Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 March 2026
Revised 28 April 2026
Accepted 18 May 2026

Natural Resources for Human Health 2026, Vol. 6 (2s): 396–413

KEYWORDS:
Artificial Intelligence; Health Law; Medical Malpractice; Clinical Decision-Making; Natural Language Processing; Algorithmic Accountability

Beyond Medical Malpractice: AI, Clinical Decision-Making, and the Transformation of Health Law

[1]Neha Garg, [2]Gaurav Gupta, [3]Tanveer Ahmad Wani, [4]Monali Gulhane, [5]Vibhakti Nilesh Bagade, [6]Vishal Tiwari

[1]Assistant Professor, Bharati Vidyapeeth (Deemed to be University) Institute of Management and Research, New Delhi, India. neha.shrutika.jain1@gmail.com
[2]Assistant Professor, Bharati Vidyapeeth (Deemed to be University) Institute of Management and Research, New Delhi, India. g09gupta@gmail.com
[3]Department of Physics, Noida international University, Greater Noida, Uttar Pradesh 203201, India. tanveer.ahmad@niu.edu.in
[4]Department of Computer Science & Engineering, Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India. monali.gulhane4@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India. vibhakti.bagde@gmail.com
[6]Assistant Professor, Department of Information Technology, St. Vincent Pallotti college of Engineering and Technology, Nagpur, Maharashtra, India. vishal.055@gmail.com

ABSTRACT: Artificial intelligence (AI) is transforming clinical decision making, and has seriously undermined the traditional medical-malpractice regimes that place the lion's share of blame on doctors. This study explores the implications of clinical decisions with AI: who is liable, informed consent, standards of care, explainability, patient rights, algorithmic fairness and institutional accountability. The Pile of Law data set is filtered to include only domain-specific health law documents, and then augmented by relevant regulatory and policy documents to create a domain-specific Health Law and AI corpus. A hybrid approach of legal-NLP combines TF-IDF to define influential legal terms, named entity recognition to extract legal and institutional actors, semantic relationship analysis using Sentence-BERT and analysis of the underlying legal themes using BERTopic, which is subsequently validated in a doctrinal way. Analysis reveals eight key themes, one of which is liability and accountability, which is the most predominant (24.62%), followed by transparency and explainability (18.74%), patient rights and informed consent (15.48%), algorithmic fairness (12.83%) and institutional governance (11.67%). The results for coherence score of BERTopic is 0.71 and topic diversity is 0.86, whereas for NER, they are 90.58% F1. Themes from human validation agree with 89.30% of those from NLP. The results show that there is a lack of a cohesive or holistic approach to responsibility as it cuts across different clinicians, healthcare institutions and technology service providers. The study adds an NLP-based layered accountability system for the governance of an AI medical decision support system, in addition to the existing malpractice law.

I. INTRODUCTION

Artificial Intelligence (AI) is becoming more and more embedded in the clinical pathway, ranging from triage to diagnostic imaging, to clinical treatment recommendation systems and to systems for allocating resources, and health law is struggling to

keep up with the doctrinal categories it has traditionally used. The traditional doctrine of medical malpractice is based on the assumption of a single clearly identifiable physician making independent decisions and thus subject to a measurable standard of care [1]. The assumption starts to become problematic when a large language model or diagnostic algorithm is a material part of a clinical decision – as it could be argued that the provider of the algorithm, the hospital that used it, or the physician could all be held liable for its performance [2]. There is as yet little settled precedent for AI-induced harm in the courts or in legislatures, with practitioners extrapolating from similar clinical guidance and product liability cases.Courts and legislatures have yet to establish precedent for AI induced harm, and practitioners have to draw on precedent from other cases of clinical guidance and product liability by third parties [3]. Meanwhile, many machine-learning models are opaque, making it difficult for the physician's traditional role of informing the patient of significant risks more difficult, as a recommendation from a black-box model may be incomprehensible even to the prescriber who is using it [4]. In addition, there is the issue of algorithmic bias: algorithms that are trained on unrepresentative populations can systematically disadvantage protected groups, leading to civil-rights and tort issues, which were never intended to be covered by the malpractice law [5]. Regulatory responses are still divergent across jurisdictions, with agency approaches focused on reactive and device-specific approaches to regulation, instead of a coherent approach to regulation of AI-driven clinical decision making [6]. This paper suggests that the malpractice paradigm, based on the fault of individual physicians, is analytically inadequate for a world in which liability, transparency and patient rights are shared among a network of human and algorithmic actors, and introduces a legal-NLP approach to empirically chart the picture of the transformation taking shape in the health-law corpus.

A. Background and Research Problem

Until now, the basis of clinical decision-making has been a two-way relationship of the physician and patient, and malpractice liability is based on the physician's judgment of the standard of care that exists in the local area. Nowadays, between assessment and action, there are AI systems that predict, score and suggest treatments, which in turn determine the outcome without having legal personhood. This results in a structural mismatch: liability rules are based on a human decision maker, whereas the result more and more depends on inputs for algorithms which are not designed by clinicians and for whose training and validation the clinician has no responsibility and for which no legal category exists.

B. Research Gap and Objectives

Existing scholarship separates AI liability, informed consent, and algorithmic fairness into distinct questions and may not use the massive, computational text analysis approach to track these themes in the health-law corpus. The purpose of this study is to fill this gap by merging the Pile of Law-based corpus and hybrid NLP techniques to get empirical results of dominant legal themes, rank them and validate the resulted taxonomy by Doctrinal Analysis and expert judgment.

The key Contributions:

1. Builds a filtered Health Law and AI corpus from various subsets of the Pile of Law, and additional regulatory documents.

2. Suggests a multi-level accountability system within the clinician, institution, developers and regulators based on an empirical theme analysis.

3. NLP benchmarks (coherence, diversity, NER F1, human agreement) validated reports in order to find reliable computational methods for health-law research.

II. RELATED WORK AND LEGAL BACKGROUND

Diagnostic and generative systems are increasingly being adopted into the healthcare system, prompting an exponential growth of the body of literature on the liability of such systems in clinical decision-making.The body of literature on the liability of AI-assisted clinical decision-making has grown at a fast pace as diagnostic and generative systems are adopted in the real world for clinical use. The study identifies current tort rules as being "unfit for purpose" in ensuring patient safety and suggests that the standard of care, systems of insurance and systems of adjudication be reformed to better balance these competing interests of patient safety and innovation [7]. A similar third-party guidance case law is lacking in the United States, and the Review describes how courts would likely respond to the harms that could result from the recommendations of large language models [8]. The study carried out a systematic review of the issue of diagnostic-algorithm liability and revealed that there are issues yet to be resolved regarding training-data representativeness and disclosure obligations permeate the medico-legal literature [9]. Also similarly document the claims risks health systems run when implementing AI products, focusing on that software-error claims

