

# Analyzing Professional Ethics of Physicians Using Online Patient Reviews: A Machine Learning Approach

Production and Operations Management  
1–22

© The Author(s) 2025

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/10591478251318885

journals.sagepub.com/home/pao



Kanix Wang<sup>1</sup> , Feng Mai<sup>2</sup> , Zhe Shan<sup>3</sup>, Dawei (David) Zhang<sup>4</sup> and Xiaosong (David) Peng<sup>4</sup>

## Abstract

The erosion of professional ethics in medicine has severe consequences for patients and society. Existing approaches often rely on retrospective analysis and lack the precision and timeliness needed to effectively identify and mitigate risks. Although patient online reviews offer a unique opportunity to proactively detect ethical issues by providing candid, unsolicited feedback on healthcare experiences, few studies have empirically established the link between patient reviews and ethical breaches in medicine. This research introduces a novel machine learning framework to derive text-based indicators of physicians' professional ethics using online patient reviews. Our approach leverages large language models to extract ethics-related comments and employs few-shot contrastive learning to train multilabel classifiers. Empirical validation studies suggest that the ethical indicators can help predict a wide range of adverse outcomes including drug-related deaths, disciplinary actions, malpractice claims, and rent-seeking behaviors. Our framework offers promising avenues for proactively managing ethical risks in healthcare and other professional services.

## Keywords

Professional Ethics, Social Media, Healthcare, Machine Learning, Natural Language Processing

Date received 31 January 2023; accepted 29 September 2024 after two revisions

Handling Editor: Responsible Data Science Special Issue Editors

## 1 Introduction

Professional ethics are of the utmost importance to health-care providers. While adherence to professional ethics is often assumed in most OM literature (Li et al., 2016; Wang et al., 2019), recent empirical studies show that breaches of medical ethics are not uncommon and can adversely affect patients (Zhao et al., 2022). In addition, the erosion of medical ethics has serious societal consequences. Between 2000 and 2020, over 270,000 people died of prescription opioid overdoses in the USA, and much of the blame has been attributed to the unchecked financial quid pro quo between pharmaceutical manufacturers and doctors who overprescribe drugs (Kornfield et al., 2022). These rent-seeking behaviors by doctors put their financial interests ahead of their patients' well-being. Yet, because of the long-standing culture of medical autonomy and self-regulation, early identification and intervention of ethical violations have been difficult (DuBois et al., 2019; Ham and Alberti, 2002).

As patients increasingly rely on online reviews to make informed decisions about providers, researchers have a unique

opportunity to explore the extent to which these reviews offer meaningful information about professional ethics. While previous research has primarily focused on the relationship between numerical star ratings and healthcare quality (Gao et al., 2015; Lantzy and Anderson, 2020; Lu and Rui, 2018; Saifee et al., 2020), the predictive utility of text comments left by patients remains largely unexplored, even though they may provide more in-depth and candid evaluations. Recent studies have highlighted the operational value of social media information in nonmedical contexts (Cui et al., 2018; Yan and Pedraza-Martinez, 2019) and demonstrated how text data

<sup>1</sup>Lindner College of Business, University of Cincinnati, Cincinnati, OH, USA

<sup>2</sup>Tippie College of Business, University of Iowa, Iowa City, IA, USA

<sup>3</sup>Farmer School of Business, Miami University, Oxford, OH, USA

<sup>4</sup>College of Business, Lehigh University, Bethlehem, PA, USA

## Corresponding author:

Xiaosong (David) Peng, College of Business, Lehigh University, Bethlehem, PA, USA.

Email: xip320@lehigh.edu

can be used for risk assessment (Wu, 2023). However, identifying meaningful signals from noisy reviews poses several challenges, such as dealing with colloquial language, ensuring model interpretability, and addressing the sparsity of relevant information. Given these considerations, our work focuses on two research questions: 1. How can we effectively extract information from online patient reviews about physicians' potential professional ethical issues? 2. What is the predictive value of the information for various ethics-related outcomes?

We propose a new natural language processing (NLP) approach to measure providers' adherence to professional ethics using online patient reviews. Drawing upon established frameworks in medical ethics literature, we identify 10 key dimensions that capture potential violations across the principles of deontology, utilitarianism, and emotivism (Lucey and Souba, 2010; Mandal et al., 2016; Rosenstein and O'Daniel, 2008). Deontology emphasizes adherence to codes or principles; utilitarianism focuses on outcomes that yield the highest net benefit; and emotivism views moral statements as expressions of personal emotions, not objective truths. Together, they inform guidelines that prioritize patient health and well-being while preserving the integrity of the medical profession. We collect a large dataset of patient reviews covering healthcare providers in the USA. Leveraging this unique data source, we develop a few-shot learning pipeline that fine-tunes a large language model (LLM) to accurately extract sentences within the reviews that pertain to these ethical dimensions. The extracted sentences serve as ethical indicators and provide quantitative measures of professional ethics adherence.

We conduct several validation studies to illustrate the practical relevance of these ethical indicators. First, we show that patient comments regarding controlled substance prescription are correlated with future drug-related deaths at aggregate local levels. Second, these indicators can predict physician sanctions by licensing boards. Third, they predict the type of injury and indemnity payments for malpractice claims. Fourth, they can help predict rent-seeking behaviors, such as healthcare providers accepting payments from pharmaceutical companies in exchange for prescribing drugs. Lastly, we extend our validation with a national clinical quality measure, revealing that while ethical indicators can predict physician quality, clinical quality alone does not predict future sanctions. Furthermore, we employ Explainable AI (XAI) methods to gain insights into our predictive algorithm and quantify the importance of each indicator. Finally, we discuss our framework's implications for the medical profession, healthcare managers, policymakers, and other professional service contexts.

Our study's primary contribution lies in bridging the literature on the operational value of social media data (Cui et al., 2018) and ethical risk management in healthcare. Specifically, we respond to the calls for proactive risk management (NEJM Catalyst, 2018) and the development of new measures that capture the interpersonal and dynamic processes (D. S. Kc et al., 2020) in healthcare. By highlighting the value of social media

data on detecting ethical lapses in medical practice—as manifested in the opioid epidemic and troubling cases of patient abuse (Kornfield et al., 2022; Whitaker, 2023)—we provide potential means to address isolated incidents before they escalate to more serious issues. This contrasts with prior healthcare operations studies that focus on singular interventions for specific problems (Bastani et al., 2019; Bobroske et al., 2022; Kc et al., 2022).

From a technical perspective, we develop an integrated machine learning framework grounded in theories. Our approach combines several methodological innovations to address the challenges in this task. Anchoring our framework in established moral philosophies enhances validity and credibility. To tackle data sparsity and class imbalance, we integrate active search strategies with contrastive learning techniques. We harness LLMs to bridge the gap between abstract ethical principles and colloquial patient language, enabling feature extraction aligned with ethical frameworks without extensive manual annotation. Importantly, we go beyond classification performance improvements by validating the predictive power of indicators against a spectrum of real-world outcomes and employing explainable AI methods to provide actionable insights. Our framework offers a promising avenue for managing similar issues across other professional services contexts.

Our third contribution lies in demonstrating the practical relevance of ethical theories to patient-centered healthcare quality. Whereas prior research focused on clinical outcomes and patient satisfaction (Nair et al., 2013), we argue that adherence to ethical principles is a crucial yet often overlooked dimension of patient-centered care. Ethical violations, unlike typical quality issues, often involve deliberate misconduct and tend to be low-probability but high-consequence events, aligning them more with risk management than quality control. By operationalizing ethical constructs into measurable indicators that predict tangible outcomes, we establish a direct connection between ethical principles and operational realities. This approach elevates the concept of professional ethics from abstract ideals into actionable elements of patient-centered care (Chandrasekaran et al., 2012).

## 2 Literature and Theoretical Background

### 2.1 Professional Ethics in Healthcare

Professional ethics refers to the values and principles that guide conducts in occupations characterized by high levels of autonomy and specialized knowledge (Chadwick, 2016). In healthcare, this autonomy is particularly pronounced, as workers exercise considerable discretion in their work (Kc et al., 2020). This high level of autonomy, combined with the high-stakes nature of healthcare, necessitates a strong ethical framework. Consequently, professional ethics in healthcare becomes a critical issue with far-reaching implications for patient outcomes, public trust, and the overall efficiency of the

system. A recent Gallup (2023) survey ranks medical professionals at the top among all professions in terms of perceived honesty and ethics, which highlights the exceptional ethical expectations placed on them.

However, despite this high level of trust, significant ethical challenges persist. The US healthcare system faces alarming rates of medical errors due to negligence (Bastani et al., 2019). A study reveals that nearly 20% of doctors have encountered impaired or incompetent colleagues over three years, yet many instances went unreported (Roland et al., 2011). Some medical systems failed to take appropriate action against egregious cases of sexually abusive doctors (Whitaker, 2023). Financial relationships between physicians and pharmaceutical companies have driven wasteful spending and contributed to the opioid epidemic (Kornfield et al., 2022). Conventional methods in ethical monitoring, such as whistleblowing (Blenkinsopp et al., 2019), auditing (Busch, 2012), and training programs (Jimenez and Foster, 1998) have shown limitations in effectively addressing ethical breaches. These approaches often rely on retrospective analysis of historical data and manual review processes, which can lack precision and timeliness needed to identify and mitigate ethical issues due to lengthy investigation processes, limited resources, and fragmented reporting systems (Kumaraswamy et al., 2022).

The healthcare operations management literature has focused primarily on clinical outcomes, measured through objective metrics like patient outcomes and guideline adherence (Chandrasekaran et al., 2012; Nair et al., 2013), and experiential aspects, assessed through patient satisfaction surveys (Peng et al., 2020). While several studies have examined the effects of process and policy changes on specific ethical issues such as opioid overuse and upcoding (Bastani et al., 2019; Bobroske et al., 2022; Kc et al., 2022), they tend to focus narrowly on the causal effects of singular interventions on singular problems. As a result, there is a lack of a comprehensive framework for assessing the broad spectrum of ethical risks. This gap is crucial given that ethical breaches, unlike quality problems, range from intentional misconduct to unintentional negligence and are more subjective in nature (Kaptein, 2008). Moreover, ethical breaches often result in low-frequency but high-impact events. Consequently, addressing these issues require shifting from traditional quality management to a new risk management paradigm.

## 2.2 Quantifying Ethical Risks Using Social Media Data

**2.2.1 Conceptual Framework: Patient-Centered Risk Perception and Unstructured Data Assessment.** Risk management in healthcare refers to systems and processes designed to detect, monitor, assess, mitigate, and prevent risks to patients (NEJM Catalyst, 2018). Conceptually, we situate our work through the lens of two key dimensions of operations risk management: *risk perception* and *risk assessment* (Cohen and Kunreuther, 2007).

