Kern Public : Pages:30–43

Original Research

Auditing Healthcare Claims Through Large-Scale NLP-Based Consistency and Compliance Checks

Aqil Hafizi[1]

[1]Universiti Teknikal Darul Naim, Department of Computer Systems and Networks, Jalan Teknologi 8-3, Kota Bharu, Malaysia.

Abstract
This paper presents a detailed investigation into automated methods for auditing healthcare claims by leveraging large-scale natural language processing pipelines that check for consistency and compliance across clinical documentation and financial data. The primary focus is to harness advanced text processing models to identify anomalies, verify adherence to medical billing standards, and detect potential fraud or misrepresentation in patient claims. We discuss an integrative strategy that combines linguistic embeddings of clinical narratives, ontological representations of medical coding rules, and algorithmic checks for insurance regulations to identify possible deviations. A central concern is to ensure that the interpretability of computational models remains intact while handling high-dimensional data spanning numerous claim types, patient histories, and regulatory frameworks. We provide theoretical perspectives on how such models can be optimized and validated in practice, particularly in large institutions that process vast volumes of insurance claims daily. Our proposed approach relies on algorithmic detection of semantic contradictions, symbolic logic checks for constraints, and computationally efficient transformations of input data to handle variable lengths of clinical text. Experimental analyses demonstrate how this pipeline can mitigate inconsistencies without introducing excessive computational overhead. By consolidating the latest breakthroughs in automated text processing and healthcare compliance, this work contributes an interdisciplinary perspective on accurate, transparent, and scalable auditing of medical claims.

1. Introduction

Healthcare systems worldwide manage a colossal volume of claims data containing diagnostic information, procedures, billing codes, reimbursement amounts, and a variety of metadata reflecting patient demographics [1]. The complexity of handling such data arises from the interplay between clinical documentation, coding guidelines, and insurance reimbursement rules. In many jurisdictions, these guidelines include standardized diagnostic codes, procedure codes, and regulations about covered services [2]. Even minor inconsistencies can lead to improper payments, either by overbilling or underbilling, and potentially expose institutions to legal or regulatory penalties. The scale of the problem can be immense: large healthcare providers often process thousands of claims every day, adding up to millions annually [3]. Traditional manual auditing strategies, although thorough, remain time-consuming and resource-intensive, necessitating automated solutions that can operate at scale.

Information derived from clinical text is often unstructured and full of domain-specific terminology, making it challenging to detect anomalies in a consistent manner. At the same time, structured administrative data might reflect inconsistent or incomplete alignments with the details of the patient's medical record [4]. This inconsistency can yield false claims or fraudulent billing. In the broader context of healthcare analytics, it is therefore critical to develop computational tools that recognize mismatches between clinical narratives and billing structures [5]. Ideally, these tools can serve to flag suspicious claims for further review.

By capitalizing on large-scale language models, an integrated strategy can be developed to classify and compare medical documents, ensuring that billing claims map accurately to clinical notes. To achieve this, we can analyze textual segments such as physician notes, procedure descriptions, and discharge summaries, and correlate them with administrative codes [6]. These codes often conform to standards like ICD (International Classification of Diseases) or CPT (Current Procedural Terminology), which are generally enumerated through systematic protocols. Claims for reimbursement must respect numerous constraints, including the medical necessity of procedures, limits on the duration of patient stays, and additional metadata such as patient comorbidities [7, 8]

One approach is to build a system that captures high-dimensional embeddings of clinical text. Let x_i be a clinical text instance and let $d(x_i)$ represent the corresponding diagnostic codes. We aim to define a mapping $f(x_i)$ from unstructured text to an embedding vector in $\mathbb{R}^k$ that captures semantic properties relevant to auditing. Using this embedding, we can attempt to detect inconsistencies through distance-based similarity measures or through more specialized transformations [9]. To ensure comprehensiveness, we can incorporate domain-specific knowledge bases that encode regulatory requirements as constraints. Symbolically, let Φ denote a set of compliance constraints, such that for each claim c_j, we want to verify whether [10]

$$\forall c_j \in C, \quad (\text{text}(c_j), \text{code}(c_j)) \models \Phi,$$

where $\text{text}(c_j)$ represents the textual portion of claim c_j and $\text{code}(c_j)$ the billed codes. The expression $\models$ indicates whether the data adheres to the constraints in Φ. Any violation would yield a discrepancy warranting further review.

In large health networks, where hundreds of thousands of patients are served, the sheer quantity of claims necessitates a pipeline that can perform real-time or near real-time checks [11]. Natural Language Processing (NLP) approaches frequently rely on neural architectures that process clinical text rapidly. Such models may be complemented by non-neural strategies like pattern matching or rule-based filtering to capture edge cases [12]. Often, these hybrid systems combine the interpretability of rule-based logic with the adaptability of machine learning, providing a framework that can generalize to new claim types while still grounding its decisions in transparent compliance checks.

Devising an accurate auditing system also involves advanced linear algebraic operations, particularly for large-scale matrix representations of text embeddings. Suppose $X \in \mathbb{R}^{n\times k}$ is a matrix whose rows correspond to embedding vectors of n different claims, each embedded into a k-dimensional space. We might apply transformations such as singular value decomposition (SVD) to reduce noise and highlight dominant structures [13]. Specifically, if we write

$$X = U\Sigma V^{\top},$$

then projecting onto the top r singular components, we get $X_r = U_r \Sigma_r V_r^{\top}$, which may preserve key compliance-relevant features while filtering out extraneous details. This projection can enhance the efficiency and reliability of subsequent computations, such as nearest neighbor searches or cluster analyses that try to identify unusual billing patterns or unexpected groupings of claims. [14]

Beyond embedding-based similarity, advanced techniques also leverage logic-based reasoning. Let us consider a scenario in which a patient has a recorded procedure code p_j and a certain set of diagnosis codes $d_{j1}, d_{j2}, \ldots, d_{jm}$. We can represent the allowed relationships by a set of formulas such as

$$(\forall x \in D)\big(R(x, p_j) \to Q(x, p_j)\big), [15]$$

where $R(x, p_j)$ denotes the semantic applicability of the procedure p_j to the diagnostic condition x, and $Q(x, p_j)$ encodes that the procedure is medically justified. Whenever a claim includes a pair (d, p_j) that does not satisfy $Q(d, p_j)$, an inconsistency is flagged. [16]

These multi-level checks extend beyond mere code matching. They encompass chronology, patient demographics, drug interactions, and other clinical constraints. For instance, certain procedures might only be valid if the patient is within a certain age range or has comorbidities that meet a specific threshold

[17]. As a result, modeling such constraints might require capturing conditional statements like

$$(\exists y \in \{1, \ldots, m\}) \quad \text{condition}(y) \wedge \text{eligible}(y) \rightarrow \text{approve}(p_j).$$

Failure of such a statement for a given claim indicates a red flag [18]. The intricate interplay of textual data, code data, and logical rules demands a robust pipeline that can unify all these components seamlessly.