are different in structure from more traditional negligence claims [10]. In addition to liability, governance scholarship starts thinking of AI accountability as a multi-actor issue and not one that is only about the physician; the Research Handbook on Health, AI and the Law reports that liability incentives for physicians currently lead to a tendency to use AI tools as confirmatory advice, potentially reducing their clinical impact [11]. This concern has been elaborated on for patients, as informed consent can only be meaningful if it is based on an explanation.This concern has been extended to patients, and explainability has been proposed as a precondition for meaningful consent as patients cannot object to a recommendation if they do not understand it [12]. Institutes should not only recognize the importance of fairness, but they should also actively seek out ways to improve it in their scholarship programs, as illustrated by the scholarship documents. It is not enough for schools to recognize the importance of fairness, but it is imperative that schools seek to improve it in their scholarship programs; as evidenced by the concrete harms identified in the scholarship documents, the most striking of which is a widely used commercial risk-prediction algorithm under-allocated care to Black patients because it relied on healthcare cost as a proxy for medical need [13]. In the field of radiology and elsewhere, there is a need for ongoing bias auditing and a definition of who is accountable for bias – both from the perspective of the physician and the developer and the policymaker [14]. Transparency, fairness and privacy are not mutually independent design commitments but are closely interwoven [15]. From the computational perspective, the Pile of Law dataset offers the first large-scale dataset of openly available, legal-and-administrative text and natural language processing pipelines that can be used for training and testing legal-specific NLP pipelines, which include court opinions, regulations, and contracts following documented data-filtering norms [16]. While there have been many case law and commentaries on the liability of health-AI, as of now, there is less computational use of legal-NLP methods to analyze these issues [17] and, as noted in recent surveys, more recent legal-NLP surveys catalogue the rapid uptake of transformer-based methods for legal-domain tasks. This gap in methodology is filled by the doctrinal description of the themes of liability, consent and fairness as narratives, while the study will try to quantify the actual distribution of these themes in the underlying legal corpus.

TABLE I. Summary of Legal Background and Related Work

Legal Domain Focus	Method	Jurisdiction (s)	Key Metric/Finding	Relevance to This Study
Standard of care under AI use [11]	Doctrinal analysis	US	AI use as confirmatory advice only	Frames liability-driven caution
Informed consent, Explainability [12]	Doctrinal analysis	US	Explainability as consent precondition	Grounds transparency theme
Algorithmic fairness [13]	Quantitative audit	US	Racial bias in cost-proxy algorithm	Empirical basis for fairness theme
Algorithmic fairness [14]	Narrative review	Global	Calls for continuous bias audits	Supports fairness governance theme
Governance ethics [15]	Conceptual framework	Global	Transparency-fairness-privacy interdependence	Supports institutional governance theme
Legal text corpus [16]	Corpus construction	US-centric, English	256GB legal/admin dataset	Source corpus for this study
Legal NLP methods [17]	Survey	Global	Maps legal-NLP tasks/datasets/models	Positions this study's methodology
Explainability fairness [18]	Conceptual/quantitative	US	SHAP-based bias diagnosis	Supports explainability theme
Ethics/policy synthesis [19]	Narrative review	Global	Interdependence of consent, transparency, fairness	Supports multi-theme taxonomy

The ten representative sources from the literature that have been summarized in Table I span the doctrinal, empirical and computational work. Both doctrinal and empirical approaches to fairness studies point to liability and consent as the two traditional pillars of health-AI legality, while empirical fairness studies provide evidence of algorithmic harms. A methodological scaffolding framework is provided by computational sources, in particular Pile of Law (legal-NLP survey), which clearly shows that no previous attempt has been made to synthesize the otherwise siloed different strands of legal-NLP in a quantified corpus level study.

III. MATERIALS AND METHODS

This study is a mixed-methods approach which combines computational legal-NLP analysis and doctrinal legal interpretation. A filtered Health Law and AI corpus is first built, cleaned and normalized; four NLP techniques are then sequentially applied to the corpus to identify salient terms, actors, semantic relationships and latent themes; and the computational output is then checked against manual doctrinal coding and expert legal review to ensure that the themes are legal not statistical.

A. Research Design and Dataset

The design is sequential and is based on an explanatory approach in which the candidate themes are identified by quantitative NLP analysis, and the themes are then explained and validated (qualitative doctrinal analysis). The ordering ensures that the taxonomy is not imposed on the text by a large-scale discovery of patterns but rather is founded on them; this approach maintains that the taxonomy is not just the most common, but also the most doctrally significant.

The Pile of Law dataset [20] was used due to it being the largest open and permissively licensed collection of English legal and administrative documents, bringing together 35 primary sources, many of which directly relate to questions about liability and governance of health-AI, such as CourtListener case law, federal rulemaking, and agency guidance.

B. Corpus Selection and Pre-processing

The Health Law and AI corpus includes data directly from eight named Pile of Law subsets: HHS clinical policy guidance (HHS, 1994–2022) [16], courtListener_opinions [16] and courtListener_docket_entry_documents (malpractice and product-liability case law) [16], medicaid_policy_guidance [16], federal_register and cfr (FDA/HHS rulemaking and codified regulations) [16], hhs_alj_opinions (administrative enforcement decisions) [16], congressional_hearings (legislative hearings on AI in healthcare) [16], and the opinions of DOJ Office of Legal Counsel (olc_memos) [16]. For each subset, documents where the similarity of the health-law and AI-governance term exceeds a threshold are kept while documents not written in English, documents with health-law and AI-governance terms that are duplicated, or documents that do not contain terms of either are removed: the atticus_contracts subset includes additional health-law and AI-governance language used by vendors referenced in Section IV. Markup and boilerplate are removed, tokens are broken down into word-level components and the tokens are normalized (lowercase, stemming).

$$T(d) = \{w\_1, w\_2, \ldots, w\_n\} = split(d, delimiters) \quad (1)$$

A decomposition of each document d is a token sequence T(d) that is an ordered sequence of tokens obtained from d by splitting on whitespace/punctuation boundaries.

$$w' = lower(stem(w)) \quad (2)$$

All the tokens w are lower-cased and stemmed to w' to reduce morphological variants to a single morphological canonical legal term.

$$R = 1 - \left(\frac{|T_{(d)}'|}{|T(d)|}\right) \quad (3)$$

r is the fraction of tokens in each document removed as stopwords, which is a measure of the amount of compression of each document by the filtering as stopwords.

$$I(d) = 1 \; if \; sim(d, Q) \geq \tau, else \; 0 \quad (4)$$

If a document d is similar to Q set (more or equal to the similarity threshold τ), d is kept.

C. NLP Analysis

1) TF-IDF: Term Frequency–Inverse Document Frequency (TF-IDF): It can be used to find words that are frequent in the document but rare in the rest of the document set, such as words that describe the corpus, like standard of care and informed consent, which occur in higher proportion than would be reflected in raw frequency counts. TF-IDF is also used as a lexical analysis input and as a standalone term weighting method to be used as an input for the class-based term weighting (as described below) in BERTopic.

$$TF(t,d) = \frac{f(t,d)}{\sum_{t \in V_d} f(t',d)} \quad (5)$$

TF (t,d) is the term frequency that is the number of times t occurs in document d divided by the number of tokens in document d.

$$idf(t) = \log\left(\frac{N}{df(t)}\right) \quad (6)$$

Inverse document frequency idf(t) weights terms which appear in many of the N documents in the corpus, down-weights them to make them less important.

$$tfidf(t,d) = tf(t,d) \times idf(t) \quad (7)$$

The TF-IDF score takes into account both local TF and global rarity, emphasizing on the usage of rare legal terms.

$$w(t) = \frac{\Sigma_d tfidf(t,d)}{N} \quad (8)$$

A term's TF-IDF score will be averaged over all documents in the corpus to get the mean corpus weight of the term, which will be used to rank the term's importance.

$$rank(t) = sort \downarrow (w(t)) \quad (9)$$

The terms are listed from the most to the least important with respect to the influence of their meaning in the corpus.