Risk perception focuses on how different stakeholders understand, view, and act on risks. In the context of data science models for risk perception, this translates to selecting data sources that represent different stakeholder perspectives. We categorize existing work into *provider-centered* (or more broadly, business-centered) and *patient-centered* (or customer-centered) approaches. Provider-centered risk perception relies on healthcare organization data through formal channels, with most work focused on detecting fraud and misconduct (Bauder and Khoshgoftaar, 2018; Ekin et al., 2021; Herland et al., 2018; Kumaraswamy et al., 2022). However, these approaches have limitations, as they can be neutralized by organizational cultures where providers address issues independently rather than through official channels (Blenkinsopp et al., 2019). In contrast, patient-centered risk perception focuses on the experiences and perceptions of healthcare consumers, often expressed through unofficial channels such as social media. The literature on online patient reviews has shown mixed results regarding their usefulness in assessing quality and risk. Some studies (Gao et al., 2015; Lantzy and Anderson, 2020; Lu and Rui, 2018) find that online ratings can provide valuable insights into patient experiences. Others caution that reviews may not accurately reflect all aspects of care, particularly for services with credence attributes (Saifee et al., 2020).

Risk assessment involves evaluating the likelihood and consequences of risks using data, expert judgments, and probabilistic methods. These models can be categorized into those using *structured* data such as sales, inventory, and electronic health records (Ekin et al., 2021; Herland et al., 2018; Markou and Corsten, 2021), and those that rely on *unstructured* data such as text (Abrahams et al., 2015; Liu et al., 2023; Wu, 2023). Structured data is easier to integrate into existing ERP and business intelligence systems (Araz et al., 2020), thus enabling more precise risk assessment. Conversely, unstructured data captures specifics of risk scenarios that might not be evident in structured data, thus enabling greater granularity of risk assessment (Wu, 2023). Text analytics can add significant value to predictive models for ethical and compliance monitoring, as it combines two important mechanisms of leveraging big data (Cohen, 2018). Crucially, in our context, unstructured data can contain subtle indicators of professional misconduct that are often difficult to reflect in existing structured data collection systems.

Our work belongs to the quadrant of patient-centered risk perception using unstructured text data for risk assessment (Table 1). Among this quadrant, the conceptual novelty of our work lies in leveraging unstructured online patient reviews to quantify ethical risks in healthcare, a hitherto overlooked link. Given conflicting evidence on the value of patient reviews and anecdotal reports of patients inadvertently rewarding unethical practices (e.g., giving high ratings to those who are willing to write opioid prescriptions) (Macy, 2018), it is imperative to empirically test this link. By mining reviews for ethics-related

**Table 1.** Key literature and positioning of our approach.

		Risk assessment	
		Structured data	Unstructured data
Risk perception	Provider/business-centered	Bauder and Khoshgoftaar. (2018); Ekin et al. (2021); Herland et al. (2018, 2019); Kumaraswamy et al. (2022); Markou & Corsten (2021)	Wu (2023); Liu et al. (2023)
	Patient/customer-centered	Lantzy & Anderson (2020)	This paper Yang et al. (2014); Abrahams et al. (2015); Abbasi et al. (2019); Mejia et al. (2021); Xie et al. (2021); Zhang et al. (2022); Li et al. (2023)

comments and evaluate their predictive power for multiple outcomes, our approach aligns with emerging trends in healthcare risk management that shift from reactive strategies towards more proactive methods that consider risk across the entire ecosystem (NEJM Catalyst, 2018).

**2.2.2 Methodological Contributions: Addressing Challenges in Ethical Risk Quantification.** The patient-centered, unstructured data approach outlined in our conceptual framework introduces unique challenges. First, extracting features anchored in ethical theories is crucial for model interpretability and actionable insights. This approach enhances conceptual validity by aligning measurements with established moral values, lending credibility when engaging with interdisciplinary audiences, practitioners, or policymakers. Anchoring features in ethical theories ensures that quantified risks reflect principled moral reasoning rather than mere statistical artifacts. Second, the heterogeneous and nuanced expressions of ethical concerns in patient reviews require advanced natural language understanding (NLU) capabilities. The model must bridge the gap between abstract ethical principles and diverse, colloquial patient expressions to accurately map narratives onto ethical constructs. Third, the unstructured nature of text data results in sparsity within the input space, while the infrequency of ethical complaints leads to imbalanced output labels. This combination of sparse inputs and skewed outputs presents challenges for model training. An integrated approach is needed to handle both dispersed signals and uneven class distribution while balancing the identification of rare ethical violations against overfitting risks. Fourth, demonstrating real-world relevance requires evidence that patient reviews can predict multiple ethics-related outcomes while providing interpretable insights. Ethical violations in healthcare can lead to diverse adverse consequences, e.g., patient harm, legal liabilities, emotional distress, and financial losses. In this high-stakes context, the principles of Explainable AI, i.e., the models are understandable, justifiable, and actionable, are crucial (De Bock et al., 2023).

We compare our study with the literature along these dimensions (see Table EC.1).<sup>1</sup> Most studies (except Zhang et al., 2022) do not have a theoretical foundation underpinning their model architectures. Regarding NLU models, prior work has employed lexicon-based approaches (Abbasi et al., 2019; Abrahams et al., 2015; Yang et al., 2014), static word embeddings (Wu, 2023; Xie et al., 2021; Zhang et al., 2022), topic models (Ko et al., 2019), and LSTM (Liu et al., 2023). While effective in their domains, these models have limitations in our context. Lexicon-based methods struggle with diverse patient language; static embeddings fail to capture context; and topic models lack granularity for specific ethical issues. Deep learning models like LSTMs require large, labeled datasets, but obtaining these is difficult given the sparsity of ethical concerns. To address sparsity, some studies use heuristics (Abbasi et al., 2019; Xie et al., 2021), data augmentation (Li et al., 2023), or undersampling (Zhang et al., 2022). However, while some approaches handle imbalanced target classes, acquiring sufficient training data remains challenging due to sparse ethical expressions. In terms of real-world relevance and interpretability, while existing studies demonstrate superior classification performance against benchmarks, most either lack real-world outcome prediction beyond test-set documents or focus on a single outcome type.

Our study contributes novel solutions to address these limitations. First, we ground the measurement of professional ethics in the “big three” ethical theories.<sup>2</sup> Integrating them into the same measurement framework allows us to compare and contrast their practical utility in assessing risks. Second, we leverage the representational power of LLMs to understand the semantics of ethical concerns expressed in patient’s words, going beyond the limitations of NLU techniques employed in most extant studies. Third, facing more extreme sparsity issues, we devise a two-pronged solution: active search to efficiently discover sparse signals, and contrastive learning to further amplify these signals. This combination of techniques is novel in the literature. Finally, our work goes beyond simply demonstrating superior classification performance. We validate the predictive power of ethical indicators against a much

**Table 2.** Examples of ethical indicators based on three ethical theories.

Theory	Ethical indicator	Example sentences
Deontology	Improper prescribing of controlled substances (IPCS)	“Reputation for prescribing controlled drugs.” “He pushes narcotics for pain.”
	Sexual abuse of patients (SAP)	“She said she really wanted me to have the drug and wanted to help out her drug rep.” “Made sexual advances that were inappropriate and unwelcomed.” “He made inappropriate remarks, one being of a sexual nature.” “Abusive and completely inappropriate.”
	Unnecessary invasive procedures (UIP)	“Like another patient reported, was touched inappropriately.” “Beware of diagnosis as this doc seems to be interested in making money from surgeries.” “Finally, he managed to do an awful lot of work of questionable necessity, some of which was high risk.”
	Negligence or incompetence (NI)	“Pressures patient into what seems to be an unnecessary invasive procedure.” “Not knowledgeable enough to call himself a specialist.” “Failed to inform a patient of a disease that the doctors tests discovered and failure to provide treatment and referrals necessary for the patient to conquer his disease.” “Was negligent that the proper medication wasn’t being administered and that orders weren’t being followed.”
	Attitudinal or communication unprofessionalism (ATT)	“I would never want to be treated by a doctor who doesn’t extend the smallest kindness or compassion to someone who is seeking his services.” “He just doesn’t care or listen.”
	Fraud or inappropriate billing practices (FIBP)	“This doctor clearly lacks human sympathy and empathy.” “They abuse the medical billing system and take advantage of patients.” “Insurance billing is done incorrectly, often the same visit is billed multiple times and then patients are billed for amounts they do not owe.” “Billing process is unbelievable and they have no clue how insurance companies work or what their procedures cost.”
Utilitarianism	Positive cure (POSC)	“Saved my life and stabilized me prior to transfer to ICU” “Kept me alive for over 20 years with a heart condition” “Helped me right away in an emergency”
	Negative cure (NEC)	“Left me in pain for almost a year.” “He miscalculated the dosage and caused me severe nerve damage AND a reaction to the meds.” “Will say anything to minimize the deleterious effects of injury that he causes.”
Emotivism	Love (LOV)	“what an awesome doctor” “amazing doctor all around” “I love this doctor”
	Hate (HAT)	“He is the worst doctor I have ever seen in my life.” “I cannot express enough hatred I have for this doctor.” “there are not words to describe my disdain for this so called dr”

broader set of real-world outcomes compared to existing work and provide rich interpretable insights using XAI methods.

### 2.3 Theoretical Background: The Three Lenses of Ethical Theory

We draw upon three ethical theories to construct a set of professional ethics measures anchored in the patient experience. Table 2 presents a summary of these measures and example reviews.

The first theoretical lens, *deontology*, concerns the ethics of duty, that is, what one person should or should not do in

relation to another (Garbutt and Davies, 2011). This theory is rooted in the belief that any act can be judged on its own merit, rather than outcome (Gal et al., 2022). A basic criterion for such assessment is whether the act conforms to a moral norm. In healthcare, deontology is strongly reflected in professional codes of conduct (Fineschi et al., 1997). These codes of conduct often outline the moral and legal obligations that healthcare professionals have to their patients. They also prohibit specific violations and behaviors that are deemed unacceptable by the medical community, legal system, and society. Accordingly, we develop six deontological measures. While different medical specialty associations define different codes, the literature suggests that the three most commonly reported

violations are improper prescribing of controlled substances (IPCS), sexual abuse of patients (SAP), and unnecessary invasive procedures (UIP) (DuBois et al., 2019). These serious violations can have severe consequences for patients, including opioid addiction and overdose deaths, as well as physical, emotional, and financial harm. Additionally, a national study indicates that other three common reasons for disciplinary action by medical boards are negligence or incompetence (NI), attitudinal or communication unprofessionalism (ATT), fraud or inappropriate billing practices (FIBP) (Papadakis et al., 2005). Table EC.1 provides references to repercussions associated with these deontological indicators.