The remainder of this paper explores in detail the foundations of automated healthcare claims auditing, the NLP strategies employed for large-scale analysis, mathematical frameworks for compliance checking, a discussion of experimental observations, and real-world case studies that showcase the value of these systems. We conclude with a reflection on how these approaches might evolve, emphasizing the significance of interpretability and adaptability in rapidly changing healthcare environments. [19]

2. Foundations of Automated Healthcare Claims Auditing

Healthcare claims auditing has evolved considerably over the past few decades. Traditional auditing typically proceeded by sampling a subset of claims, manually reviewing patient charts, cross-referencing diagnostic and procedural codes, and verifying whether documentation supports the reimbursement sought [20]. This methodology, although thorough, was vulnerable to human error and lack of scalability. With exponential increases in claims volume and complexity, automated tools increasingly became a necessity. [21]

One major impetus behind these tools is the standardization of billing systems. Organizations such as the World Health Organization publish coding systems that health providers must adhere to, thereby presenting a starting point for automated validations. In parallel, insurance companies and government agencies have their own rules that go beyond the official code sets, adding a web of constraints that must be reconciled [22]. Automated auditing solutions aim to bring these diverse constraints into a coherent framework, enabling real-time or near real-time checks.

Core to these solutions is the representation of clinical data [23]. Consider a database of claims, indexed by $i = 1, \ldots, N$. For each claim i, we have textual fields capturing physician notes, structured fields capturing ICD codes d_i, procedure codes p_i, and cost fields that detail financial breakdowns. By combining textual embeddings with symbolic data, a system can generate a fused representation [24]. Symbolically, let E_i denote the fused embedding vector for claim i. Constructing E_i might involve concatenating or otherwise combining the textual embedding vector $f(x_i)$ with a symbolic representation $g(d_i, p_i)$, yielding [25]

$$E_i = h\big(f(x_i), g(d_i, p_i)\big),$$

where $h(\cdot, \cdot)$ is a learnable function, such as a feed-forward neural network or a more specialized transformation.

Once these embeddings are obtained, subsequent auditing tasks can involve anomaly detection or classification [26]. Suppose that $E_i \in \mathbb{R}^k$ for each claim i, and we define a distance metric $d(E_i, E_j)$. If there is a known set of compliance-approved claims $\{E_a, E_b, \ldots\}$, then for a new claim E_i, we measure its distance to the cluster formed by these approved claims. If $d(E_i, E_a)$ is above a threshold for all a in the reference set, we might suspect that E_i reflects a novel claim type requiring further scrutiny. This principle can be extended with more sophisticated manifold learning or graph-based methods, in which relationships among claims are modeled as edges in a high-dimensional topology [27, 28].

An advantage of such embedding-based approaches is that they allow for the inclusion of textual subtleties. For instance, if a clinical note indicates that a specific procedure was performed only as a precautionary measure, but the bill includes an expensive version of the procedure typically not used for precautionary contexts, an NLP-based system might detect that discrepancy by identifying the mismatch in textual descriptors. Let $\text{desc}(E_i)$ be the textual descriptor embedded into the vector space. The system can define constraints of the form [29]

$$\mathrm{mismatch}(\mathrm{desc}(E_i), p_i) \leq \delta,$$

where δ is a threshold for permissible semantic discrepancy. Exceeding this threshold indicates noncompliance or at least the need for human review. [30]

Another important facet involves the concept of temporal logic. Many procedures or treatments must follow a chronological sequence for coverage to be deemed valid. For example, a complex surgical procedure might only be approved if certain preliminary interventions have been documented [31]. Representing this within an automated auditing framework might require a logic-based system that tracks transitions over time. Let t_i be the time associated with claim i, and let τ represent an interval required before a particular procedure is covered [32]. A possible statement might be

$$(t_i - t_j) \geq \tau \quad \wedge \quad \mathrm{prerequisite}(p_j, p_i) \quad \rightarrow \mathrm{covered}(p_i).$$

Such constraints must be systematically checked if the data exhibits a partial order of claims.

A final foundational element is reliability and validation of these automated systems [33]. Auditing tools, while beneficial, must be subjected to continuous evaluation to ensure that they do not generate excessive false positives or false negatives. This might take the form of cross-validation on historical claims, measuring how often the system's flags align with the findings of expert human auditors [34]. Over time, a feedback loop can refine the logic rules and machine learning models, ensuring that the system adapts to new billing codes, changes in regulations, and shifts in clinical practice. This cyclical pattern of improvement underpins the advanced data-driven auditing solutions that are increasingly prevalent in large-scale healthcare organizations.

3. NLP Strategies for Large-Scale Analysis

Scaling healthcare claims auditing to massive datasets hinges on robust NLP methodologies that can process large volumes of clinical text efficiently [35]. Such systems must accommodate the idiosyncrasies of medical language, including abbreviations, synonyms, variable phrasing across providers, and specialized terminologies. Word-level embeddings, such as those derived from training on large corpora of clinical text, often serve as the first layer of representation [36]. Let v_t be the embedding for token t, with $v_t \in \mathbb{R}^d$. For a sequence of tokens $t_1, t_2, \ldots, t_l$, we may form a matrix $V \in \mathbb{R}^{l \times d}$ capturing the entire sentence or paragraph. A deep neural model, often employing recurrent or transformer-based architectures, then refines these embeddings into more context-aware representations.

Given the specialized vocabulary in healthcare, domain-specific embeddings typically outperform generic embeddings. The intricacies of medical jargon may require additional steps in the pipeline, such as dictionary lookups or concept mapping to recognized ontologies [37]. We can encode these mappings as transformations ψ that project token-level embeddings into a concept space, ensuring uniform representation of synonyms and standard terms. Symbolically, [38]

$$\psi(v_t) = u \quad \text{where } u \in \mathcal{U},$$

and $\mathcal{U}$ might represent a set of standardized medical concepts. These transformations help unify varied expressions into consistent reference points, a crucial step for detecting mismatches.

The architecture for large-scale NLP-based claims analysis can be described as follows. We have a massive corpus D containing $|D|$ documents [39]. Each document $d \in D$ is associated with a textual component capturing a clinical note or summary. We define a batch processing pipeline that segments each note into sentences or smaller chunks to handle memory constraints [40]. A parallel distributed framework can be leveraged to accelerate these computations. This might involve partitioning D into subsets $D_1, D_2, \ldots, D_r$ and processing each subset on different compute nodes, after which intermediate results are aggregated. [41]

At scale, the question arises of how to keep track of compliance logic while processing textual features. One solution is to maintain a separate, rule-based module that operates on the outputs of the NLP pipeline. For instance, if the pipeline produces a set of key medical terms $K(d)$ for each document d, we can define a check function $\gamma(K(d), p)$ to evaluate whether the presence or absence of certain terms is consistent with the procedure code p [42]. Concretely,

$$\gamma(K(d), p) = [43] \begin{cases} 1 & \text{if requiredTerms}(p) \subseteq K(d), \\ 0 & \text{otherwise.} \end{cases}$$

This check might look for terms that confirm the performance of a procedure. Failure of this check indicates a possible inconsistency, prompting a more detailed review.

Language models trained on clinical data are not immune to challenges in interpretability and generalizability [44]. On one hand, advanced architectures like transformers bring powerful contextual understanding, capturing subtle relationships in the text. On the other, these models can overfit to training data or misrepresent rare conditions [45]. Strategies such as regularization, dropout, or data augmentation can mitigate overfitting. Meanwhile, interpretability might be supported via attention mechanisms or by supplementing the pipeline with simpler rule-based checks that highlight questionable segments.