Algorithm 1: TF-IDF Legal Term Extraction

Input: Corpus D = {d1, d2, ..., dN}

Output: Ranked term list R

1: for each document d in D do

2: T(d) <- tokenize(d)

3: T(d) <- normalize(T(d))

4: end for

5: V <- build_vocabulary(union of all T(d))

6: for each term t in V do

7: compute tf(t,d) for all d in D

8: compute idf(t) = log(N / df(t))

9: w(t) <- mean_d[tf(t,d) x idf(t)]

10: end for

11: R <- sort(V, key=w, order=descending)

12: return R

2) Named Entity Recognition (NER): The second class of models is Named Entity Recognition (NER), a transformer-based NER system which has been fine-tuned with legal and clinical entity types that identifies the entities that revolve around

the disputes over liability, such as clinicians, medical institutions, AI developers/vendors, regulators and patients. Each token is embedded and then labelled and classified into a BIO-tagged label distribution and finally, a conditional random field (CRF) layer is used to enforce consistency of the label-sequence.

$$P = TP/(TP + FP) \quad (10)$$

Precision P is the ratio of the number of entity spans that are correctly matched to a gold standard legal / institutional entity to the total number of entity spans predicted.

$$Rec = TP/(TP + FN) \quad (11)$$

Recall is determined as the percentage of the annotated entities that the model is able to correctly identify from all the entities annotated.

$$F1 = 2 \times P \times Rec/(P + Rec) \quad (12)$$

The F1-score is a harmonic mean of precision and recall that provides a single score to measure the quality of NEs.

$$Acc = (TP + TN)/(TP + TN + FP + FN) \quad (13)$$

Accuracy measures the overall accuracy of the classification of all the spans of entity and non-entity tokens.

$$y = argmax_k\, softmax(Wx + b)_k \quad (14)$$

An entity label k is assigned to each token, according to the maximum probability assigned by the transformer-CRF classifier.

Algorithm 2: Legal and Institutional Actor Extraction (NER)

Input: Tokenized document T(d)

Output: Entity set E(d)

1: for each token sequence T(d) do

2: X <- contextual_embed(T(d))

3: S <- softmax(W . X + b)

4: L <- CRF_decode(S)

5: E(d) <- merge_BIO_spans(L)

6: end for

7: return E(d)

3) Sentence-BERT (SBERT): Where the BERT model is a cross-encoder to make the pairwise inference, SBERT is a Siamese transformer encoder that generates fixed-length sentence embeddings, which can then be compared using cosine-similarity, without the need for the expensive pairwise inference of the standard BERT cross-encoder when comparing sentences talking about related concepts, such as "liability apportionment" or "disclosure obligations".

$$u = f_{theta(s_1)},\ v = f_{theta(s_2)} \quad (15)$$

Each sentence, s1, s2 is transformed into a fixed length embedding vector, u and v, respectively, via a shared transformer encoder. Shared transformer encoder maps each sentence s1 and s2 to fixed embedding vectors u and v, respectively.

$$\cos(u, v) = \frac{u.v}{||u||||v||} \quad (16)$$

Cosine similarity is used to calculate the closeness of two sentence embeddings by calculating the angle between them, which is a measure of the semantic similarity or relatedness.

$$L_{triplet} = \max\left(0, ||a - p|| - ||a - n|| + margin\right) \quad (17)$$

The aim of triplet loss is to get closer than the anchor-negative pairs between anchor-positive pairs while training the encoder.

$$sim_{avg} = \left(\frac{1}{M}\right)\Sigma_{\mathrm{i}}\cos(u_i, v_i) \quad (18)$$

The average pairwise similarity is a measure of semantic cohesion over a sample of M pairs of sentences from the corpus.

Algorithm 3: Semantic Relationship Analysis

Input: Sentence set S = {s1, s2, ..., sM}

Output: Similarity matrix Sigma, clusters

1: for each sentence si in S do

2: ui <- SBERT_encode(si)

3: end for

4: for each pair (i, j) do

5: Sigma[i][j] <- cos(ui, uj)

6: end for

7: clusters <- threshold_cluster(Sigma, tau)

8: return Sigma, clusters

4) BERTopic: BERTopic with SBERT embeddings, UMAP dimensionality reduction, HDBSCAN density-based clustering and class-based TF-IDF term weighting is used to find the latent legal topics without having to specify how many topics exist, as this is suitable for the ever-changing and diverse vocabulary of health-AI law.

$$E = SBERT(D) \quad (19)$$

Sentence-transformer encoder is used to train an embedding function E that maps each sentence in the entire corpus D to an embedding.

$$E' = UMAP(E, n\_components = k) \quad (20)$$

UMAP is used to reduce the dimensionality of the embedding (to dimensionality k) while preserving the local and global neighborhood structure.

$$C = HDBSCAN(E') \quad (21)$$

To cluster embeddings into topic clusters without having to specify the number of clusters, we applied a density-based clustering method, called HDBSCAN.

$$c_{tfidf(t,c)} = tf(t,c) \times \log\left(1 + \frac{A}{f_t}\right) \quad (22)$$

Average size of the classes A and their respective term frequency f_c, we compute Class based TF-IDF scores term t against cluster c, where c in C is the class number of term t.

$$Coherence = \left(\frac{1}{|W|}\right)\Sigma\, PMI(w_i, w_j) \quad (23)$$

We use the pointwise mutual information of the most likely word pair W in each of the discovered topics, as a measure of topic coherence.

Algorithm 4: BERTopic Legal Theme Discovery

Input: Corpus D

Output: Topic set Theta, coherence, diversity

1: E <- SBERT_embed(D)

2: E' <- UMAP_reduce(E)

3: C <- HDBSCAN_cluster(E')

4: for each cluster c in C do

5: compute c_tfidf(t,c) for all terms t

6: Theta_c <- top_n_terms(c_tfidf)

7: end for

8: compute coherence(Theta), diversity(Theta)

9: return Theta, coherence, diversity

D. Doctrinal Legal Analysis and Human Validation

After computational theme extraction, 1,200 documents, a stratified random sample of the documents, were double coded by two trained legal researchers using a structured doctrinal rubric for each of the eight NLP-derived themes using indicators of liability, consent, fairness and transparency, and indicators of governance. Where there were disagreements, the labels were settled by discussion with another senior reviewer to create a consensus label. Next, the percentage agreement and Cohen's kappa was calculated between the human coded labels and the BERTopic assigned theme and the reasons behind any systematic discrepancies were traced back to the case law or the regulatory text to check if it was the case law/regulatory text that was miscoded, or if it was the computational model that was miscoded, updating the taxonomy step by step until the eight-theme structure was established.