The second theoretical lens, *utilitarianism*, holds the best action is the one that promotes overall better consequences. In healthcare, utilitarianism reflects the outcome of the treatment from the patient's perspective, which may include the impact that a diagnosis or treatment has on the patient's overall well-being and quality of life. In other words, outcomes determine the morality of the intervention (Mandal et al., 2016). This lens can be used as an indicator of physician professional ethics, as it reflects the extent to which doctors are fulfilling their ethical obligations to provide the best possible care. We thus develop two utilitarianism measures: positive care (POSC) and negative care (NEC). They capture the degree to which patients believe that their conditions have improved or deteriorated from treatment.

The third theoretical lens, *emotivism*, holds that moral judgments express positive or negative feelings rather than relying on reasoning or objective evaluation (Bandman, 2003). In healthcare, emotivism reflects how patients feel about their providers and treatments. Though seemingly more subjective than other lenses, emotivism is crucial in patient-centered healthcare as it acknowledges the importance of emotional reactions to healthcare experiences (Husted and Husted, 2005). Patient satisfaction and emotional well-being are important quality indicators (Manary et al., 2013) and can impact clinical outcomes (Robertson et al., 2012). Emotions and sentiments are particularly important on social media channels for providing operational feedback (Cui et al., 2018; Gour et al., 2022). Patients' emotions expressed in reviews, even when nonspecific, can still be related to the professional ethics of doctors. For example, patients' expressions of frustration and complaints may indicate a disregard of respect and empathy. Based on this theory, we develop two measures: love (LOV) and hate (HAT), measuring the degree of positive and negative feelings toward providers respectively.<sup>3</sup>

### 3 Data

Our primary data source is a set of patient reviews from Vitals (<http://www.vitals.com>), a popular online platform for healthcare provider information. We used a custom web crawler to extract 1,167,455 reviews from July 2005 to December 2014. These reviews cover 449,116 unique providers across various medical specialties. The dataset provides broad coverage of

all 50 US states and Washington, D.C. Each review entry contains three key components: an ordinal star rating (1–5 stars), physician metadata (e.g., years of experience, specialty, and gender), and free-text comments explaining the ratings. The textual component averages 52 words per review. We present detailed summary statistics of the patient review data in Table EC.4. The credibility of the data source stems from Vitals' quality control measures. These include mandatory board certification for listed physicians and a rigorous review authentication process. Vitals also imposes a 30-day cooldown period between submissions from the same reviewer. Anonymity on the platform may reduce social desirability bias and encourage candid feedback.

To validate our ethical indicators and demonstrate their practical utility, we employ several additional datasets. First, we use state- and county-level drug poisoning mortality data from the Centers for Disease Control and Prevention (CDC). This dataset allows us to quantify opioid-related deaths across all 50 states and Washington, DC from 1999 to 2016 (Giles et al., 2023; Janssen and Zhang, 2023).

Second, we collect physician discipline information from 21 state medical boards' public websites. We convert these state-specific records into a machine-readable format through extensive data processing.<sup>4</sup> We present detailed characteristics of these disciplinary records in Table EC.5.

Third, we use physicians' payment-prescription sensitivity as a measurement for rent-seeking behavior (Parsons et al., 2018). This measure reflects the relationship between payments made by pharmaceutical firms to doctors and the value of prescriptions written by those doctors for drugs produced by those firms. To construct this measure, we combine two datasets. The first is ProPublica's Dollars for Docs data, which contains the payments made by these companies to providers in the form of speaking fees, consulting fees, dinner, etc. from August 2013 to December 2016. The data is compiled from publicly available information, including Open Payments data from the Centers for Medicare and Medicaid Services (CMS) and voluntary disclosures from the firms. We then collect the Medicare Part D Prescriber Data provided by CMS to merge with the Dollars for Docs data. The Prescriber Data constitutes detailed records of prescription drugs in the Medicare Part D program. The data contains the National Provider Identifier (NPI) of the healthcare provider, the prescribed drug, the brand names (which we use to identify manufacturers), and the cost incurred under Medicare. Prior literature has shown that this data can be a useful resource for understanding the conflicts of interest that may exist between healthcare providers and the industry (Brennan et al., 2006).

For a given doctor-firm pair, we aggregate the payment amount and prescription cost. The payment data ( $pay_{ij}$ ) is the total amount of money spent on doctor  $i$  by firm  $j$ . The prescription cost ( $drug_{ij}$ ) is the total value of prescriptions written by doctor  $i$  for all the drugs produced by firm  $j$ . The payment-prescription sensitivity ( $Sensitivity_i$ ) for doctor  $i$  is the Pearson correlation between two vectors  $[pay_{i1}, pay_{i2}, \dots, pay_{iJ}]$  and

$[drug_{i1}, drug_{i2}, \dots, drug_{iJ}]$ , when the doctor received payments from  $J$  firms. If the correlation is positive, it indicates that the doctor is more likely to write prescriptions for drugs from firms from whom they have received payments. The inclination to prescribe drugs in exchange for financial rewards signals a potential rent-seeking behavior. A similar payment-prescription sensitivity measure is shown to be highly correlated with other financial misconducts at the city level, such as political corruption (Parsons et al., 2018). Table EC.6 presents summary statistics for this measure.

Fourth, as a measure of more detailed financial and legal repercussions of medical misconduct, we access the Professional Liability Tracking Database from the state of Florida. The database contains 35,632 claims linked to Florida state licenses, and provides more granular information than state medical board sanction records, including injury classifications (permanent, temporary, or emotional) and indemnity payment details.<sup>5</sup> These payments comprise settlements, court-ordered judgments, and associated legal fees covered by insurers, thus allowing us to analyze the spectrum of patient-reported adverse events and their associated costs. Table EC.7 details the claims' characteristics.

Fifth, to validate our ethical indicators in the context of clinical quality, we utilize measures from the CMS Merit-Based Incentive Payment System (MIPS). Established by the Medicare Access and CHIP Reauthorization Act (MACRA) in 2017, MIPS evaluates clinician performance on a scale from 0 to 100 based on CMS-approved criteria. The criteria encompass preventive care, chronic disease management, care coordination, patient safety, patient engagement, and efficient use of clinical resources. MIPS assesses performance across three domains: quality, promoting interoperability, and improvement activities. Our analysis focuses on the quality component of the MIPS scores from 2018, as it is the first year with stable enrollment following the inaugural year of 2017.

We merge the above datasets (see Figure EC.1) by first matching the doctor records in the review dataset with the Medicare provider data from CMS, using doctors' full names and the city to locate their NPI. We acknowledge that mapping the review dataset to the NPI can be noisy. Nevertheless, we employ a stringent criterion requiring both name and city matching. For the majority of doctors (65.91%) in the review dataset, we are able to find exact matches. To mitigate the issue of unmatched NPIs, we conducted a robustness test in Section 5.3, which demonstrates that our model's performance remains robust even in the presence of unmatched NPIs. We then merge the Dollars for Docs and Medicare Part D Prescriber data using NPI. For physician disciplinary records, we use NPI when available. In cases where NPI is unavailable, we match records using the physician's name and licensed state. We also utilize state license numbers and other identifying information for manual deduplication when necessary. We standardized names, states, and other entities during the merging process to ensure consistency.

## 4 Methods

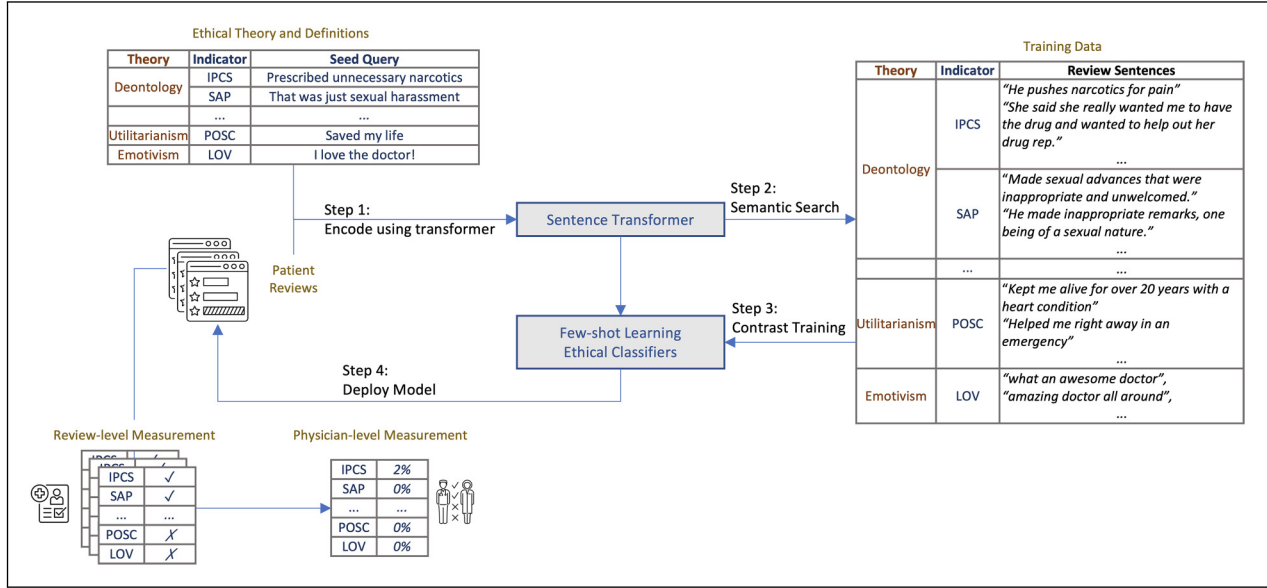
### 4.1 Overview of Methods

Figure 1 presents our framework's flowchart addressing the challenges laid out in Section 2.2.2. Our approach begins by grounding the analysis in the three theoretical lenses of professional ethics—deontology, utilitarianism, and emotivism—and identifying 10 dimensions that embody these principles (detailed in Section 2.3). To operationalize these dimensions, we first develop a set of generic, representative *seed* statements that depict the corresponding ethical standards (see Table EC.9). For example, we use “*prescribed unnecessary narcotics*” for improper prescription of controlled substances (IPCS), and “*saved my life*” for positive care (POSC).

Our next goal is to construct a training dataset. This task presents a challenge due to the *sparsity* of ethics-related concepts in the web-scale review corpus. Annotating a random sample as a training set would likely yield very few, if any, review sentences directly related to the ethical indicators. To overcome this challenge, we leverage the representation power and NLU ability of LLMs to conduct a cost-effective *active search* (Jiang et al., 2019)—a labeling strategy that targets the identification of positive examples within a large unlabeled dataset. Active search is a specialized form of active learning (Saar-Tsechansky and Provost, 2007) designed for highly skewed class distributions. It focuses on the maximization of minority class discovery. Our approach is partly inspired by Coleman et al. (2022), which prioritizes the nearest neighbors of currently labeled examples to enhance efficiency by avoiding exhaustive scans of all unlabeled data. We adapt this idea from image to text domain by using semantic search, an information retrieval technique that identifies relevant documents matching the meaning of a query, to find the nearest neighbors of labeled ethics-related reviews.