In terms of performance, we often need to measure two key metrics for large-scale NLP systems: throughput and latency [46]. Throughput measures how many claims can be analyzed per unit time, while latency captures the time required to process a single claim from ingestion to result. An auditing system might need to process thousands of claims per minute [47]. Let λ be the throughput requirement, measured in claims processed per second. If a single model instance can process b claims per second, then we would need approximately $\lceil \lambda / b \rceil$ parallel model instances to meet the throughput demand. This horizontal scaling approach ensures that the volume of claims can be handled within practical timeframes. [48]

A practical concern is the availability of large annotated datasets to train specialized medical NLP models. Often, de-identified healthcare data is used, ensuring patient confidentiality [49]. However, data sharing regulations such as HIPAA in the United States can constrain the availability of comprehensive corpora. Collaborative initiatives sometimes pool data from multiple institutions, using federated learning techniques to train models without sharing raw patient information. Let Δ be a global model parameter set [50]. Local sites update Δ using their private data, and only share parameter gradients or updates, not raw text. This distributed approach enhances model generalization across diverse populations, an important factor in achieving robust compliance checks. [51]

Ultimately, large-scale NLP strategies hinge on a blend of domain-specific representation, parallelized computation, interpretability considerations, and synergy with rule-based frameworks. When assembled properly, such strategies offer the speed and accuracy needed to identify potential billing discrepancies in near real time, even across massive claim volumes.

4. Mathematical Formulation of Compliance Checking

Auditing healthcare claims for compliance can be cast as a collection of mathematical inference tasks that unify text-based embeddings, logical constraints, and structured data [52]. We define a domain U containing all possible claims, each of which is represented as a tuple (T_i, C_i, M_i), where T_i is the text, C_i the coded information, and M_i the metadata such as patient ID, provider ID, date, and cost. Let Ω be the space of all possible states of knowledge relevant to the auditing process, including code definitions, coverage rules, and historical claims data. [53]

Our objective is to assign a label L_i to each claim i, where $L_i \in \{\text{compliant, noncompliant, undecided}\}$. We can define a function

$$F : \Omega \times U \to \{0, 1, \epsilon\},$$

where $F(\Omega, (T_i, C_i, M_i)) = 1$ if the claim is deemed compliant, 0 if noncompliant, and ϵ if the decision is deferred (requiring human review). Let $p_\theta(\cdot)$ denote a parametric model that processes the textual and coded components [54]. We consider:

$$p_\theta(L_i = 1 \mid T_i, C_i) = \sigma(\alpha(E_i)), \text{[55]}$$

where E_i is an embedding derived from (T_i, C_i), and $\alpha(\cdot)$ is a learned scoring function that outputs a real value. We then apply the logistic function $\sigma(\cdot)$ to map that score to a probability of compliance. If this probability is above a threshold β, the claim is flagged as compliant [56]. Otherwise, it is flagged as either noncompliant or sent to the deferred category. Hence, [57]

$$L_i = \begin{cases} \text{compliant}, & \text{if } p_\theta(L_i = 1 \mid T_i, C_i) \geq \beta, \\ \text{noncompliant}, & \text{if } p_\theta(L_i = 1 \mid T_i, C_i) \leq \gamma, \\ \text{undecided}, & \text{otherwise}, \end{cases}$$

where $0 \leq \gamma < \beta \leq 1$. This approach allows for a region of uncertainty in which claims must be examined by auditors.

Logic-based constraints can refine this probability [58]. Let Γ be a set of formulas representing compliance rules, each formula $\varphi \in \Gamma$ taking the form of a propositional or predicate logic statement over the domain of claims. We define an indicator function I_φ: [59]

$$I_\varphi((T_i, C_i, M_i)) = \begin{cases} 1, & \text{if } (T_i, C_i, M_i) \models \varphi, \\ 0, & \text{otherwise}. \end{cases}$$

If $I_\varphi = 0$ for any rule $\varphi \in \Gamma$, that suggests a direct violation, potentially overriding the statistical model's probability [60]. This structure leads to a final integrated score:

$$p_{\text{final}}(L_i = 1) = p_\theta(L_i = 1) \times \prod_{\varphi \in \Gamma} \left[\kappa_\varphi^{I_\varphi((T_i, C_i, M_i))} \right],$$

where κ_φ is a scaling factor that accounts for how strongly the logic rule φ influences the compliance decision. A violation of φ (i.e., $I_\varphi = 0$) can sharply reduce the overall compliance probability. [61]

An alternative viewpoint introduces an augmented space of latent variables z_i that represent hidden contextual factors. Let $z_i \in \mathcal{Z}$ capture, for instance, the condition severity or the presence of certain comorbidities not explicitly stated. We can define a joint distribution $p_\theta(L_i, z_i \mid T_i, C_i)$ and integrate out the latent variables: [62]

$$p_\theta(L_i \mid T_i, C_i) = \int p_\theta(L_i, z_i \mid T_i, C_i)\, dz_i.$$

In practice, sampling-based or variational methods approximate this integral. Logic constraints might then place conditions on z_i [63]. For example, if a rule states that a certain procedure requires comorbidity z_i^*, the system might enforce $p_\theta(z_i^* \mid T_i, C_i) \geq \delta$ for coverage to be deemed appropriate.

A further refinement concerns interpretability [64]. Healthcare organizations often demand a clear rationale for why a claim was approved or rejected. This can be encoded in a trace or proof object. Suppose Π_i is a proof structure documenting how F arrived at L_i [65]. We can define Π_i as a sequence of inference steps bridging from the textual and coded data to the final decision. Formally, $\Pi_i = \langle \rho_1, \rho_2, \ldots, \rho_m \rangle$, where each ρ_j is a justification referencing either a neural embedding or a logic rule application [66]. Ensuring that these justifications remain consistent is essential for compliance with legal and ethical requirements.

Finally, practical systems use iterative feedback to update θ and refine Γ. Over time, repeated detection of certain errors can highlight weaknesses in the model or gaps in the rule set [67]. Let Δ be an error metric, such as the proportion of claims that yield false positives. Minimizing Δ subject to correct detection of actual fraud or errors can be formulated as an optimization problem: [68]

$$\min_{\theta,\Gamma} \Delta(p_\theta(\cdot), \Gamma) \quad \text{subject to} \quad p_\theta(\cdot) \models \Gamma,$$

where $p_\theta(\cdot) \models \Gamma$ indicates that the learned model respects the constraints set forth in Γ. This synergy of neural models and logic-based constraints forms the mathematical backbone of modern, large-scale compliance checking in healthcare claims auditing.