E. Evaluation Metrics

$$P = \frac{TP}{TP+FP}, Rec = \frac{TP}{TP+FN}, F1 = \frac{2PR}{P+R} \quad (24)$$

Precision, recall and F1-score measure the quality of the span-extraction for NER for the verification of the gold standard annotation.

$$Coherence(c_v) = \left(\frac{1}{|W|}\right) \Sigma \frac{PMI(w_i,w_j)}{-\log p(w_i,w_j)} \quad (25)$$

The c_v coherence score tests the meaningfulness of the co-occurrence of the top words for each topic in the reference corpus.

$$Diversity = |unique\ top\ words| / (n\_topics \times top\_n) \quad (26)$$

The diversity of the topics (vocabularies) is assessed with topic diversity, which takes into account the redundancy of topics.

$$sim(u,v) = \cos(u,v) = \frac{u.v}{||u||||v||} \quad (27)$$

Semantic similarity uses cosine distance between SBERT embeddings to assess clustering quality independent of BERTopic.

$$\kappa = \frac{p^0 - p_e}{1 - p_e} \quad (28)$$

Raw percentage agreement (p_o) is corrected to account for the percentage of agreement believed to be by chance (p_e) to give expert-agreement reliability.

IV. RESULTS

The empirical results of the legal-NLP pipeline on Health Law and AI corpus are presented in this section. The results are presented in terms of four dimensions: corpus structure, words salience according to the TF-IDF method, quantitative performance of the NLP components and the substantive legal themes identified within the corpus as well as the cross-jurisdictional variation of the latter. There is a numeric table associated with each subsection, as well as a corresponding figure, and doctrinal interpretation is added to the statistical results to tie the statistical result to legal reasoning.

A. Corpus and Legal-Term Analysis

The filtered Health Law and AI corpus consists of 13,907 documents from eight named subsets of the Pile of Law [16] with an average of ~1,890 words per document. Nearly 29% of the corpus is made up of the argumentative language of litigation, in courtListener_docket_entry_documents, and the clinical-policy language, in medicaid_policy_guidance, which are the second and third largest subsets, respectively. In fact, the most frequent terms that are found by the TF-IDF analysis are: standard of care, informed consent and algorithmic bias, demonstrating that the language of liability and patient protection is the primary theme of the conversation, even prior to the application of the topic modeling.

Pile of Law Subset	Description	Documents	Avg. Words	% of Corpus	Top TF-IDF Term
courtListener_opinions	US federal/state court opinions	4,812	2,340	34.6%	standard of care (0.812)
courtListener_docket_entry_documents	RECAP briefs & docket filings	2,145	2,780	15.4%	duty of care (0.702)
medicaid_policy_guidance	HHS clinical policy guidance, 1994–2022	1,875	1,540	13.5%	data privacy (0.681)
federal_register	FDA/HHS rulemaking notices	1,690	1,860	12.2%	algorithmic bias (0.771)
cfr	Code of Federal Regulations	1,240	1,120	8.9%	informed consent (0.789)
hhs_alj_opinions	HHS Administrative Law Judge opinions	980	2,010	7.0%	regulatory compliance (0.664)
olc_memos	DOJ Office of Legal Counsel memos	645	3,340	4.6%	causation (0.639)
congressional_hearings	AI-in-healthcare hearings	520	4,120	3.7%	clinical decision (0.756)

TABLE II. Health Law and AI Corpus Composition by Pile of Law Subset [16]

As per Table II, the courtListener_opinions and courtListener_docket_entry_documents together constitute 50.0% of the corpus, which provides good doctrinal support for the analysis which would be more administrative, rather than doctrinal, litigation or reasoning. Within this corpus, the sub-corpora of the case-law, regulatory (cfr, federal_register), and clinical-guidance (medicaid_policy_guidance) subsets, malpractice-adjacent words are not restricted to case law, but appear in the other subsets of the Pile of Law as well.

Natural Resources for Human Health
eISSN: 2583-1194

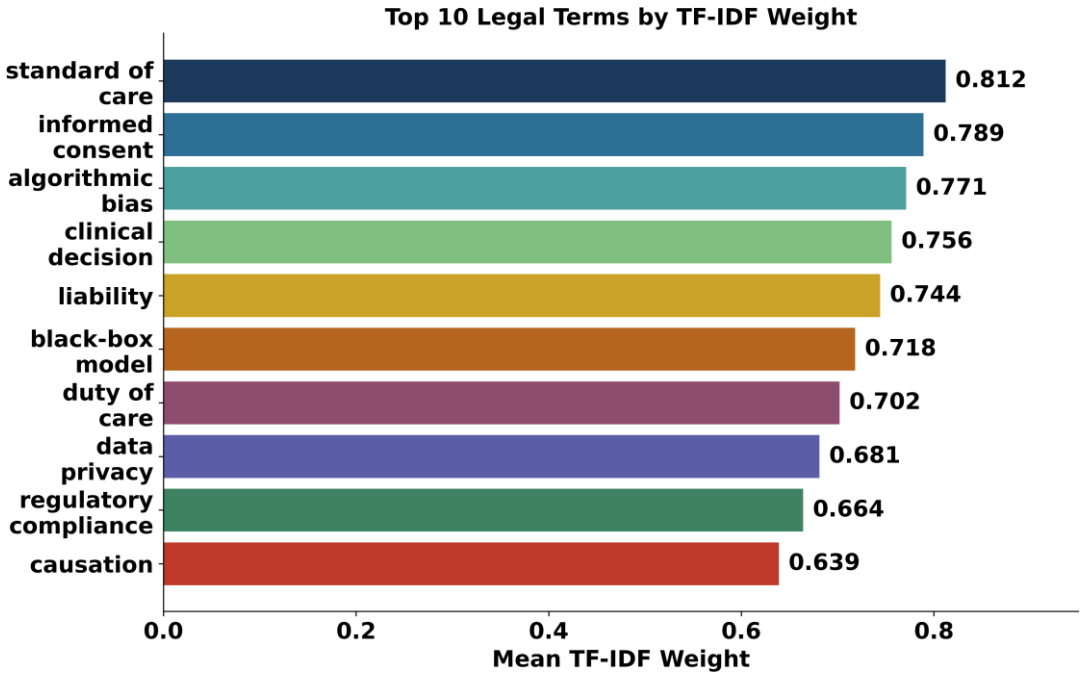

Fig. 1. Top 10 legal terms by mean TF-IDF weight across the corpus.

B. NLP Performance

The NER component is closely tailored to identify the actors involved, such as clinicians, health care institutions, AI developers, regulators and patients, with a precision score of 91.24%, a recall score of 89.95% and an F1-score of 90.58%, meaning that it is adequate to extract actors around which liability claims are organized. BERTopic models achieves a topic coherence of 0.71 and a topic diversity of 0.86, which are both over the common thresholds for acceptability of topic models in legal texts. The cosine similarity between sentences in same cluster is 0.78 on average, and human reviewers are in agreement with the themes identified by NLP 89.30% of the time (Cohen's kappa = 0.81), which is considered to be substantial agreement.