Furthermore, we encounter a *multiclass few-shot learning* problem. Each review sentence may pertain to multiple ethical indicators, with some indicators having a limited number of positive sentences even after employing an active search labeling strategy. Although GPT-style models are technically feasible for such tasks (Brown et al., 2020), their associated cost and latency render them unsuitable for our context. We opt for a *contrastive learning* approach that can outperform GPT-style models in few-shot classification tasks (Tunstall et al., 2022). Contrastive learning focuses on improving the representation ability of a model, so that they can better distinguish between similar and dissimilar pairs of sentences. After contrastive learning, a classification head is then trained on top of a frozen transformer model (i.e., their parameters are not updated during training). Research has shown that freezing the parameters in a transformer model can improve the robustness of the model, particularly when distribution shift is a concern such as in few-shot learning (Kumar et al., 2021).

Finally, as our eventual goal is to enable regulators and healthcare managers to transform online review data into insights, establishing clear reasonings for decision-making



**Figure 1.** Analysis flowchart.

is crucial. Rather than developing an end-to-end model for directly predicting real-world ethical violations from review texts, we first construct review-level measurements and subsequently assess their predictive power in downstream tasks. This approach provides a clearer understanding of the model's mechanisms; it also enables the generation of theoretical and managerial insights into which types of ethical indicators are most indicative of actual behavior. We next describe the implementation of each step in detail.

## 4.2 Constructing Curated Dataset Using Semantic Search With Sentence-Transformer

**Step 1: Encode seed sentences in the same semantic space as the patient review text.** We employ a sentence-transformer model (all-mpnet-base-v2) (Song et al., 2020), to encode both seed statements and patient reviews. This sentence-transformer model is built upon the MPNet architecture, a pre-trained encoder-only LLM that refines the widely-used BERT model (Devlin et al., 2019). It is specifically optimized for semantic textual similarity tasks. While BERT-like encoders excel at various NLP tasks, they often struggle to capture subtle semantic differences in text. Sentence-transformer models are fine-tuned on specialized datasets like human-annotated text similarity corpora to overcome this limitation (Reimers and Gurevych, 2019).

**Step 2a: Semantic search.** We adopt an active search strategy within the patient review corpus. This addresses the challenge posed by the scarcity of explicit ethical content in these reviews (Jiang et al., 2019). In the first step, we embed all reviews and seed statements into sentence vectors using the sentence-transformer. We then index these vectors using a semantic search engine to facilitate semantic comparison.

The search utilizes cosine similarity to measure the closeness between the embedding vectors of our seed sentences and sentences in the patient review corpus. This approach bypasses the limitations of traditional keyword matching by prioritizing semantic relevance over exact term alignment. Given a query, the semantic search engine yields a ranked list of review sentences based on their semantic pertinence to our queries.

For each of the 10 ethical indicators, our initial queries are a set of seed statements that represent unethical or unprofessional physician behaviors (see Table EC.9). Guided by these seed statements, we interactively search all review sentences to find those pertinent to the ethical notions, such as "Dr. wanted to perform unnecessary procedures," "gross negligence of patient care," "deliberately misleading and unethical billing practices," "kept me alive for over 20 years with a heart condition," or "caused nothing but pain and agony." This step essentially bridges the linguistic gap between the more professional terminology of the seed statements and the more colloquial language of patient reviews.

**Step 2b: Dataset finalization.** The construction of the dataset involves a multi-round, iterative active search process. In the first round, the top 100 review sentences most semantically similar to the seed statements are manually annotated for relevance to the corresponding ethical indicator. The sentences deemed relevant are then used as additional queries in the second round of search, along with the original seed statements. This process is repeated for a third round, with the relevant sentences from the second round serving as additional queries. The iterative process allows for the discovery of a wider variety of relevant expressions that patients use to describe ethical concerns, which may not be captured by searching the initial seed statements alone. The result is a collection of 150–300



manually validated sentences for each ethical indicator, totaling 1729 sentences across all indicators. We also add 500 randomly drawn negative examples per ethical indicator from the review corpus to form the final dataset.

To address potential underreporting of ethical concerns, especially for sensitive issues like sexual misconduct, our data collection strategy captures a diverse set of training sentences covering a wide range of misconduct behaviors. For instance, we include sentences such as “*made me uncomfortable by stating inappropriate comments and touching,*” and “*inappropriate behavior, made me feel uncomfortable as a female patient*” to capture subtle forms of sexual misconduct. As a result, the incidence rate of sexual misconduct indicators (SAP) is of similar magnitude (0.18%) as fraud and improper billing (FIBP) 0.22%, improper prescription of controlled substance (IPCS) 0.19%, hate (HAT) 0.19%, or negative care (NEC) 0.20% in the sanctioned cases (see Table EC.4). This suggests our data collection strategy effectively captures a representative distribution of ethical violations, even for sensitive issues prone to underreporting.

### 4.3 Training and Deploying Classifiers

*Step 3: Few-shot learning using contrastive training.* To construct a classifier for ethical indicators, we turn to few-shot learning, a methodology that trains language models for classification tasks using a relatively small number of labeled examples. This approach has been effective in various natural language processing applications, such as text classification and question answering (Brown et al., 2020). Specifically, we employ the SetFit method. SetFit conducts contrastive fine-tuning of pretrained sentence embeddings, and is shown to be more efficient than traditional fine-tuning and GPT-style in-context learning methods (Tunstall et al., 2022).

SetFit employs a two-stage training approach for the development of a professional ethics indicator classifier. Initially, the sentence-transformer (ST) is fine-tuned using 80% of the curated dataset from step 2. This contrastive fine-tuning process enhances the embedding model’s discriminative capabilities between different ethical classes. Subsequently, to avoid overfitting and maintain the stability of the embeddings, the transformer is frozen while a classification head is trained on these embeddings. Classifier performance is assessed using the remaining 20% of the data as a hold-out set.

Specifically, in the contrastive fine-tuning phase, given a small set of  $K$  labeled examples  $D = \{(x_i, y_i)\}$  representing input sentences and their corresponding class labels, the model generates sets of  $R$  positive and negative triplets for each class label  $c \in C$ . Positive triplets, denoted as  $T_p^c = \{(x_i, x_j, 1)\}$ , comprise pairs of sentences randomly selected from the same class  $c$ , while negative triplets,  $T_n^c = \{(x_i, x_j, 0)\}$ , consist of sentences from different classes. For example, a positive pair might include sentences like “*unnecessary medical procedures were performed*” and “*pressures patients into an unnecessary invasive procedure,*” both indicating UIP practices. The

model learns to bring the embeddings of these sentences closer in the vector space. Conversely, the model will distance the embeddings of a negative pair, comprising either one sentence about one type of unethical practice and another random sentence, or two different types of unethical practices. The resulting contrastive fine-tuning dataset  $T$  is formed by concatenating these triplets across all class labels:  $T = \{(T_p^0, T_n^0), (T_p^1, T_n^1), \dots, (T_p^{|C|}, T_n^{|C|})\}$ , where  $|C|$  is the number of class labels. The total number of pairs  $|T|$  equals  $2R \cdot |C|$ . This method of data creation effectively expands the data in few-shot settings, as the potential size of the ST fine-tuning set  $T$  is  $\frac{K(K-1)}{2}$ , significantly larger than  $K$ . As a result, contrastive training amplifies the learning signal from each example. The model becomes more adept at recognizing and distinguishing between various ethics-related expressions in patient reviews.

After the ST is fine-tuned contrastively, the original labeled training data  $x_i$  is encoded into sentence embeddings per training sample,  $Emb^{x_i} = ST(x_i)$ . These embeddings and their class labels form the training set for the classification head  $T^{CH} = \{(Emb^{x_i}, y_i)\}$ . The classification head is tailored for multiclass-multilabel classification. It essentially applies binary logistic regression classifiers to multi-target classification, fitting one classification head per ethical indicator. This allows each sentence to be associated with multiple ethical indicators.

The trained model is made accessible on the HuggingFace platform.<sup>6</sup> For training, we set epochs to 4 and use an Adam optimizer with a  $2e^{-05}$  learning rate. A learning rate scheduler is employed for linear rate increases during the warmup phase. To prevent overfitting, the model incorporates a weight decay of 0.01 as a form of L2 regularization. Full hyperparameter details are available in Table EC.10.

*Step 4: Deploy the classifiers for downstream tasks.* Once the multiclass professional ethics indicator classifier is trained, it is deployed on all review sentences. Each sentence in the text is then labeled as either related to one of the 10 ethical indicators or not. Depending on the downstream task, the ethical indicators are aggregated either at the physician level or at the state/county-year level.

To improve the prediction accuracy of physician sanction status and payment-prescription sensitivity, we turn to XGBoost, a machine learning technique that aggregates the predictions of multiple decision trees (Chen and Guestrin, 2016). In economics and operations management, XGBoost has been used to predict police misconduct (Chalfin et al., 2016) and has demonstrated superior performance compared to other tree-based algorithms and machine learning approaches (Krauss et al., 2017; Mohri et al., 2018). Other benchmark techniques, e.g., random forests, decision trees, logistic regression, linear regression, and ridge regression, are also tested (see Section EC.3 for more details).

We split our data into an 80% training sample and a 20% test sample and use five-fold cross-validation to fine-tune the hyperparameters of machine learning models. To mitigate the effects of this rarity of sanctions, we adopt class

weighting and the Synthetic Minority Over-sampling Technique (SMOTE). To measure the performance of classification tasks, we report the out-of-sample receiver operating characteristic (ROC) curve's area under the curve (AUC), precision, recall, and F1 score. Measures for regression tasks, for example, MSE, RMSE, and MAE, are also computed.

We use a common XAI method, SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to interpret the trained XGBoost models. SHAP values are assigned to each feature based on the classic Shapley value from game theory. The values allow us to rank ethical indicators based on their predictive power. SHAP uses the conditional expectation  $E_{X_j|X_{-j}}(\hat{f}(x)|x_j)$  to estimate the effects of a feature  $x_j$  that contributes to pushing a model's output,  $\hat{f}(x)$ , away from its base value. It is defined as the difference between the expected output of the model with the feature absent  $X_{-j}$  and the current output of the model with the feature present  $X_j$ . We provide local SHAP figures to illustrate how the model uses various ethical indicators to make predictions for each physician. When applied across all physicians for all features, SHAP values also explain the features' global importance.

## 5 Results

### 5.1 Measurement: Search and Classification Performance

We benchmark the encoder's search performance against human annotation following the information retrieval literature. We compare the all-mpnet-base-v2 sentence-transformer, applied herein, against two other encoding models. The first, TF-IDF, assigns weights to words based on their frequency within a specific document relative to their frequency across an entire corpus. This provides a sparse vector representation for each review sentence. The second, word2vec, employs a single-hidden-layer neural network to derive word associations from the review corpus. Its outputs are static word embeddings that encapsulate semantic relationships between words (Mikolov et al., 2013). We train a 400-dimensional word2vec model on the review corpus for five epochs; the hyperparameters are set to default in the *gensim* package (Rehurek and Sojka, 2011). We represent each review sentence using the mean pooling of its word2vec word embeddings.