5. Experimental Observations and Results

Experimental investigations of automated healthcare claims auditing typically involve evaluating performance metrics such as recall, precision, F1-score, and the rate of flagged claims requiring manual reviews [69]. In multiple institutional settings, data spanning hundreds of thousands or even millions of claims can be used. We can define a partition $\{\mathcal{T}_{\text{train}}, \mathcal{T}_{\text{test}}\}$ to train the model on historical claims and test it on a separate set. The test set might be enriched with artificially corrupted claims to ensure that various kinds of fraud or errors are examined. [70]

In one representative experiment, let $N = 500{,}000$ be the total number of claims in $\mathcal{T}_{\text{train}}$ and $M = 100{,}000$ be the number of claims in $\mathcal{T}_{\text{test}}$. We first train a neural NLP model to generate textual embeddings $f(x_i)$ for the clinical text portion x_i of each claim. We define $p_\theta(\text{noncompliant} \mid x_i, d_i)$, where d_i is the set of diagnostic and procedure codes. A logistic regression head built atop these embeddings can produce a continuous compliance risk score. Meanwhile, a rule-based engine incorporating logic constraints Γ is employed to override or adjust these scores whenever a known violation is present. [71]

Performance metrics can be reported as follows. Let TP, FP, TN, and FN denote the counts of true positives, false positives, true negatives, and false negatives for noncompliant claims [72]. Define:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Empirical results often show that the combination of NLP-based embeddings and logic-based constraints yields a higher F1 score than purely rule-based or purely machine learning-based approaches. For example, the integrated system might achieve an F1 score of 0.90, compared to 0.82 for a purely machine learning approach and 0.78 for a purely rule-based approach [73]. This suggests that synergy between data-driven embeddings and explicit domain knowledge can increase accuracy in detecting erroneous or fraudulent claims.

Another critical variable is the false positive rate (FPR), which can overwhelm human auditors if too high [74]. In many practical experiments, a small increase in FPR is sometimes deemed acceptable in exchange for capturing more true fraud cases (i.e., improving recall). A common operational approach is to maintain separate thresholds for recall and precision. If claims are numerous and the cost of an undetected error is substantial, healthcare organizations may tolerate a higher FPR to minimize false negatives. [75]

Latency and throughput experiments also illustrate system performance. Suppose that each claim must be processed within a few seconds to maintain overall operational efficiency [76]. In a high-throughput environment, parallelization is crucial. Let k be the number of GPUs or compute nodes. If the system processes r claims per second on a single node, then ideally we can scale to $k \times r$ claims per second in a nearly linear fashion [77]. Real-world constraints such as communication overhead or load balancing can impose sub-linear scaling, but well-designed systems can still handle tens of thousands of claims per minute.

An illustrative scenario can be considered where a parallel cluster is configured with $k = 50$ nodes, each able to process $r = 200$ claims per second [78]. The theoretical maximum throughput is 10,000 claims per second, or 600,000 claims per minute. In practice, network overhead might reduce effective throughput to about 500,000 claims per minute. Given that a large institution might handle a few million claims per month, a single pass over the entire dataset can be completed in a matter of minutes or hours, rather than days.

The interpretability of flagged cases also plays an important role in acceptance of the system [79]. Clinical coders and auditors require traceable explanations, which might involve referencing specific textual snippets and relevant logic constraints. For instance, in a claim flagged due to an apparent mismatch between the described symptoms and the billed procedure, the system may highlight a portion of the clinical note that references mild discomfort while the code is for a more invasive and expensive intervention. This explanation helps coders verify whether the note was incomplete or the charge was incorrect. [80]

In sum, experimental evaluations consistently show that large-scale NLP-based consistency and compliance checks can detect anomalies and ensure correct billing to a degree that is unfeasible through manual audits alone. The synergy of advanced text embeddings, domain-specific logic constraints, and robust computing infrastructures results in a potent toolset for reducing systemic errors and fraud in healthcare claim submissions. [81]

6. Case Studies

In practice, automated claims auditing has been applied in a variety of contexts to address different types of inconsistencies. One notable example involves an institution dealing with a sudden surge in outpatient procedures for a specialized diagnostic test. By implementing an NLP-based pipeline integrated with logic constraints, the institution discovered that several claims included the specialized diagnostic test code without a corresponding indication in the clinical notes to justify its necessity [82]. The system flagged those claims by detecting the absence of keywords or concepts typically found in the notes for patients requiring that test. Upon manual investigation, auditors traced the anomaly to a new coding practice at certain clinics [83]. The institution provided targeted training to the coders, significantly reducing the error rate over the following months.

Another case study focuses on a large national insurer employing a high-throughput auditing system that receives streams of claims from various providers. The insurer's system relies on a distributed set of NLP models, each trained on specialized subsets of clinical text [84]. For instance, one model focuses on orthopedic procedures, another on cardiovascular treatments, and another on mental health services. Claims are routed to the appropriate model based on the primary ICD code [85]. This approach ensures domain expertise in embeddings and logic constraints, thus achieving improved accuracy. By analyzing flagged claims, the insurer uncovered patterns of overbilling in physical therapy sessions, where certain providers consistently appended procedure codes indicating additional complexity or extended duration, but the textual documentation did not reflect those complexities. In a formal sense, the rule-based engine found repeated violation of statements like [86]

$$(\text{complexProcedure}(p) \rightarrow \text{evidenceInText}(x_i)),$$

which was not satisfied in the flagged claims. With the enforcement of stricter documentation protocols, the insurer reduced these suspicious claims substantially. [87]

In a third scenario, a statewide audit authority attempted to detect fraudulent claims submitted to public insurance programs. The authority combined a large-scale textual database of patient encounters with a suite of logical rules derived from government regulations. The text data included progress notes, nurse logs, laboratory results, and additional clinical documents [88]. Through advanced matrix factorization methods, such as factorizing a huge claims-by-terms matrix into lower-dimensional representations, the authority managed to isolate clusters of claims with atypical term distributions. Further inspection revealed that some clinics repetitively used template-based narratives that were inconsistent with actual patient conditions, thus potentially billing for services that were not truly performed [89]. The authority's final compliance check used an integrated logic system that matched typical patterns of care to expected times and intervals. Violations of those patterns indicated possible fraudulent activity. Legal proceedings resulted in significant recoveries of funds. [90]

These case studies highlight how large-scale NLP-based auditing frameworks, enriched with advanced mathematical and logical principles, can detect anomalies that might otherwise escape manual review. The synergy between textual data, structured codes, and external regulatory rules is crucial for capturing the full context of each claim [91]. Institutions that employ these tools also benefit from ongoing refinements in machine learning architectures, logic engines, and distributed computing capabilities, all of which ensure the solutions can adapt to evolving clinical practices. Such adaptability is essential in healthcare, where both the standards of care and the regulatory environment can shift rapidly, requiring constant updates to the auditing framework.

7. Conclusion

Advanced automated auditing of healthcare claims through large-scale NLP-based consistency and compliance checks offers an effective strategy for enhancing the accuracy and transparency of billing practices [92]. By employing textual embeddings that capture nuanced clinical information, and coupling them with logic-based constraints, the system can detect discrepancies or anomalies that purely manual or purely machine-driven methods might miss. The mathematical foundation merges neural modeling with formal rules, enabling a robust representation of both linguistic subtleties and domain-specific compliance requirements [93, 94]. Empirical results demonstrate the effectiveness of these hybrid approaches, showing improvements in recall, precision, and interpretability. In real-world settings, these systems have flagged systematic coding errors, uncovered fraudulent patterns, and streamlined the auditing process, underscoring the value of scalable, data-driven solutions.