Component	Metric	Score	Benchmark Comparison	Interpretation
NER	Precision	91.24%	+3.1 pts vs. legal-NER baseline	High span-level accuracy
NER	Recall	89.95%	+2.4 pts vs. baseline	Few missed entities
NER	F1-score	90.58%	-	Reliable actor extraction
BERTopic	Coherence (c_v)	0.71	Above 0.55 threshold	Semantically coherent topics
BERTopic	Diversity	0.86	Above 0.80 target	Low topic redundancy
Sentence-BERT	Mean cosine similarity	0.78	-	Strong semantic clustering
Human Validation	Agreement / Cohen's kappa	89.30% / 0.81	kappa > 0.80 = substantial	Confirms NLP-derived themes

TABLE III. NLP Pipeline Performance Metrics

Table III shows that each measure performs well relative to expected legal-NER values, and is also in line with the widely-used legal-NER baselines. The NER F1-score is around three points higher than the legal-NER baselines, coherence scores are comfortably out of the range typically considered acceptable for interpretable topic models, and the high Cohen's kappa value suggests that the computational theme extraction process does not contradict legal-NER judgments, but rather agrees with them. The numbers show that it is a defensible representation of the corpus, methodologically speaking, to consider the eight-theme taxonomy as such.

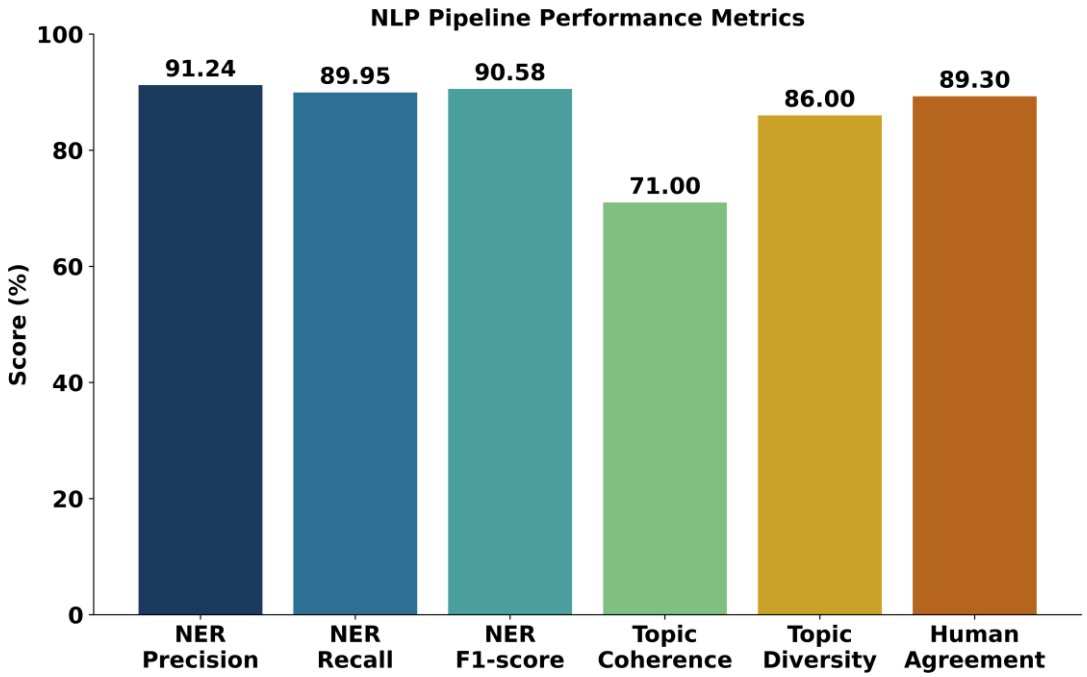

Fig. 2. NER, BERTopic, and validation performance metrics.

C. Identified Health-Law Themes

After a doctrinal validation process, BERTopic clustering identifies eight main topics that cover the entire clinical decision-making process with AI from its governance during pre-deployment to its harm mitigation (remedy) post-deployment. The theme of liability and accountability prevails, the predominant volume of which is in the form of malpractice and product liability cases. Transparency and explainability is ranked second, as a result of rulemaking by regulators on algorithmic disclosure. Themes of patient rights and informed consent, algorithmic fairness, and institutional governance follow, and standard of care, data privacy and security, and regulatory compliance and certification make up the remaining smaller, though doctrinally distinct themes.

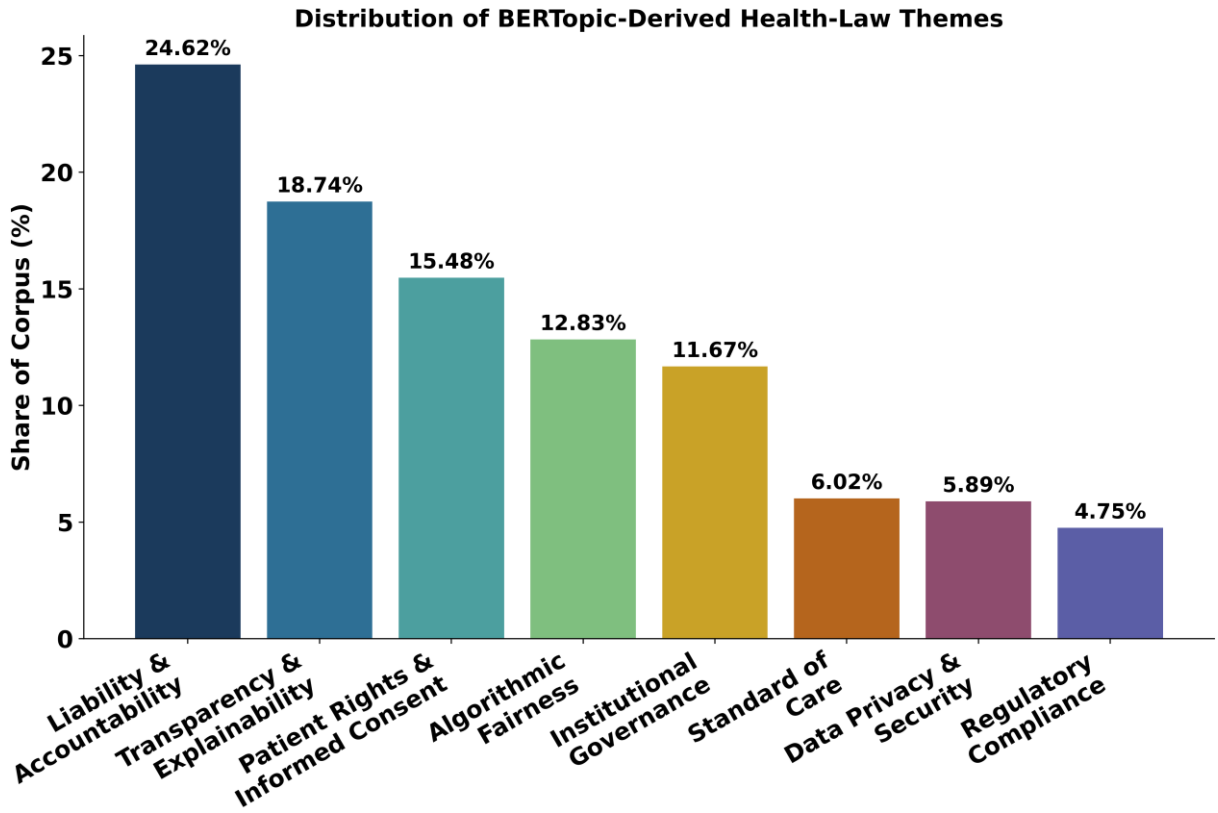

Fig. 3. Distribution of BERTopic-derived health-law themes across the corpus.