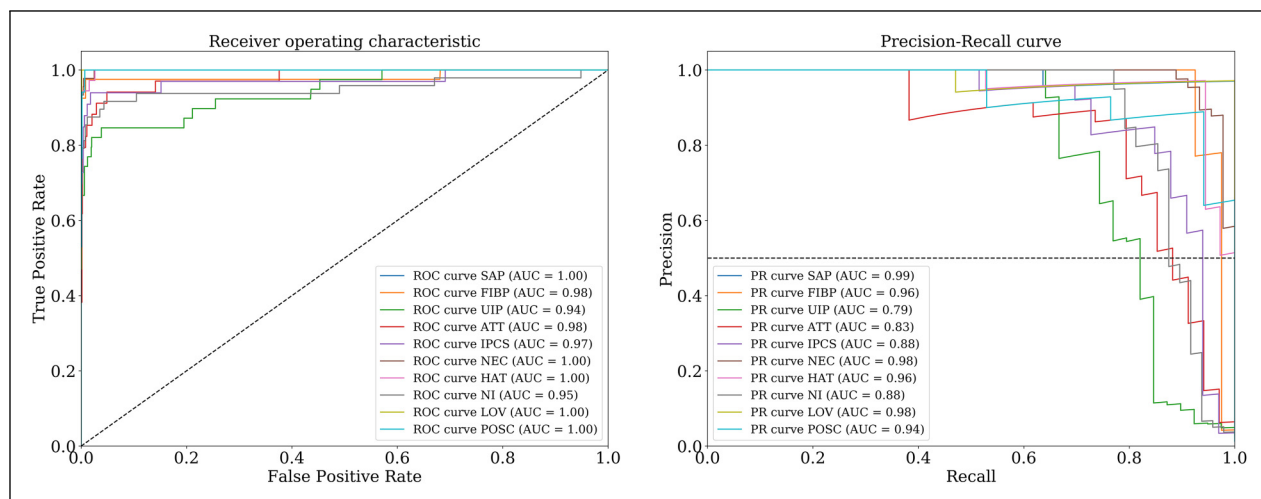
Given the three encoders, we process all reviews and seed statements, and retrieve the top 20 sentences ranked by each encoder. We subsequently shuffle these sentences randomly and request that two human annotators assess the relevance between each query and the corresponding sentences. The annotators use a 1–3 scoring system and take into account the general context and the specific definition of each ethical indicator. We use Normalized Discounted Cumulative Gain (NDCG) to evaluate the ranking efficacy of three encoder models.<sup>7</sup> We find that the sentence-transformer outperforms both TF-IDF and word2vec across all ethical indicators as measured by its concordance to the human annotators and

the NDCG scores (Table EC.11). The sentence-transformer's superior performance may be ascribed to its refinement through human-annotated paraphrase pairs and similarity data, which more closely resonates with the task at hand. The only cases when the benchmark models show comparative performance to the sentence-transformer are emotivism categories (love and hate). This result may stem from the less intricate nature of emotivism than deontology and utilitarianism indicators, where the need to comprehend contextual meanings is more challenging for simpler models.

We find that when processing indicators like improper prescription of controlled substances (IPCS), word2vec captures general prescription and controlled substances elements but struggles with the nuances. For example, it often provides irrelevant results such as strict prescription practices or patients being treated as addicts. A word2vec IPCS search yields matches like *"He believes that controlled substances should never ever be prescribed,"* which does not accurately reflect IPCS. Similarly, for sexual misconduct indicators, alternative methods find it challenging to distinguish between patients recounting past abuse experiences and actual complaints of physician misconduct. A notable misalignment can be seen in sentences like *"what she didn't bother to find out was that I was a victim of sexual abuse"* being erroneously matched to the sexual abuse by physicians (SAP) category. These patterns are consistently observed across various ethical indicators. They indicate the limitations of simpler encoding models in representing complex ethical notions.

In terms of classification performance, our few-shot learning model consistently achieves impressive results across all indicators (average ROC AUC 0.98, minimum ROC AUC 0.94, maximum ROC AUC 1.0, average PR AUC 0.92, minimum PR AUC 0.79, maximum PR AUC 0.99, see Figure 2).<sup>8</sup> This model's high performance is also evident through various evaluation metrics such as precision, recall, and F1 score (Table 3). The ROC curves demonstrate the model's strong ability to distinguish between various cases of ethical notions and negative examples (Figure 2). The precision, recall, and F1 scores provide a more detailed view of the model's performance. Given these strong metrics, we conclude that the model can accurately identify ethics-related comments with a high degree of accuracy.

To benchmark our classifier, we compare it against several other models (details in Table EC.12). These benchmarks span a range of text classification techniques: a naive BERT multiclass single-label classifier, a BERT-based embedding with an RNN for a single-label classification head, a word2vec embedding with a single-class XGBoost classifier, and a TF-IDF vector-based single-class XGBoost classifier. The results demonstrate that our few-shot learning approach achieves superior or comparable results in all dimensions. Another advantage of our model is its ability to quickly adapt to new data and improve its performance over time (Tunstall et al., 2022). By actively selecting and labeling a small number of



**Figure 2.** ROC and PR curves for 10 professional ethics indicators.

**Table 3.** Performance of the few-shot learning model in extracting ethical indicators.

Ethical indicator	Precision	Recall	F1 score
Improper prescribing of controlled substances (IPCS)	83.0%	76.4%	79.2%
Sexual abuse of patients (SAP)	94.3%	100.0%	97.1%
Unnecessary invasive procedures (UIP)	90.1%	67.3%	76.4%
Negligence or incompetence (NI)	100.0%	75.0%	85.7%
Attitudinal or communication unprofessionalism (ATT)	88.5%	67.7%	76.7%
Fraud or inappropriate billing practices (FIBP)	97.3%	92.5%	94.9%
Positive care (POSC)	86.7%	76.5%	81.3%
Negative care (NEC)	97.6%	88.9%	93.0%
Love (LOV)	97.1%	100.0%	98.6%
Hate (HAT)	97.1%	94.4%	95.8%

examples, our model is able to learn and make accurate predictions on new data more efficiently than a traditional model. This makes it an especially useful tool for identifying and addressing new forms of ethical violations if the need arises.

## 5.2 External Validations and Prediction Performance

**5.2.1 Prediction of State-Level and County-Level Drug Poisoning Mortality.** To illustrate the aggregate-level predictive power of our ethical indicators, we conduct a panel-data regression analysis to predict state-level drug poisoning deaths using improper prescription of controlled substances (IPCS) as our main independent variable. We merge state-level drug poisoning data from the CDC with our aggregated, state-level IPCS measures to obtain a matched panel of 50 states and Washington DC across 7 years (2008–2014). We use feasible generalized least squares (FGLS) to estimate the relationship between IPCS and drug poisoning deaths, adjusting for heteroskedasticity. We include year and state fixed-effects to account for time-invariant and state-invariant factors, and use the one-year lag of IPCS to mitigate simultaneity concerns. We also control for state-level per-capita income, population,

and a time trend. The results (Table 4) show the IPCS coefficients are highly significant and consistent across models, with the lagged model indicating one more IPCS-related comment is associated with about two more drug poisoning deaths, representing a strikingly strong predictive relationship.

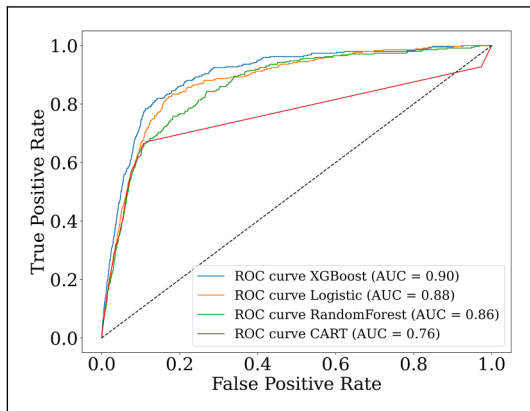
We further validate these findings using a county-level analysis covering 1908 counties from 2005 to 2014 (Table EC.3). The results remain qualitatively the same, with 10 more IPCS-related comments associated with 3.5 more drug poisoning deaths at the county-year level.

**5.2.2 Prediction of Provider Sanctions.** Next, we investigate the predictive power of the ethical indicators along with star ratings and physician metadata in identifying provider sanctions across all sample periods. Employing a range of machine learning algorithms, including XGBoost, Logistic Regression, Random Forest, and Decision Tree, we assess their performance using the ROC AUC. The results in Figure 3 demonstrate strong predictive capabilities across all models, with ROC AUCs ranging from 0.76 to 0.90. This consistent performance highlights the robustness of the ethical indicators in

**Table 4.** Results of IPCS and drug poisoning mortality.

	Dependent Variable: drug poisoning mortality		
	(1) OLS	(2) FGLS	(3) FGLS with lagged DV
IPCS	0.779*** (0.247)	0.721*** (0.252)	2.169*** (0.265)
Per-capita income	−0.002 (0.006)	−0.006** (0.003)	−0.000 (0.003)
Population	−0.000 (0.000)	0.000* (0.000)	−0.000*** (0.000)
Year	30.882*** (4.564)	22.195*** (2.397)	18.276*** (3.444)
Year FE		Yes	Yes
State FE		Yes	Yes
N	357	357	306

Note. Robust standard errors in parentheses. \*\*\* $p < .01$ ; \*\* $p < .05$ ; \* $p < .1$ .

**Figure 3.** ROC curves for predicting physician sanctions.

identifying instances of sanctions. Among the evaluated models, XGBoost exhibits the best performance with an AUC of 0.90. Consequently, we employ XGBoost for the subsequent analysis of predicting sanctions after 2016, that is, for cases beyond our review sample period.

We use the Global SHAP graphs to highlight the importance of individual ethical indicators in the XGBoost model's predictions. Sentences related to improper prescribing of controlled substances, fraud/billing problems, sexual misconduct, and neglect and incompetence are highly influential in predicting physician sanctions. Improper prescribing of controlled substances (IPCS) emerges as one of the most significant predictors of sanctions, both overall and specifically for cases after 2016 (Figure 4).