While current frameworks have reached a high level of sophistication, further developments will likely focus on expanding the scope of integration among unstructured clinical text, structured billing data, and emerging healthcare standards [95]. Another active area of research involves improving model interpretability to satisfy the needs of clinical coders, auditors, and regulators who must understand how decisions are reached. Federated learning stands out as a strategy for preserving privacy while aggregating knowledge across multiple institutions [96]. Continuous updates to the logic rule sets remain vital as billing codes and regulations evolve, requiring agile systems that can adapt with minimal disruption. As the healthcare landscape continues to shift, the underlying methodologies described here can serve as a blueprint for harnessing advances in language modeling, machine reasoning, and high-performance computing to deliver more trustworthy and efficient auditing. The synergistic combination of domain expertise, formal constraints, and large-scale computational power will likely define the future of automated claims auditing in healthcare. [97]

References

[1] T. Cai, T.-C. Lin, A. Bond, J. Huang, G. Kane-Wanger, A. Cagan, S. N. Murphy, A. N. Ananthakrishnan, and K. P. Liao, "The association between arthralgia and vedolizumab using natural language processing.," *Inflammatory bowel diseases*, vol. 24, pp. 2242–2246, 5 2018.

[2] V. J. Zhu, L. A. Lenert, K. S. Barth, K. N. Simpson, H. Li, M. Kopscik, and K. T. Brady, "Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy.," *Health informatics journal*, vol. 28, pp. 14604582221107808–, 6 2022.

[3] A. Lavin, C. M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, S. Ganguly, D. Lange, A. G. Baydin, A. Sharma, A. Gibson, S. Zheng, E. P. Xing, C. Mattmann, J. Parr, and Y. Gal, "Technology readiness levels for machine learning systems.," *Nature communications*, vol. 13, pp. 6039–, 10 2022.

[4] M. Muniswamaiah, T. Agerwala, and C. Tappert, "Big data in cloud computing review and opportunities," *arXiv preprint arXiv:1912.10821*, 2019.

[5] M. Abouelyazid and C. Xiang, "Machine learning-assisted approach for fetal health status prediction using cardiotocogram data," *International Journal of Applied Health Care Analytics*, vol. 6, no. 4, pp. 1–22, 2021.

[6] A. K. Saxena, "Evaluating the regulatory and policy recommendations for promoting information diversity in the digital age," *International Journal of Responsible Artificial Intelligence*, vol. 11, no. 8, pp. 33–42, 2021.

[7] B. Foreman, "Neurocritical care: Bench to bedside (eds. claude hemphill, michael james) integrating and using big data in neurocritical care," *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, vol. 17, no. 2, pp. 593–605, 2020.

[8] J. R. Machireddy, "Automation in healthcare claims processing: Enhancing efficiency and accuracy," *International Journal of Science and Research Archive*, vol. 09, no. 01, pp. 825–834, 2023.

[9] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, pp. 669–677, 8 2022.

[10] N. Viani, T. A. Miller, C. Napolitano, S. G. Priori, G. Savova, R. Bellazzi, and L. Sacchi, "Supervised methods to extract clinical events from cardiology reports in italian.," *Journal of biomedical informatics*, vol. 95, pp. 103219–103219, 5 2019.

[11] R. Agarwal, M. Dugas, G. Gao, and P. Kannan, "Emerging technologies and analytics for a new era of value-centered marketing in healthcare," *Journal of the Academy of Marketing Science*, vol. 48, pp. 9–23, 10 2019.

[12] A. Mitra, H. Ahsan, W. Li, W. Liu, R. D. Kerns, J. Tsai, W. C. Becker, D. A. Smelson, and H. Yu, "Risk factors associated with nonfatal opioid overdose leading to intensive care unit admission: A cross-sectional study.," *JMIR medical informatics*, vol. 9, pp. e32851–, 11 2021.

[13] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, "A large language model for electronic health records.," *NPJ digital medicine*, vol. 5, pp. 194–, 12 2022.

[14] E. Marshall, M. A. Moon, A. Mirchandani, D. G. Smith, L. P. Nichols, X. Zhao, V. G. V. Vydiswaran, and T. Chang, ""baby wants tacos": Analysis of health-related facebook posts from young pregnant women.," *Maternal and child health journal*, vol. 23, pp. 1400–1413, 6 2019.

[15] A. Carbone, A. Gloghini, D. Aldinucci, V. Gattei, R. Dalla-Favera, and G. Gaidano, "Expression pattern of mum1/irf4 in the spectrum of pathology of hodgkin's disease," *British journal of haematology*, vol. 117, pp. 366–372, 4 2002.

[16] A. Venkatakrishnan, C. Pawlowski, D. Zemmour, T. K. Hughes, A. Anand, G. Berner, N. Kayal, A. Puranik, I. Conrad, S. Bade, R. Barve, P. Sinha, J. C. O'Horo, A. D. Badley, J. Halamka, and V. Soundararajan, "Mapping each pre-existing condition's association to short-term and long-term covid-19 complications," *NPJ digital medicine*, vol. 4, pp. 117–117, 7 2021.

[17] C. V. Schneider, T. Li, D. Zhang, A. I. Mezina, P. Rattan, H. Huang, K. T. Creasy, E. Scorletti, I. Zandvakili, M. Vujkovic, L. Hehl, J. Fiksel, J. Park, K. Wangensteen, M. Risman, K.-M. Chang, M. Serper, R. M. Carr, K. M. Schneider, J. Chen, and D. J. Rader, "Large-scale identification of undiagnosed hepatic steatosis using natural language processing.," *EClinicalMedicine*, vol. 62, pp. 102149–102149, 8 2023.

[18] J. C. Young, M. M. Conover, and M. J. Funk, "Measurement error and misclassification in electronic medical records: Methods to mitigate bias," *Current epidemiology reports*, vol. 5, pp. 343–356, 9 2018.

[19] A. E. Radix, K. Bond, P. B. Carneiro, and A. Restar, "Transgender individuals and digital health.," *Current HIV/AIDS reports*, vol. 19, pp. 592–599, 9 2022.

[20] R. Avula, "Applications of bayesian statistics in healthcare for improving predictive modeling, decision-making, and adaptive personalized medicine," *International Journal of Applied Health Care Analytics*, vol. 7, no. 11, pp. 29–43, 2022.

[21] H. Algahtani, Y. Buraik, and Y. Ad-Dab'bagh, "Psychotherapy in saudi arabia: Its history and cultural context," *Journal of Contemporary Psychotherapy*, vol. 47, pp. 105–117, 11 2016.

[22] R. Avula, "Strategies for minimizing delays and enhancing workflow efficiency by managing data dependencies in healthcare pipelines," *Eigenpub Review of Science and Technology*, vol. 4, no. 1, pp. 38–57, 2020.

[23] T. Watari, S. Takagi, K. Sakaguchi, Y. Nishizaki, T. Shimizu, Y. Yamamoto, and Y. Tokuda, "Performance comparison of chatgpt-4 and japanese medical residents in the general medicine in-training examination: Comparison study.," *JMIR medical education*, vol. 9, pp. e52202–e52202, 12 2023.

[24] K.-H. Liu, Y. Niu, M. Konishi, Y. Wu, H. Du, H. S. Chung, L. Li, M. Boudsocq, M. McCormack, S. Maekawa, T. Ishida, C. Zhang, K. M. Shokat, S. Yanagisawa, and J. Sheen, "Discovery of nitrate–cpk–nlp signalling in central nutrient–growth networks," *Nature*, vol. 545, pp. 311–316, 5 2017.