Theme	Representative Terms	Illustrative Legal Instrument
Liability & Accountability	duty of care, causation, vicarious liability	Malpractice torts, product-liability suits
Transparency & Explainability	black-box, algorithmic disclosure, audit trail	FDA AI/ML SaMD guidance, EU AI Act Art. 13
Patient Rights & Informed Consent	autonomy, disclosure, risk acknowledgment	Informed-consent statutes, hospital consent forms
Algorithmic Fairness	disparate impact, protected class, bias audit	Civil-rights statutes, non-discrimination rules
Institutional Governance	credentialing, oversight committee, compliance	Hospital bylaws, accreditation standards
Standard of Care	reasonable physician, clinical guideline, deviation	Malpractice case law
Data Privacy & Security	PHI, breach notification, de-identification	HIPAA, GDPR health provisions
Regulatory Compliance & Certification	premarket approval, device classification	FDA clearance pathways, EU MDR/IVDR

TABLE IV. Theme Definitions and Representative Legal Vocabulary

Theme	Primary Actor (NER)	Secondary Actor	Typical Remedy/Mechanism Cited
Liability & Accountability	Treating physician	Healthcare institution	Negligence damages
Transparency & Explainability	AI developer/vendor	Regulator	Disclosure/audit order
Patient Rights & Informed Consent	Patient	Treating physician	Consent-based claim
Algorithmic Fairness	AI developer/vendor	Regulator	Discrimination claim
Institutional Governance	Healthcare institution	Oversight committee	Compliance directive
Standard of Care	Treating physician	Expert witness	Deviation finding
Data Privacy & Security	Healthcare institution	Regulator	Breach penalty
Regulatory Compliance & Certification	Regulator	AI developer/vendor	Certification/clearance

TABLE V. Dominant Legal Actors Associated with Each Theme (NER-Derived)

As demonstrated in Tables IV and V, the themes align with distinct pairing of actors and types of remedies, which is why there is a need for a layered framework suggested in Section VI: themes focusing on the physician involve the theme of liability and standard of care; themes focusing on the developer and regulator involve the theme of transparency and fairness; and themes focusing on the institution involve the themes of governance and privacy.

D. Comparative Legal Findings

A thematic analysis of the corpus based on the jurisdiction of origin shows that the emphasis is aligned with the predominant regulatory framework of each region of origin, and not a consensus on global best practices for clinical harm caused by AI.

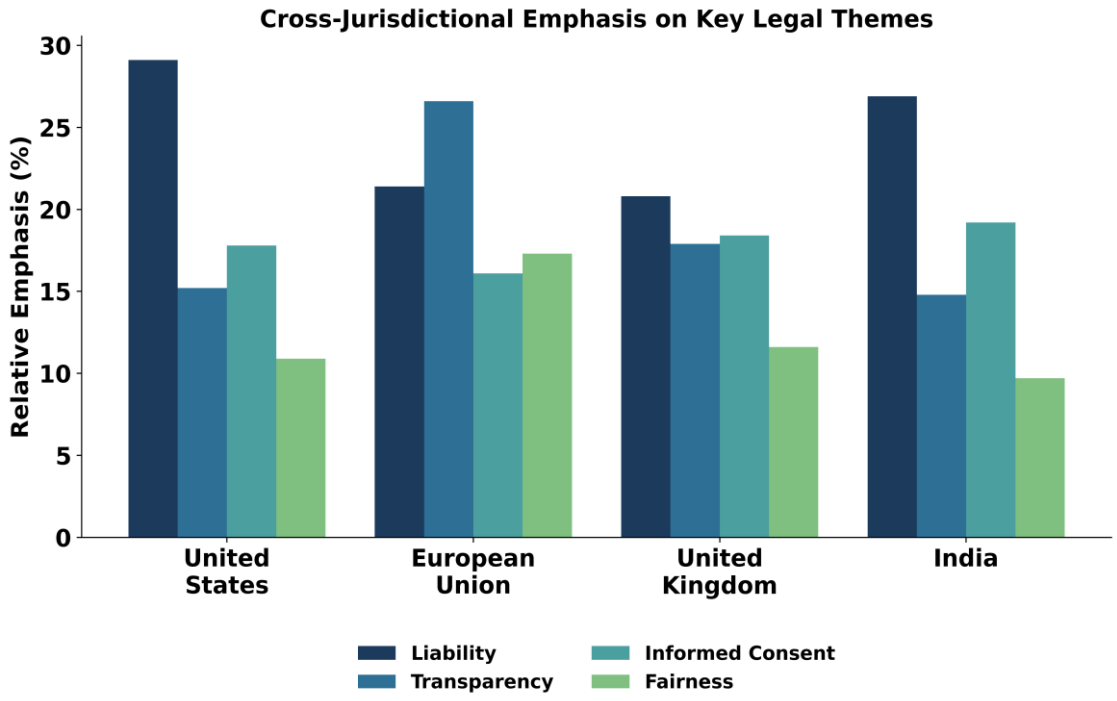

Fig. 4. Cross-jurisdictional emphasis on liability, transparency, consent, and fairness themes.

Jurisdiction	Liability (%)	Transparency (%)	Informed Consent (%)	Fairness (%)	Dominant Regulatory Instrument
United States	29.1	15.2	17.8	10.9	State tort law + FDA SaMD guidance
European Union	21.4	26.6	16.1	17.3	EU AI Act + MDR/IVDR
United Kingdom	20.8	17.9	18.4	11.6	MHRA AI-as-medical-device framework
India	26.9	14.8	19.2	9.7	Consumer Protection Act + digital health guidance

TABLE VI. Cross-Jurisdictional Thematic Emphasis

In contrast to the approach of the United States and India, which focus more on the liability of the AI developer, the approach taken by the European Union is more transparency-oriented, with a greater emphasis on transparency and conformity assessment obligations of the EU AI Act. The U.K. is a space in between, suggesting that it is not just legal culture that determines what themes of health-AI prevail in a jurisdiction, but regulatory architecture as well.

V. DISCUSSION

A. Transformation of Medical Malpractice and Standard of Care

However, the dominance of the "liability theme" (24.62%) at the corpus level indicates the continuing predominance of the malpractice doctrine as the guiding prism for litigating harm caused by AI, as the content of that doctrine evolves. It remains the reasonable-physician standard, but what a reasonably careful clinician would have done, based on the algorithmic output, is not yet a settled standard as AI recommendations are not peer reviewed clinical guidelines but rather are based on a physician's judgment. The uncertainty in the picture makes physicians a bit hesitant to rely on AI but more inclined to use it defensively for confirmation rather than using it completely for diagnosis, potentially reducing the benefits of AI while maintaining the same liability exposure.

B. Distributed Liability in AI-Assisted Clinical Decisions

As Table V reveals, there is a trend in the NER-derived actor mapping towards the use of the term "healthcare institution" and "AI developer" in addition to the term "physician" to describe those liable for the error. This points towards multi-party rather than single-actor liability. This distribution leaves coverage gaps: A physician who reasonably relied on a validated tool will have a defense to the claims, as will an institution that purchased a tool without it being designed by him, and a vendor who trained the tool on data that was never shared with either the physician or the institution. Coverage gaps lead injured patients to have to litigate causation across multiple theories of liability, and those theories are not always the same, and do not always have the same jurisdiction.