To illustrate the impact of ethical complaints on the likelihood of an individual physician being sanctioned, we provide examples of local SHAP plots. Figure 5(a) illustrates the case of a physician with a base log odds of approximately  $-4.76$ . The presence of improper prescription practice complaints has the most substantial effect on increasing the sanctioned odds

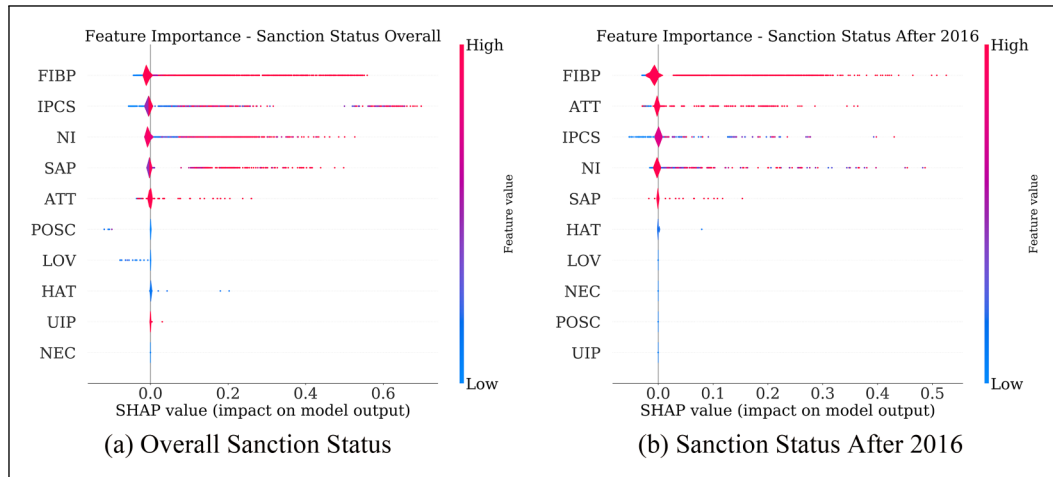
above the baseline, with an increase of 0.8 in the log odds (from approximately  $-4.76$  to  $-3.97$ ).

**5.2.3 Prediction of Future Sanctions.** To further validate the predictive power of our ethical indicators, we conduct an analysis to predict sanction actions after 2016 using review data before 2015. The results show that the model achieves an out-of-sample AUC of 0.90 for sanctions after 2016 (Figure EC.2), indicating that it is effective at predicting future sanctions. The XGBoost model shows the highest accuracy. Figure 4(b) reveals that improper prescribing of controlled substances is again one of the most important factors for predicting future sanctions after 2016. Figure 5(b) illustrates a physician with reviews complaining about sexual misconduct, which significantly increases their sanctioned log-odds. Together, the SHAP plots point to the importance of addressing the issue of improper prescribing of controlled substances (IPCS) and the effectiveness of our model in identifying and capturing such trends. Additionally, reviews related to sexual misconduct (SAP), and fraud or improper billing practices (FIBP), are also associated with a higher likelihood of future sanction.

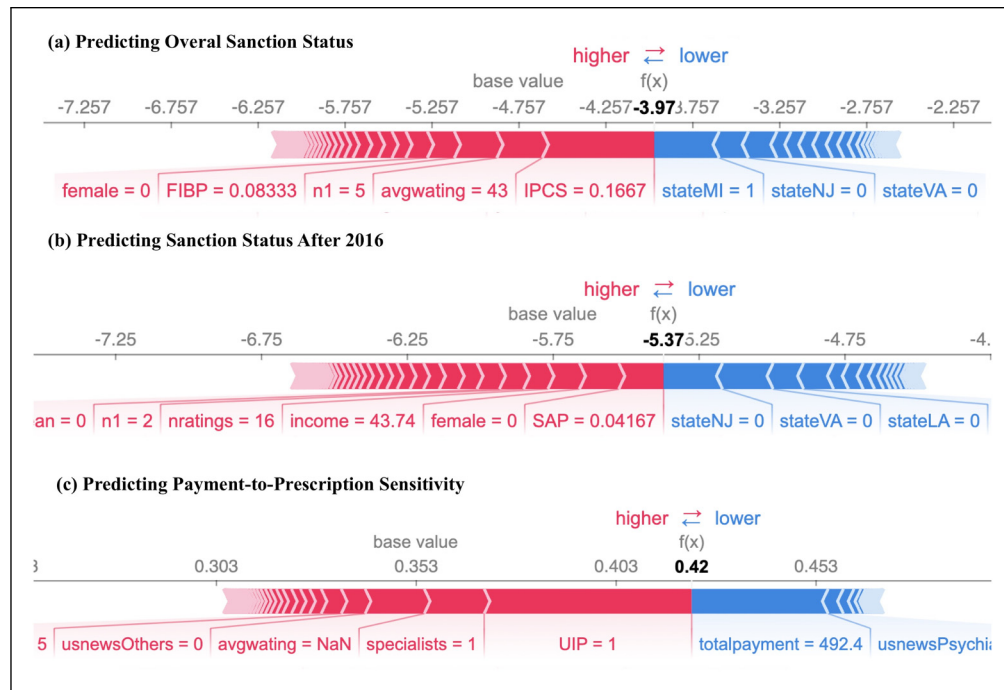
The heightened importance of IPCS in predicting future sanctions aligns with the increased scrutiny during the ongoing opioid crisis. The widespread abuse of prescription opioids has led to stricter regulations and oversight of prescription practices (Barre et al., 2019; DuBois et al. 2016). This increased focus on preventing prescription drug abuse and improving patient safety (Degenhardt et al., 2019; Rutkow et al., 2015) may result in more severe penalties for physicians who engage in inappropriate prescription practices, such as prescribing outside the scope of professional practice, failing to properly monitor patients for signs of addiction, or prescribing excessive amounts of controlled substances.

#### 5.2.4 Prediction of Future Injuries and Professional Liability.

We utilize Florida's Professional Liability Tracking Database to demonstrate how ethical indicators can predict specific consequences of unethical behaviors in healthcare, including psychological, physical, legal, and financial outcomes. We select medical malpractice claims submitted after 2016 to evaluate the predictability of our ethical indicators generated from patient reviews before 2015. Using a multinomial logistic regression model, we examine the relationship between the type of injury (categorized as emotional only, temporary, or permanent) and these ethical indicators. The results in Table 5 reveal significant correlations between ethical indicators and specific injury types. The sexual misconduct indicator (SAP) substantially increases the likelihood of emotional-only injuries, with an odds ratio of 2.35. The unnecessary invasive procedures (UIP) indicator also shows a large effect, increasing the odds of such injuries by over 50%. In contrast, the negative care indicator (NEC) demonstrates a stronger association with physical injuries: it increases the odds of temporary injuries by 63% and permanent injuries by 25%. These



**Figure 4.** SHAP variable importance plots for predicting physician sanctions (global interpretability).



**Figure 5.** SHAP variable importance plots (local interpretability).

findings highlight the predictive specificity of our ethical indicators. For example, while sexual misconduct is more strongly associated with emotional harm, negative care incidents (as indicated by reports of pain and suffering) are more likely to predict future physical injuries.

We further examine the relationship between ethical indicators and indemnity payments using  $\ln(\text{Indemnity})$  (Table 6). We find that most indicators, particularly improper prescription of controlled substances (IPCS) and negative care (NEC), significantly predict indemnity amounts. When all ethical indicators are included in the same model, the negative care

(NEC) and neglect and incompetence (NI) indicators remain significant predictors, with coefficients indicating 78% and 17% increases in the amount paid, respectively.

**5.2.5 Prediction of Rent Seeking.** To examine the relationship between physician rent-seeking behaviors and ethical indicators, we conduct two regression analyses. The first analysis employs an OLS regression model with payment-prescription sensitivity as the dependent variable. The second analysis uses logistic regression, where the dependent variable is an indicator for physicians with a sensitivity above the sample

**Table 5.** Ethical indicators and injury types from Florida Claims Data after 2016.

	Dependent Variable: types of injury					
	Emotional only		Permanent		Temporary	
	OR	SE	OR	SE	OR	SE
IPCS	0.00	0.000	0.97	0.133	1.26	0.181
SAP	2.35**	0.336	0.87	0.157	1.06	0.218
UIP	1.45***	0.136	1.07	0.043	1.12*	0.062
NI	1.27	0.168	1.04	0.044	1.10	0.061
ATT	0.72	0.206	1.00	0.041	0.98	0.065
FIBP	1.25	0.241	1.06	0.073	0.72*	0.169
POSC	3.35	1.00	1.40	0.302	1.05	0.582
NEC	0.60	0.479	1.25***	0.077	1.63***	0.095
LOV	0.00	0.000	0.83	0.253	1.57	0.297
HAT	0.64	0.510	1.03	0.101	1.04	0.159
Gender male	1.62	0.368	2.37***	0.076	1.75***	0.117
Overall rating	0.69***	0.139	0.97	0.031	1.02	0.053
YOE	1.01	0.011	1.00**	0.002	1.00	0.004
lnPopulation	0.71	0.255	0.94	0.051	0.87	0.086
lnIncome	1.00	0.219	0.99	0.044	1.11	0.072
Specialists FE		Yes		Yes		Yes
N		29,447		29,447		29,447

Note. OR = odds ratio; SE = standard error. \* $p < .1$ ; \*\* $p < 0.05$ ; \*\*\* $p < .01$ .

median (0.315).<sup>9</sup> The empirical results in Table 7 present a clear pattern: Indicators for improper prescription of controlled substances (IPCS) and fraudulent billing practices (FIBP) show significant associations with the dependent variables in both models, suggesting reviews about these behaviors correlate with higher payment-prescription sensitivity. Unnecessary invasive procedures (UIP) are significant in the OLS model, implying that physicians who perform more of these procedures may also exhibit rent-seeking behaviors. The significant coefficients of IPCS, FIBP, and UIP affirm the validity of the indicators as financial incentives from pharmaceutical companies are likely to influence these behaviors. In contrast, the effects of other indicators such as sexual abuse of patients (SAP) and attitudinal or communication unprofessionalism (ATT) are not significant.

We next compare several prediction models using the payment-prescription sensitivity as the target variable. Table EC.14 demonstrates comparable performance among the models, with XGBoost showing a slight advantage. The out-of-sample  $R^2$ s of the models are relatively low, with the highest equal to 0.119. This indicates that subtler individual behaviors are more difficult to predict using online review data. The SHAP plot (Figure EC.4) highlights the influence of various features in the XGBoost model. Interestingly, while the more direct indicators like UIP, IPCS, and FIBP are, as expected, strong predictors due to their financial implications, nonfinancial indicators such as ATT (attitudinal or communication unprofessionalism) and HAT (hate) are also important. This finding suggests that, when predicting rent-seeking behaviors, the model benefits from a broader spectrum of indicators.

**5.2.6 Prediction of Clinical Quality.** We employ the 2018 CMS MIPS quality score to evaluate the ethical indicators' association with provider quality. Again, we employ both an OLS and logistic regression, where the latter predicts physicians with quality scores above 90%. Our analyses reveal that most ethical indicators significantly predict future clinical quality, as detailed in Table 8. Notably, indicators for unnecessary invasive procedures (UIP) and fraud and improper billing practices (FIBP) correspond to the largest decreases in quality scores, reducing them by 0.90% and 0.88% respectively out of a total 100%. Additionally, overall rating scores (1–5) are also significant predictors, with each additional rating point correlating to a 1% increase in the MIPS quality score. However, emotivism indicators show no significant association with the MIPS quality score. The lack of correlation between emotive measures from patient reviews and MIPS quality scores likely stems from the subjective nature of these indicators, which often focus more on personal experiences and perceptions than on measurable clinical quality.