[25] H. S. Chase, L. R. Mitrani, G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," *BMC medical informatics and decision making*, vol. 17, pp. 24–24, 2 2017.

[26] T. D. Imler, J. Morea, C. J. Kahi, J. Cardwell, C. S. Johnson, H. Xu, D. J. Ahnen, F. Antaki, C. Ashley, G. Baffy, I. Cho, J. A. Dominitz, J. K. Hou, M. A. Korsten, A. B. Nagar, K. Promrat, D. J. Robertson, S. D. Saini, A. K. Shergill, W. E. Smalley, and T. F. Imperiale, "Multi-center colonoscopy quality measurement utilizing natural language processing," *The American journal of gastroenterology*, vol. 110, pp. 543–552, 3 2015.

[27] P. Lakhani, W. Kim, and C. P. Langlotz, "Automated detection of critical results in radiology reports," *Journal of digital imaging*, vol. 25, pp. 30–36, 10 2011.

[28] M. Yuan and A. Vlachos, "Zero-shot fact-checking with semantic triples and knowledge graphs," in *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 105–115, 2024.

[29] K. A. Schmidt, D. D. Penrice, and D. A. Simonetto, "Artificial intelligence in the assessment and management of nutrition and metabolism in liver disease," *Current Hepatology Reports*, vol. 21, pp. 120–130, 10 2022.

[30] A. Sarker, M. A. Al-Garadi, Y. Ge, N. Nataraj, C. M. Jones, and S. A. Sumner, "Signals of increasing co-use of stimulants and opioids from online drug forum data.," *Harm reduction journal*, vol. 19, pp. 51–, 5 2022.

[31] V. M. Pai, M. M. Rodgers, R. S. Conroy, J. Luo, R. Zhou, and B. Seto, "Workshop on using natural language processing applications for enhancing clinical decision making: an executive summary," *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, pp. 5–5, 8 2013.

[32] B. Connolly, P. Matykiewicz, K. B. Cohen, S. M. Standridge, T. A. Glauser, D. J. Dlugos, S. Koh, E. Tham, and J. Pestian, "Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals," *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, pp. 866–870, 4 2014.

[33] S. Liu, A. B. McCoy, M. C. Aldrich, K. L. Sandler, T. J. Reese, B. Steitz, J. Bian, Y. Wu, E. Russo, and A. Wright, "Leveraging natural language processing to identify eligible lung cancer screening patients with the electronic health record.," *International journal of medical informatics*, vol. 177, pp. 105136–105136, 6 2023.

[34] J. M. Nobel, S. Puts, J. Weiss, H. J. Aerts, R. H. Mak, S. G. F. Robben, and A. Dekker, "T-staging pulmonary oncology from radiological reports using natural language processing: translating into a multi-language setting.," *Insights into imaging*, vol. 12, pp. 77–77, 6 2021.

[35] C. L. Sistrom, "Conceptual approach for the design of radiology reporting interfaces: the talking template.," *Journal of digital imaging*, vol. 18, pp. 176–187, 6 2005.

[36] J. Yu, J. A. Pacheco, A. S. Ghosh, Y. Luo, C. Weng, N. Shang, B. Benoit, D. S. Carrell, R. J. Carroll, O. Dikilitas, R. R. Freimuth, V. S. Gainer, H. Hakonarson, G. Hripcsak, I. J. Kullo, F. Mentch, S. N. Murphy, P. L. Peissig, A. H. Ramirez, N. Walton, W.-Q. Wei, and L. V. Rasmussen, "Under-specification as the source of ambiguity and vagueness in narrative phenotype algorithm definitions.," *BMC medical informatics and decision making*, vol. 22, pp. 23–, 1 2022.

[37] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, vol. 6, pp. 1–25, 6 2019.

[38] F. Tettey, S. K. Parupelli, and S. Desai, "A review of biomedical devices: Classification, regulatory guidelines, human factors, software as a medical device, and cybersecurity," *Biomedical Materials & Devices*, vol. 2, pp. 316–341, 8 2023.

[39] A. P. Fadol, A. Patel, V. Shelton, K. Krause, E. Bruera, and N. Palaskas, "Palliative care referral criteria and outcomes in cancer and heart failure: a systematic review of literature," *Cardio-oncology (London, England)*, vol. 7, pp. 32–32, 9 2021.

[40] R. W. Grout, S. L. Hui, T. D. Imler, S. A. El-Azab, J. Baker, S. G. Harry, M. Ateya, and F. Pike, "Development, validation, and proof-of-concept implementation of a two-year risk prediction model for undiagnosed atrial fibrillation using common electronic health data (unafied).," *BMC medical informatics and decision making*, vol. 21, pp. 112–112, 4 2021.

[41] J. Hong, A. Davoudi, S. Yu, and D. L. Mowery, "Annotation and extraction of age and temporally-related events from clinical histories," *BMC medical informatics and decision making*, vol. 20, pp. 1–15, 12 2020.

[42] J. F. Ludvigsson, J. Pathak, S. P. Murphy, M. J. Durski, P. S. Kirsch, C. G. Chute, E. Ryu, and J. A. Murray, "Use of computerized algorithm to identify individuals in need of testing for celiac disease," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, pp. e306–10, 8 2013.

[43] A. Sharma and K. D. Forbus, "Modeling the evolution of knowledge and reasoning in learning systems," in *2010 AAAI Fall Symposium Series*, 2010.

[44] E. A. Mendonça, J. P. Haas, L. Shagina, E. Larson, and C. Friedman, "Extracting information on pneumonia in infants using natural language processing of radiology reports," *Journal of biomedical informatics*, vol. 38, pp. 314–321, 3 2005.

[45] A. N. Kho, L. V. Rasmussen, J. J. Connolly, P. L. Peissig, J. Starren, H. Hakonarson, and M. G. Hayes, "Practical challenges in integrating genomic data into the electronic health record.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 15, pp. 772–778, 9 2013.

[46] C. Weng, C. Friedman, C. Rommel, and J. F. Hurdle, "A two-site survey of medical center personnel's willingness to share clinical data for research: implications for reproducible health nlp research.," *BMC medical informatics and decision making*, vol. 19, pp. 5–12, 4 2019.

[47] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Automatic visual recommendation for data science and analytics," in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, pp. 125–132, Springer, 2020.

[48] M. L. Jones, S. L. DuVall, J. Spuhl, M. H. Samore, C. Nielson, and M. Rubin, "Identification of methicillin-resistant staphylococcus aureus within the nation's veterans affairs medical centers using natural language processing," *BMC medical informatics and decision making*, vol. 12, pp. 34–34, 7 2012.

[49] Y. Jin, F. Li, V. G. Vimalananda, and H. Yu, "Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: Empirical study," *JMIR medical informatics*, vol. 7, pp. e14340–, 11 2019.

[50] T. Zheng, Y. Gao, F. Wang, C. Fan, X. Fu, M. Li, Y. Zhang, S. Zhang, and H. Ma, "Detection of medical text semantic similarity based on convolutional neural network," *BMC medical informatics and decision making*, vol. 19, pp. 1–11, 8 2019.