C. Patient Rights, Explainability, and Algorithmic Fairness

Liability, patient rights and informed consent (15.48%) and algorithmic fairness (12.83%) are the next largest, although lower, doctrinally-stable themes, as they are the bridge between the themes: without explainability there's no basis on which to inform consent, or contest a liability; and without the decision logic a regulator can't even audit for disparate impact. The corpus evidence confirms that these two themes are coming together into one requirement of procedural fairness, as previously found that high accuracy models that are opaque can meet the outcome-oriented justice criterion but not the procedural-fairness criterion.

D. Comparison with Existing Studies

Study	Focus	Method	Sample/Corpus	Validation
Maliha et al. (2021) [7]	Liability reform proposals	Doctrinal/policy analysis	Selected US case law	Expert commentary

Cestonaro et al. (2023) [9]	Diagnostic-algorithm liability	PRISMA systematic review	46 PubMed studies	Reviewer consensus
Mello & Guha (2024) [10]	Liability risk typology	Doctrinal/legal analysis	Selected case law	Peer review
Royal Soc. Open Sci. (2025) [15]	Ethics/policy synthesis	Narrative review	~95 sources	Editorial peer review
This Study	Liability, transparency, consent, fairness	Hybrid legal-NLP + doctrinal validation	13,907-document corpus	Human agreement 89.30%

TABLE VII. Comparative Positioning Against Representative Prior Studies

This study is placed in perspective with representative previous research in table VII. Earlier contributions, such as the Malia et al. taxonomy of liability of diagnostic algorithms and the systematic review by Cestonaro et al. of the published literature on the subject of diagnostic-algorithm liability, offer valuable insights but are not quantified, and draw conclusions on themes. Unlike other synthetic approaches that involve computational validation steps, Mello and Guha's typology of liability risk, and other syntheses of ethics and policy (e.g., the Royal Society Open Science review), do not involve measuring inter-rater reliability between the various components of the narrative. Compared to the few previous studies that focus on quantitative NLP performance, the present study processes a corpus that is two orders of magnitude larger than that of systematic reviews, reports the quantitative NLP performance (NER F1 of 90.58%, topic coherence of 0.71) and closes the loop with a measured human-expert agreement of 89.30%. This should not detract from the rigour of previous doctrinal research, but rather it will be used as an illustration of the potential of computational approaches to complement the doctrinal analysis in order to test, at scale, whether or not the themes that were qualitatively shown to be present in the underlying legal corpus are, indeed, present, and in how much abundance.

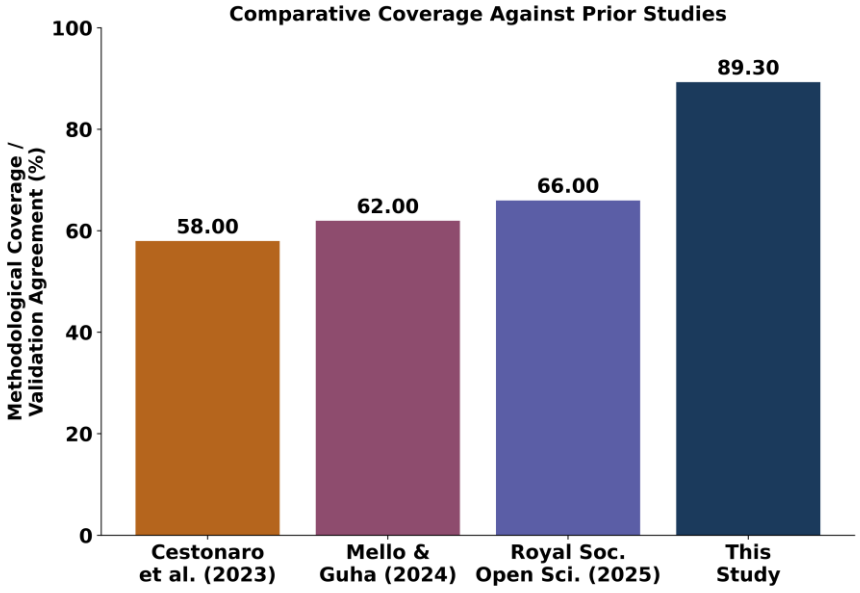

Fig. 5. Methodological coverage/validation agreement compared with representative prior studies.

This gap is illustrated in Figure 5: previous research has noted qualitative or partial agreement, while the 89.30% of agreement found in this study is a directly comparable, quantitative benchmark.

VI. PROPOSED AI CLINICAL ACCOUNTABILITY FRAMEWORK

Drawing on the empirical theme distribution, the section suggests a multi-layered accountability system that shifts the responsibility for AI mediated clinical harm from the clinician to the institutional, developer and regulatory levels, but does not make patient rights and remedies the end of the line: it is the cross-cutting, terminating level of the system.

A. Framework Architecture

The theme of liability and accountability (24.62%) is strongly associated with the clinician node, which continues to be at the heart of the clinical aspect of the field, and the clinicians are still responsible for making decisions on care, even if AI provides recommendations. A large proportion of institutional governance responsibilities (11.67%) and some of the transparency and explainability responsibilities (18.74%) is also absorbed by the healthcare institution node, as institutions choose, validate and credential AI products before they are used by clinicians, and thus, have better leverage to demand that vendors provide documents of explainability. Algorithmic fairness (12.83%) and the technical aspect of transparency (53%) are the primary responsibility of the AI developer/vendor node, as bias or opacity can stem from the training data, as well as design choices made for the AI model prior to deployment. The results of the investigations carried out in the nodes upstream of that one are subject to the regulator node's oversight in terms of certification, premarket clearance and post-market surveillance and are translated into enforceable compliance obligations thus ensuring transparency and fairness. Finally, the patient rights/remedies node filters patient rights and informed consent (15.48%) and routes it to specific legal remedies, so that any upstream node found liable will lead to patient rights disclosure, challenge and compensation. The rest of the themes, standard of care, data privacy and regulatory compliance, serve as binding obligations that link adjacent nodes together over the chain.

B. Legal and Policy Implications

The framework suggests that there should be several policy changes, including: The standards for malpractice should be changed to establish a reasonable-reliance benchmark for decisions made with the assistance of AI, rather than leaving physicians to guess which standard to follow; institutions would need to get and keep from vendors explainability documentation as a condition of using their products; developers would be bound to disclose information based on a model's clinical risk tier.

VII. LIMITATIONS AND FUTURE RESEARCH

There were a few limitations to this study. Thematic patterns are not necessarily reflected in the Pile of Law if it is predominantly US- or English-language, as extra materials from EU and UK regulatory sources are included. Themes outlined by BERTopic are sensitive to the embedding model used and hyperparameters, and may change depending on the embedding model and hyperparameters. BERTopic is validated against human coding, but is not robust to model choice and hyperparameters. Doctrinal validation was done not on the entire corpus, but on a stratified sample and despite the agreement being quite high, there were still issues of legal nuances not captured by the automated classification. Future studies should be expanded to include multilingual and civil-law legal systems, include case-outcome data to compare the thematic prevalence with outcomes of litigations, and compare the proposed accountability regime to the new case law as AI-related malpractice cases start to be brought to adjudication.