Furthermore, we investigate whether the quality score in 2018 could predict future sanction post-2018. The analysis yields a ROC AUC of 0.55 (Figure EC.5), which is only marginally better than a random guess (0.5). This suggests that, while our ethical indicators can predict the physician's quality, relying solely on quality scores is insufficient in forecasting future medical misconduct.

In summary, our ethical indicators demonstrate strong predictive power across a wide range of external validations, as summarized in Table 9. Collectively, these findings attest to the robustness and versatility of our approach.

**Table 6.** Ethical indicators and indemnity payment amount from Florida Claims Data after 2016.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
IPCS	0.21* (0.086)										0.12 (0.088)
SAP		0.09 (0.099)									−0.07 (0.102)
UIP			0.10*** (0.028)								0.03 (0.031)
NI				0.13*** (0.027)							0.07* (0.031)
					(0.026)						(0.029)
FIBP						0.08 (0.050)					−0.01 (0.053)
POSC							−0.06 (0.225)				−0.05 (0.225)
NEC								0.32*** (0.059)			0.25*** (0.062)
LOV									−0.06 (0.149)		−0.07 (0.149)
HAT										0.19** (0.068)	0.08 (0.072)
Gender male	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)	0.37*** (0.038)
Overall rating	−0.02 (0.019)	−0.02 (0.019)	−0.01 (0.019)	0.00 (0.020)	−0.01 (0.020)	−0.02 (0.019)	−0.02 (0.019)	−0.02 (0.019)	−0.02 (0.019)	−0.02 (0.019)	0.01 (0.020)
YOE	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)
lnPopulation	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)	−0.09** (0.033)
lnIncome	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)	−0.01 (0.029)
Specialist FE						Yes					
N						29,447					

Note. Robust standard errors in parentheses. \* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ .

### 5.3 Robustness Tests

We conduct five robustness tests. First, to ensure that unmatched NPI data does not skew our model's performance, we implement a propensity score trimming approach. We train a logistic regression classifier to predict the likelihood of missing NPI and generate propensity scores for each physician. By sorting matched physicians by these scores and removing the bottom 10%, that is, physicians who are least likely to have missing NPI, this approach yields comparable characteristics between “trimmed matched” and unmatched groups. The model trained on the trimmed matched set exhibits negligible performance differences compared to the models trained on the entire matched set. Second, we evaluate the sanction prediction model's performance across different states using state-specific review data. This allows us to assess if the influence of diverse state-specific factors such as regulatory, cultural, societal, and demographic factors impact model training and prediction. We find that the model has relatively satisfactory performance despite geographic variations. Third,

we stratify the sanction prediction model by the number of reviews a physician receives and find consistent performance across different review count categories. The model shows a decline in precision for physicians with only one review. Still, the model's ability to correctly rule out sanctions (TNR) and identify true instances of sanctions (TPR) remains relatively stable. Fourth, we show our sanction model strikes a balance between precision (avoiding false positives) in the very high-risk group, recall (identifying true positive cases) in high and medium-risk groups, and specificity/TNR (correctly identifying true negative cases) in the low-risk group. Fifth and finally, we assess the impact of training data size on model accuracy. We find consistent performance across different subsets, even with as little as 25% of the review data. Precision appears to be most sensitive to the amount of training data used. The details of the tests are reported in Section EC.2.



**Table 7.** Ethical indicators and rent-seeking behaviors.

	Dependent variables	
	Payment-prescription sensitivity (1) OLS	High sensitivity (2) Logistic
IPCS	0.023***	0.083**
SAP	−0.007	−0.054
UIP	0.008**	0.013
NI	−0.001	0.015
ATT	−0.002	−0.003
FIBP	0.013**	0.070***
POSC	0.003	−0.076
NEC	0.013*	0.052
LOV	0.021	0.069
HAT	−0.003	−0.014
Gender male	0.012***	0.049**
Overall rating	−0.004	−0.013
YOE	−0.001***	−0.002**
lnPopulation	0.015***	0.063***
lnIncome	0.001	0.039***
Specialist FE	Yes	Yes
N	64,919	64,919

Note. Robust standard errors in parentheses. \* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ .

## 6 Discussion and Conclusion

### 6.1 Implications for Literature

Our research aims to identify potential medical ethics violations by considering patient reviews as a primary source of information. Based on a comprehensive review of the literature, we develop a set of ethical indicators grounded on three ethics theories. To support our analyses, we merge data from multiple sources, including patient reviews, physician sanction data, financial relationships with pharmaceutical companies, drug poisoning mortality, malpractice claims, as well as national quality measures. We design and validate a machine learning framework capable of predicting state and county-level poisoning mortality, physician sanction status, malpractice injury payments, payment-prescription sensitivity, and clinical quality.

Our results provide compelling evidence for the power of patient-generated content in identifying ethical risks. Our findings add to the ongoing debate in the literature regarding the value of online reviews in healthcare. By demonstrating that textual content from patient reviews contains valuable information beyond star ratings, our work suggests that these reviews can serve as an early warning system for ethical breaches, thereby broadening their utility in operations management.

Moreover, our analysis sheds light on the relative importance of different ethical theories in identifying unethical behavior in healthcare. The SHAP analysis (Figure 4) and regression models (Tables 5 to 8) consistently reveal that deontological indicators are the strongest predictors of various outcomes. The insight contributes to the ongoing discourse in

**Table 8.** Ethical indicators and CMS MIPS quality measure.

	Dependent Variables	
	Quality Score (1) OLS	High Quality (2) Logistic
IPCS	−0.345***	−0.009*
SAP	−0.310***	−0.012**
UIP	−0.904***	−0.053***
NI	−0.320***	−0.025***
ATT	−0.161**	−0.002
FIBP	−0.883***	−0.045***
POSC	−0.106	−0.004
NEC	−0.527***	−0.035***
LOV	0.029	0.001
HAT	−0.104	−0.004
Gender male	−2.010***	−0.163***
Overall rating	1.031***	0.071***
YOE	−1.954***	−0.076***
lnPopulation	−0.589***	−0.014***
lnIncome	−0.122*	0.012**
Specialist FE	Yes	Yes
N	170,173	170,173

Note. Robust standard errors in parentheses. \* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ .

the ethics literature about the practical applicability of different ethical frameworks in professional settings (Beauchamp, 2003; Mandal et al., 2016), offering support for the high relevance of duty-based ethical considerations. In contrast, we show that emotivism has the least predictive power. This finding diverges from previous social media research, which often suggests emotions as a key factor in shaping behavior. Our results also underscore the multifaceted nature of ethical risks. The model's ability to predict a range of outcomes highlights the interconnectedness of different types of ethical violations. Relatedly, the results imply that ethical lapses in one area may be indicative of broader behavioral patterns, a perspective that has been underexplored in the literature.

Furthermore, our findings have implications for the literature on professional compliance and regulation. The ability to predict ethical violations using patient-generated content suggests a potential shift in the dynamics of professional oversight. It indicates that patients, through their collective feedback, can play a more significant role in identifying misconduct. Their voice can complement traditional regulatory mechanisms. This opens up new avenues for research on the role of patient feedback in professional governance.

### 6.2 Managerial and Policy Implications

From a risk management perspective, preventing medical malpractices and unethical physician behavior is less costly, both economically and socially, than addressing their aftermath. While our model relies on reviews written after incidents occur, it enables faster detection of patterns compared to many



**Table 9.** Summary of external validations and prediction performances.

External Validation	Description	Key Results
State-level drug poisoning mortality	Employing state-level IPCS to predict drug poisoning deaths	IPCS coefficients: 0.779 (OLS), 0.721 (FGLS), 2.169 (lagged IPCS)
County-level drug poisoning mortality	Utilizing IPCS to predict county-level drug poisoning deaths	Similar results to state-level with economically significant effects
Overall provider sanctions	Employing ethical indicators to predict sanctions using machine learning models	ROC AUC ranging from 76% to 90% (XGBoost: 90%); key indicators: IPCS, FIBP, SAP, NI
Future sanctions	Using pre-2015 review data to predict post-2016 sanctions	AUC: 90%; key indicators: IPCS, FIBP, SAP
Future injuries and professional liability	Employing ethical indicators to predict injury types and financial liability in post-2016 Florida Professional Claims	SAP: odds ratio 2.35 for emotional injuries; NEC: 63% increase for temporary injuries; predicts indemnity payments
Rent seeking	Investigating the relationship between ethical indicators and rent-seeking behavior	Significant predictors: IPCS, FIBP, UIP; modest predictive performance
Clinical quality	Utilizing 2018 CMS MIPS quality score to evaluate the association with ethical indicators	UIP and FIBP: largest decreases in quality scores; emotive indicators: no significant association; MIPS scores: minimal predictive power for future sanctions

Note. IPCS = improper prescribing of controlled substances; FIBP = fraudulent and improper billing practices; SAP = sexual abuse of patients; NI = neglect and incompetence; UIP = unnecessary invasive procedures; NEC = negative care.

traditional regulatory mechanisms, which often lag significantly behind violations. Unethical behaviors rarely occur in isolation—early identification through patient reviews can help detect concerning patterns before they escalate into more serious violations or harm additional patients. To this end, our model provides a foundation for early warning and prevention systems for such adverse incidents. The economic implications could be significant. Take malpractice for example, over the period of 2010 to 2019, \$42 billion was paid to victims of medical malpractice in the USA (Justpoint, 2021), with the average settlement amount ranging from \$425,000 to \$1 million (Medscape, 2013). Another example is the opioid crisis which is in part due to unethical prescription practices. The economic cost of opioid use disorder was estimated to be \$471 billion in the USA (Luo, 2021). Early-prevention systems can identify practitioners at risk of malpractice, allowing for more proactive strategies and avoiding costly legal and financial consequences. When integrated with complaint records and internal reports, the system becomes invaluable in risk management and resource allocation. For example, it enables hospitals to direct resources towards areas with higher risks of ethical violations or invest in training programs aimed at preventing such issues.

Likewise, our approach leads to management strategies that can reduce the impact of unethical practices. The strategies can work through three avenues. First, early detection of ethical lapses can not only mitigate their negative impacts but also enhances patient trust and satisfaction, ultimately leading to cost reductions. Policymakers can incentivize the adoption of such models through grants or recognition programs

for institutions that actively contribute to model development or maintenance of review platforms. Second, drawing from the economic theory of crime (Becker, 1968), it is important to communicate to providers the concrete benefits of ethical compliance and the repercussions of violations. More targeted educational campaigns or mandatory training programs can be informed by a predictive model's findings. Lastly, our approach serves as a promising tool for empowering victims of unethical practices, which often remain underreported due to power dynamics or fear of reprisal (Roland et al., 2011). Policymakers can promote the use of online review platforms as legitimate channels for patient feedback while ensuring that these platforms are safeguarded against retaliation. Given the importance of deontological indicators, these platforms could prompt patients to share specific instances of adherence or violations of code of ethics. Integrating this data into health-care oversight mechanisms increases visibility and scrutiny of violations. This could catalyze a shift towards improved compliance and stronger patient advocacy.