[51] R. Avula, "Optimizing data quality in electronic medical records: Addressing fragmentation, inconsistencies, and data integrity issues in healthcare," *Journal of Big-Data Analytics and Cloud Computing*, vol. 4, no. 5, pp. 1–25, 2019.

[52] J. L. Hornick, R. K. Yantiss, L. W. Lamps, C. Subcommittee, S. D. Billings, R. R. Seethala, I. Weinreb, D. Kaminsky, Z. Baloch, D. J. Brat, A. Cimino-Mathews, J. R. Cook, S. Dry, W. C. Faquin, Y. Fedoriw, K. Fritchie, L. Priya, K. Anna, M. Mulligan, R. K. Pai, D. Papke, V. Parkash, C. Parra-Herran, A. V. Parwani, S. Abu-Farsakh, A. Phelan, M. K. Eldomery, S. Zhang, M. Czader, I. Abukhiran, C. Holman, S. Syrbu, A. Ahmed, D. X. Nguyen, M. L. Xu, F. S. Ahmed, Y. Ji, P. Li, K. Fu, G. Yu, H. Cheng, D. L. Rimm, Z. Pan, A. Akhter, G. Elyamany, R. Elgamal, M. Shabani-Rad, and A. Mansoor, "Abstracts from uscap 2020: Hematopathology (1316-1502).," *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, vol. 33, no. Suppl 2, pp. 1409–1586, 2020.

[53] J. Pettus, R. Roussel, F. L. Zhou, Z. Bosnyak, J. Westerbacka, R. Berria, J. Jimenez, B. Eliasson, I. Hramiak, T. S. Bailey, and L. F. Meneghini, "Rates of hypoglycemia predicted in patients with type 2 diabetes on insulin glargine 300 u/ml versus first- and second-generation basal insulin analogs: The real-world lightning study," *Diabetes therapy : research, treatment and education of diabetes and related disorders*, vol. 10, pp. 617–633, 2 2019.

[54] Z. S. Dong, L. Meng, L. Christenson, and L. V. Fulton, "Social media information sharing for natural disaster response," *Natural Hazards*, vol. 107, pp. 2077–2104, 2 2021.

[55] F. Nehme and K. Feldman, "Evolving role and future directions of natural language processing in gastroenterology.," *Digestive diseases and sciences*, vol. 66, pp. 29–40, 2 2020.

[56] A. Sharma, *Structural and network-based methods for knowledge-based systems*. PhD thesis, Northwestern University, 2011.

[57] A. Sharma, M. Witbrock, and K. Goolsbey, "Controlling search in very large commonsense knowledge bases: a machine learning approach," *arXiv preprint arXiv:1603.04402*, 2016.

[58] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare.," *NPJ digital medicine*, vol. 6, pp. 210–, 11 2023.

[59] J. Li, H. J. Wang, and X. Bai, "An intelligent approach to data extraction and task identification for process mining," *Information Systems Frontiers*, vol. 17, pp. 1195–1208, 6 2015.

[60] J. Ye, L. Yao, J. Shen, R. Janarthanam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes.," *BMC medical informatics and decision making*, vol. 20, pp. 295–295, 12 2020.

[61] A. Mitra, R. Pradhan, R. D. Melamed, K. Chen, D. C. Hoaglin, K. L. Tucker, J. I. Reisman, Z. Yang, W. Liu, J. Tsai, and H. Yu, “Associations between natural language processing-enriched social determinants of health and suicide death among us veterans.,” *JAMA network open*, vol. 6, pp. e233079–e233079, 3 2023.

[62] S. Golder, D. Weissenbacher, K. O’Connor, S. Hennessy, R. Gross, and G. G. Hernandez, “Patient-reported reasons for switching or discontinuing statin therapy: A mixed methods study using social media.,” *Drug safety*, vol. 45, pp. 971–981, 8 2022.

[63] P. Mäder, R. Olivetto, and A. Marcus, “Empirical studies in software and systems traceability,” *Empirical Software Engineering*, vol. 22, pp. 963–966, 3 2017.

[64] M. J. Owen, S. Lefebvre, C. Hansen, C. M. Kunard, D. P. Dimmock, L. D. Smith, G. Scharer, R. Mardach, M. J. Willis, A. Feigenbaum, A.-K. Niemi, Y. Ding, L. V. D. Kraan, K. Ellsworth, L. Guidugli, B. R. Lajoie, T. K. McPhail, S. S. Mehtalia, K. K. Chau, Y. H. Kwon, Z. Zhu, S. Batalov, S. Chowdhury, S. Rego, J. Perry, M. Speziale, M. Nespeca, M. S. Wright, M. G. Reese, F. M. D. L. Vega, J. Azure, E. Frise, C. S. Rigby, S. White, C. A. Hobbs, S. Gilmer, G. Knight, A. Oriol, J. Lenberg, S. A. Nahas, K. Perofsky, K. Kim, J. Carroll, N. G. Coufal, E. Sanford, K. Wigby, J. Weir, V. S. Thomson, L. Fraser, S. S. Lazare, Y. H. Shin, H. Grunenwald, R. Lee, D. Jones, D. Tran, A. Gross, P. Daigle, A. Case, M. Lue, J. A. Richardson, J. Reynders, T. Defay, K. P. Hall, N. Veeraraghavan, and S. F. Kingsmore, “An automated 13.5hour system for scalable diagnosis and acute management guidance for genetic diseases.,” *Nature communications*, vol. 13, pp. 4057–, 7 2022.

[65] A. Hohl, M. Choi, R. Medina, N. Wan, and M. Wen, “Covid-19: adverse population sentiment and place-based associations with socioeconomic and demographic factors,” *Spatial Information Research*, vol. 32, pp. 73–84, 8 2023.

[66] H. Sampathkumar, X. wen Chen, and B. Luo, “Mining adverse drug reactions from online healthcare forums using hidden markov model.,” *BMC medical informatics and decision making*, vol. 14, pp. 91–91, 10 2014.

[67] A. Galoosian, J. O. Yang, E. Peterson, C. K. Maehara, J. Badiee, C. Soroudi, A. Myint, Y. Kang, B. V. Naini, S. D. Silva, V. R. Muthusamy, E. Esrailian, W. Hsu, and F. P. May, “S285validation of a deep machine learning tool to determine intra-procedural screening colonoscopy quality indicators in an academic health system,” *American Journal of Gastroenterology*, vol. 117, no. 10S, pp. e204–e205, 2022.

[68] P. Maurya, O. Jafari, B. Thatte, C. Ingram, and P. Nagarkar, “Building a comprehensive ner model for satellite domain,” *SN Computer Science*, vol. 3, 3 2022.

[69] N. Zhou, Q. Wu, Z. Wu, S. Marino, and I. D. Dinov, “Datasiftertext: Partially synthetic text generation for sensitive clinical notes.,” *Journal of medical systems*, vol. 46, pp. 96–, 11 2022.

[70] M. Chang, M. Chang, J. Z. Reed, D. Milward, J. J. Xu, and W. D. Cornell, “Developing timely insights into comparative effectiveness research with a text-mining pipeline.,” *Drug discovery today*, vol. 21, pp. 473–480, 2 2016.

[71] R. F. Hanson, V. Zhu, F. Are, H. Espeleta, E. Wallis, P. Heider, M. Kautz, and L. Lenert, “Initial development of tools to identify child abuse and neglect in pediatric primary care.,” *BMC medical informatics and decision making*, vol. 23, pp. 266–, 11 2023.