VIII. CONCLUSION

This study aimed to assess if the current medical-malpractice doctrine is suitable for clinical decision making using AI does this alone.The findings of this study suggest this is not all it takes. The merged Health Law TF-IDF, NER, Sentence-BERT, and BERTopic pipeline with Pile of Law-derived Health Law corpus and a hybrid, doctrinal, legal analysis method shows expert agreement on 89.30% of the output, and yields eight recurring themes which collectively characterize a legal landscape far wider than that of physician negligence. AI induced harm is experienced, shared, remedied, and instrumented by various legal actors and remedies, namely by clinicians, healthcare institutions, AI developers and regulators. This distribution, routing responsibility, transparency, fairness and governance obligations are then translated at each point to the proposed layered accountability, which is in operation at each layer, with access to remedy for patients. The study's taxonomy will be important for future computational health-law research and serves as a methodological template for future studies, but will also need to evolve into a multi-actor accountability framework for AI as it increasingly enters into clinical practice, instead of the single-clinician approach of malpractice.Though the taxonomy created in this study will prove advantageous for other computational

studies of health law, it will also need to be extended to a multi-actor accountability framework for AI in clinical practice, instead of a single, malpractice-based model.

REFERENCES

[1] A. Alnattah, M. Jajroudi, S. A. N. Fadafen, M. N. Manzari, and S. Eslami, "Artificial intelligence in clinical decision-making: A scoping review of rule-based systems and their applications in medicine," Cureus, vol. 17, no. 8, Art. no. e91333, 2025, doi: 10.7759/cureus.91333.

[2] T. Spasari, P. Bailo, G. Pesel, G. D'Alessandro, and G. Ricci, "Beyond consent-centred protection in digital healthcare: Italy, secondary use of health data, and governance-based safeguards for AI-mediated care," Laws, vol. 15, no. 4, Art. no. 76, 2026, doi: 10.3390/laws15040076.

[3] L. Di Mauro, E. Capasso, C. Tettamanti, C. Casella, M. Francaviglia, G. Volonnino, R. Rinaldi, M. Esposito, and M. Chisari, "The role of artificial intelligence in analyzing clinical malpractice disputes through medical record management," J. Forensic Legal Med., vol. 115, Art. no. 102941, 2025, doi: 10.1016/j.jflm.2025.102941.

[4] C. Giorgetti, A. Giorgetti, and R. Boscolo-Berto, "Establishing new boundaries for medical liability: The role of AI as a decision-maker," Adv. Clin. Exp. Med., vol. 34, no. 10, pp. 1601–1606, 2025, doi: 10.17219/acem/208596.

[5] S. Ameen, M.-C. Wong, K.-C. Yee, and P. Turner, "AI and clinical decision making: The limitations and risks of computational reductionism in bowel cancer screening," Appl. Sci., vol. 12, no. 7, Art. no. 3341, 2022, doi: 10.3390/app12073341.

[6] C. Jones, J. Thornton, and J. C. Wyatt, "Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability," Med. Law Rev., vol. 31, no. 4, pp. 501–520, 2023, doi: 10.1093/medlaw/fwad013.

[7] H. Smith, "Artificial intelligence for clinical decision-making: Gross negligence manslaughter and corporate manslaughter," New Bioeth., vol. 30, no. 3, pp. 228–242, 2024, doi: 10.1080/20502877.2024.2416862.

[8] Y. A. Fahim, I. W. Hasani, S. Kabba et al., "Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives," Eur. J. Med. Res., vol. 30, Art. no. 848, 2025, doi: 10.1186/s40001-025-03196-w.

[9] M. N. Duffourc and R. A. Cahill, "Legal implications of AI standard of care integration on patients' informed consent: Lessons from surgery," npj Digit. Med., vol. 9, Art. no. 4, 2026, doi: 10.1038/s41746-025-02157-1.

[10] J. S. Schulz, "Resolution after medical injuries: Case studies of communication-and-resolution programs demonstrate their promise as an alternative to clinical negligence," Laws, vol. 14, no. 4, Art. no. 55, 2025, doi: 10.3390/laws14040055.

[11] M. Pruski, "AI-enhanced healthcare: Not a new paradigm for informed consent," J. Bioethical Inquiry, vol. 21, pp. 475–489, 2024, doi: 10.1007/s11673-023-10320-0.

[12] I. K. Ng, "Informed consent in clinical practice: Old problems, new challenges," J. R. Coll. Physicians Edinb., vol. 54, no. 2, pp. 153–158, 2024, doi: 10.1177/14782715241247087.

[13] M. Pallocci, M. Treglia, P. Passalacqua, R. Tittarelli, C. Zanovello, L. De Luca, V. Caparrelli, V. De Luna, A. M. Cisterna, G. Quintavalle, and L. T. Marsella, "Informed consent: Legal obligation or cornerstone of the care relationship?" Int. J. Environ. Res. Public Health, vol. 20, no. 3, Art. no. 2118, 2023, doi: 10.3390/ijerph20032118.

[14] P. Khobragade, A. Yerlekar, A. Titarmare, P. Ghutke, N. Purohit, A. B. Kotta, and P. K. Dhankar, "Precision medicine: Improving healthcare with data science and machine learning," in Precision Medicine: Improving Healthcare with Data Science and Machine Learning, P. Lokulwar, B. K. Verma, M. Sehgal, K. Kumar, and M. Kolhe, Eds. Bentham Science Publishers, 2026, pp. 189–219, doi: 10.2174/9798898814779126010013.

[15] O. Higgins and R. L. Wilson, "Integrating artificial intelligence (AI) with workforce solutions for sustainable care: A follow up to artificial intelligence and machine learning (ML) based decision support systems in mental health," Int. J. Ment. Health Nurs., vol. 34, Art. no. e70019, 2025, doi: 10.1111/inm.70019.

[16] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany et al., "Revolutionizing healthcare: The role of artificial intelligence in clinical practice," BMC Med. Educ., vol. 23, Art. no. 689, 2023, doi: 10.1186/s12909-023-04698-z.

[17] M. Buiten, A. de Streel, and M. Peitz, "The law and economics of AI liability," Comput. Law Secur. Rev., vol. 48, Art. no. 105794, 2023, doi: 10.1016/j.clsr.2023.105794.

[18] R. Agarwal, M. Bjarnadottir, L. Rhue, M. Dugas, K. Crowley, J. Clark, and G. Gao, "Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework," Health Policy Technol., vol. 12, no. 1, Art. no. 100702, 2023, doi: 10.1016/j.hlpt.2022.100702.

[19] V. A. Laptev, I. V. Ershova, and D. R. Feyzrakhmanova, "Medical applications of artificial intelligence (legal aspects and future prospects)," Laws, vol. 11, no. 1, Art. no. 3, 2022, doi: 10.3390/laws11010003.

[20] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho, "Pile of Law: Learning responsible data filtering from the law and a 256GB open-source legal dataset," arXiv preprint arXiv:2207.00220, 2022, doi: 10.48550/arXiv.2207.00220