuscript.
6. Human Ethics and Consent to Participate: Human Ethics and Consent to Participate declarations: not applicable.

7. Data Availability declaration: Data are however available from the authors upon reasonable request.

Summary Statement

Commercial clinical NLP tools and specialized radiograph labelers show notable differences in entity extraction and assertion classification for pediatric chest radiograph reports, emphasizing the need for task-specific validation and manual review before clinical or research implementation.