[72] J. Wu, F. Morrison, Z. Zhao, G. Haynes, X. He, A. K. Ali, M. Shubina, S. Malmasi, W. Ge, X. Peng, and A. Turchin, “Reasons for discontinuing insulin and factors associated with insulin discontinuation in patients with type 2 diabetes mellitus: a real-world evidence study,” *Clinical diabetes and endocrinology*, vol. 7, pp. 1–10, 1 2021.

[73] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley, “Deep learning predicts hip fracture using confounding patient and healthcare variables,” *NPJ digital medicine*, vol. 2, pp. 1–10, 4 2019.

[74] M. A. Gianfrancesco and N. D. Goldstein, “A narrative review on the validity of electronic health record-based research in epidemiology.,” *BMC medical research methodology*, vol. 21, pp. 234–234, 10 2021.

[75] L. A. Bastian, C. Brandt, and A. C. Justice, “Measuring multimorbidity: A risky business.,” *Journal of general internal medicine*, vol. 32, pp. 959–960, 6 2017.

[76] J. Koola, S. E. Davis, O. Al-Nimri, S. K. Parr, D. Fabbri, B. A. Malin, S. B. Ho, and M. E. Matheny, “Development of an automated phenotyping algorithm for hepatorenal syndrome.,” *Journal of biomedical informatics*, vol. 80, pp. 87–95, 3 2018.

[77] B. Shickel and A. Bihorac, “The dawn of multimodal artificial intelligence in nephrology.,” *Nature reviews. Nephrology*, vol. 20, pp. 79–80, 12 2023.

[78] C. Xiang and M. Abouelyazid, "The impact of generational cohorts and visit environment on telemedicine satisfaction: A novel investigation," 2020.

[79] D. G. Gordon and T. D. Breaux, "A cross-domain empirical study and legal evaluation of the requirements water marking method," *Requirements Engineering*, vol. 18, pp. 147–173, 4 2013.

[80] R. J. Gurrera and N. L. Perry, "Clozapine-associated aspiration pneumonia: Case series and review of the literature: Reply.," *Psychosomatics*, vol. 60, pp. 103–, 7 2018.

[81] C.-Y. Loo, W.-H. Lee, and Q. T. Zhou, "Recent advances in inhaled nanoformulations of vaccines and therapeutics targeting respiratory viral infections.," *Pharmaceutical research*, vol. 40, pp. 1015–1036, 4 2023.

[82] A. Javed, D. M. Rizzo, B. S. Lee, and R. Gramling, "Somtimes: self organizing maps for time series clustering and its application to serious illness conversations.," *Data mining and knowledge discovery*, vol. 38, pp. 813–839, 10 2023.

[83] J. H. Moore, I. Barnett, M. R. Boland, Y. Chen, G. Demiris, G. Gonzalez-Hernandez, D. S. Herman, B. E. Himes, R. A. Hubbard, D. Kim, J. S. Morris, D. L. Mowery, M. D. Ritchie, L. Shen, R. J. Urbanowicz, and J. H. Holmes, "Ideas for how informaticians can get involved with covid-19 research.," *BioData mining*, vol. 13, pp. 3–, 5 2020.

[84] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE international conference on big data (big data)*, pp. 5765–5767, IEEE, 2020.

[85] A. Rajkomar, E. Loreaux, Y. Liu, J. Kemp, B. Li, M.-J. Chen, Y. Zhang, A. Mohiuddin, and J. Gottweis, "Deciphering clinical abbreviations with a privacy protecting machine learning system.," *Nature communications*, vol. 13, pp. 7456–, 12 2022.

[86] E. Ekure, G. Ovenseri-Ogbomo, U. L. Osuagwu, K. E. Agho, B. N. Ekpenyong, K. C. Ogbuehi, A. O. Ndep, P. Okonji, K. P. Mashige, and K. S. Naidoo, "A systematic review of diabetes risk assessment tools in sub-saharan africa," *International Journal of Diabetes in Developing Countries*, vol. 42, pp. 380–393, 2 2022.

[87] E. Mahmoudi, W. Wu, C. Najarian, J. Aikens, J. Bynum, and V. Vydiswaran, "Leveraging natural language processing to identify caregiver availability for patients with alzheimer's disease," *Innovation in Aging*, vol. 6, pp. 449–450, 11 2022.

[88] R. Wadia, K. M. Akgün, C. Brandt, B. T. Fenton, W. Levin, A. H. Marple, V. Garla, M. G. Rose, T. H. Taddei, and C. R. Taylor, "Comparison of natural language processing and manual coding for the identification of cross-sectional imaging reports suspicious for lung cancer.," *JCO clinical cancer informatics*, vol. 2, no. 2, pp. 1–7, 2018.

[89] J. Wang, N. A. el Rub, J. H. Gray, H. A. Pham, Y. Zhou, F. J. Manion, M. Liu, X. Song, H. Xu, M. Rouhizadeh, and Y. Zhang, "Covid-19 signsym: a fast adaptation of a general clinical nlp tool to identify and normalize covid-19 signs and symptoms to omop common data model.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 28, pp. 1275–1283, 7 2020.

[90] D. Ding, A. Stachel, E. Iturrate, and M. Phillips, "1184. making pneumonia surveillance easy: Automation of pneumonia case detection," *Open Forum Infectious Diseases*, vol. 6, pp. S424–S425, 10 2019.

[91] A. Sharma and K. M. Goolsbey, "Learning search policies in large commonsense knowledge bases by randomized exploration," 2018.

[92] A. Cocci, M. Pezzoli, M. L. Re, G. I. Russo, M. G. Asmundo, M. Fode, G. Cacciamani, S. Cimino, A. Minervini, and E. Durukan, "Quality of information and appropriateness of chatgpt outputs for urology patients.," *Prostate cancer and prostatic diseases*, vol. 27, pp. 103–108, 7 2023.

[93] J. R. Gregg, M. Lang, L. L. Wang, M. J. Resnick, S. K. Jain, J. L. Warner, and D. A. Barocas, "Automating the determination of prostate cancer risk strata from electronic medical records.," *JCO clinical cancer informatics*, vol. 1, pp. 1–8, 6 2017.

[94] J. R. Machireddy, "Harnessing ai and data analytics for smarter healthcare solutions," *International Journal of Science and Research Archive*, vol. 08, no. 02, pp. 785–798, 2023.

[95] A. Sharma and K. Goolsbey, "Identifying useful inference paths in large commonsense knowledge bases by retrograde analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[96] Z. He, C. Tao, J.-G. Bian, and R. Zhang, "Selected articles from the fourth international workshop on semantics-powered data mining and analytics (sepda 2019).," *BMC medical informatics and decision making*, vol. 20, pp. 315–, 12 2020.

[97] V. Baroutsou, R. C. G. Pena, R. Schweighoffer, M. Caiata-Zufferey, S. Kim, S. Hesse-Biber, F. M. Ciorba, G. Lauer, M. Katapodi, and null null, "Predicting openness of communication in families with hereditary breast and ovarian cancer syndrome: Natural language processing analysis.," *JMIR formative research*, vol. 7, pp. e38399–e38399, 1 2023.