### 6.3 Generalizability to Other Sectors

Professional services, such as law, education, management consulting, and banking, are vital to modern societies and economies. Like healthcare, these fields involve high levels of customer contact and delivery specificity, where each case or problem is unique; they are also characterized by fluid operational processes, where professionals exercise judgment in determining outcomes and means (Harvey et al., 2016). Given

the similarities in the nature of these services and the importance of ethical conduct across all professional domains, it is natural to consider the broader applicability of our approach: can our approach be extended to these other professional services sectors?

On the one hand, the lenses of different ethics theories can still be relevant in other sectors. For example, consider the case of the FTX scandal in finance (Oliver, 2023): it is reported that Sam Bankman-Fried claimed a utilitarian viewpoint: “*the only moral rule that mattered was doing whatever would maximise utility*”—which clearly clashed with deontological principles and led to grave consequences. Second, the design pattern of our approach can serve as a strong foundation, being grounded in theory, mining specific indicators from large corpora, and using them as features for downstream models. On the other hand, understanding the contextual details of different sectors is crucial (Joglekar et al., 2016). Other sectors take feedback through different channels, such as client surveys or complaint registries (e.g., the CFPB). Integrating feedback from diverse sources requires additional attention. Moreover, the ethical perspectives pertinent to different professions may vary. For instance, in law, ethical indicators might emphasize fairness in representation and confidentiality, while in banking, the focus could shift to transparency and fiduciary responsibilities. Finally, careful validation should be conducted in other sectors. For instance, in legal services, validation might involve disciplinary actions from bar associations, while in financial services, it could include reports from regulatory authorities such as the SEC. In sum, extending our approach to other sectors holds promise, but will likely require adaptations to their ethical requirements, feedback mechanisms, and relevant validations.

#### 6.4 Limitations and Future Research

Our study has several limitations. First, the model developed in this research is based on reviews from a specific time period in the USA. As social media discussions evolve, the model may need updating to maintain relevance. Furthermore, our sample may not be representative of the broader patient population due to the digital divide, potentially underrepresenting less technologically savvy groups (Hao, 2015). Additionally, while our method focuses on discovering more negative aspects from reviews, which partly mitigates the issue of fake reviews, we cannot validate the authenticity of every review. Despite this, our approach is also applicable to reviews from authenticated sources like insurers and providers, and future research may explore additional validation techniques through automated algorithms or cross-referencing with other data sources. Second, although the active search strategy can handle skewed class distributions, it may miss rarer yet important ethical violations. In this vein, our approach could benefit from more granular, domain-specific theories, especially on how utilitarianism and emotivism principles are applicable in

guiding patient–provider interactions. Additionally, information retrieval methods enhanced with domain knowledge may be considered (Tamine and Goeuriot, 2021). Third and most important, it is imperative to acknowledge the limitations and risks of our ethical indicators and predictive models. Drawing from Harcourt’s (2007) critique of actuarial methods in criminal law, several cautions apply to our work. For one, merely predicting unethical behavior may not reduce such behavior—predictions require appropriate regulations and enforcement mechanisms to drive meaningful change. For another, the use of machine learning models in the field of ethics has raised complex questions about the role of technical knowledge in shaping justice, an issue that has garnered increasing attention in the broader scientific community (Christian, 2020) but out of the scope of this paper.

Despite these limitations, our work represents a meaningful proof-of-concept in discovering ethical violations and mitigating their societal costs. Future studies that aim to confirm and build on our findings may incorporate other forms of text data, such as interviews with providers, patient surveys, or expand to ethical violations of other forms. When applied responsibly, research in this area has the power to significantly improve patient outcomes, prevent harm, and increase public trust in the healthcare system.

#### Acknowledgments

The authors are grateful to the Special Issue editors, the senior editor, and anonymous reviewers for their constructive comments. They also extend thanks to the participants of the 2020 Conference on Health IT & Analytics (CHITA 2020) for the valuable feedback received.


#### Declaration of Conflicting Interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

#### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

#### ORCID iDs

Kanix Wang  <https://orcid.org/0000-0003-1355-577X>

Feng Mai  <https://orcid.org/0000-0001-6897-8935>

#### Supplemental Material

Supplemental material for this article is available online (doi: 10.1177/10591478251318885).

#### Notes

1. Some of the studies referenced in Table 1 Panel B were not originally framed explicitly as risk quantification, we have included them in our comparative analysis due to the potential adaptability of their methodologies or data to similar contexts.
2. The medical ethics literature features long-standing debates about the relationships between different ethical theories. Deontological and utilitarian approaches to doctor–patient interactions, for

instance, can lead to divergent conclusions (Garbutt and Davies, 2011; Mandal et al., 2016). The main goal of our work is not to engage in the traditional moral and philosophical analysis of these perspectives. Rather, we seek to utilize data science methods to (a) provide a more positive (descriptive) view of patient perspectives on healthcare ethics; and (b) assess the predictive value of measures constructed from these theoretical lenses.

3. While emotivism shares some similarities with sentiment analysis, they differ in their theoretical foundations. Emotivism posits that emotions are the basis for moral judgments, particularly in the context of interpersonal relationships such as patient–physician interactions. Sentiment analysis is a computational technique that assesses the overall valence of language in various domains, including social media. Sentiment scores may not fully capture the emotional and moral aspects of patient–physician relationships that are central to healthcare ethics.
4. The states include Alabama, Alaska, Arkansas, Colorado, Delaware, District of Columbia, Georgia, Florida, Louisiana, Minnesota, Missouri, Nevada, New Hampshire, New Jersey, North Carolina, South Dakota, Vermont, Virginia, West Virginia, and Wisconsin.
5. The database is available at <https://apps.fldfs.com/PLCR/Search/MPLClaim.aspx>. It is worth noting two caveats. First, the dataset excludes some claims against Health Maintenance Organizations (HMOs), which operate under different liability structures. Second, the presence of a claim or a settlement does not inherently prove medical malpractice. Rather, these records indicate instances where patients or their representatives alleged harm and sought compensation. The injury classifications and payment amounts reflect the perceived severity of the alleged misconduct and its consequences.
6. Available at <https://huggingface.co/physician-ethics-responsible-data-science/eth-setfit-multilabel-model>.
7. Normalized Discounted Cumulative Gain (NDCG) measures search result quality by weighting the relevance and ranking of items. It computes cumulative gain from relevance scores and applies penalties for lower-ranked items using a logarithmic scale. The formula at a given rank  $p$  is with  $rel_i$  representing the relevance at position  $i$ . NDCG normalizes the Discounted Cumulative Gain (DCG) by the Ideal DCG (the DCG for a perfectly ordered result list). Its values range from 0 to 1, where 1 denotes a ranking consistent with humans.
8. To provide a more balanced perspective on utilitarianism and emotivism ethical indicators, which one may suspect naturally contain fewer specific topics compared to deontology, we conducted post hoc topic modeling analyses. The results showed considerable granularity within these indicators beyond simple sentiment. For the positive care (POSC) comments, we identified topics on the positive impact of the care, such as diagnosis, surgical care, emotional support, and the duration of care provided. Similarly, in the emotivism-themed comments, we observed a diverse range of expressions, such as appreciation for the doctors and accolades for other aspects of the care delivery. Detailed breakdowns of topics are provided in Table EC.15, and word clouds highlighting the themes within each topic are presented in Figures EC.14–EC.17.
9. As Table EC.6 illustrates, the majority of states show an average value significantly greater than zero, indicating that at the population level, there is a significant correlation between physician prescription behavior and the compensation received from pharmaceutical companies.

## References

- Abbasi A, Li J, Adjeroh D, et al. (2019) Don't mention it? Analyzing user-generated content signals for early adverse event warnings. *Information Systems Research* 30(3): 1007–1028.
- Abrahams AS, Fan W, Wang GA, et al. (2015) An integrated text analytic framework for product defect discovery. *Production and Operations Management* 24(6): 975–990.
- Araz OM, Choi T, Olson DL, et al. (2020) Role of analytics for operational risk management in the era of big data. *Decision Sciences* 51(6): 1320–1346.
- Bandman B (2003) *The Moral Development of Health Care Professionals: Rational Decision Making in Health Care Ethics*. Westport, CT: Praeger.
- Barre L, Gallo J and McDonald JV (2019) Review of disciplinary actions regarding controlled substances, Rhode Island 2012–2017. *Journal of Medical Regulation* 105(1): 22–27.
- Bastani H, Goh J and Bayati M (2019) Evidence of upcoding in pay-for-performance programs. *Management Science* 65(3): 1042–1060.
- Bauder R and Khoshgoftaar T (2018) Medicare fraud detection using random forest with class imbalanced big data. In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 06–09 July 2018, Salt Lake City, UT, pp. 80–87.
- Beauchamp TL (2003) Methods and principles in biomedical ethics. *Journal of Medical Ethics* 29(5): 269–274.
- Becker GS (1968) Crime and punishment: An economic approach. *Journal of Political Economy* 76(2): 169–217.
- Blenkinsopp J, Snowden N, Mannion R, et al. (2019) Whistleblowing over patient safety and care quality: A review of the literature. *Journal of Health Organization and Management* 33(6): 737–756.
- Bobroske K, Freeman M, Huan L, et al. (2022) Curbing the opioid epidemic at its root: The effect of provider discordance after opioid initiation. *Management Science* 68(3): 2003–2015.
- Brennan TA, Rothman DJ, Blank L, et al. (2006) Health industry practices that create conflicts of interest: A policy proposal for academic medical centers. *Journal of the American Medical Directors Association* 295(4): 429–433.
- Brown T, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877–1901.
- Busch RS (2012) *Healthcare Fraud: Auditing and Detection Guide*. Hoboken, NJ: John Wiley & Sons.
- Chadwick R (2016) Professional ethics. In: Craig E (ed) *Routledge Encyclopedia of Philosophy* (1st ed.). London: Routledge, pp. 314–315.
- Chalfin A, Danieli O, Hillis A, et al. (2016) Productivity and selection of human capital with machine learning. *American Economic Review* 106(5): 124–127.
- Chandrasekaran A, Senot C and Boyer KK (2012) Process management impact on clinical and experiential quality: Managing tensions between safe and patient-centered healthcare. *Manufacturing & Service Operations Management* 14(4): 548–566.
- Chen T and Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: Association for Computing Machinery, pp. 785–794.
- Christian B (2020) *The Alignment Problem: Machine Learning and Human Values*. New York, NY: WW Norton & Company.

- Cohen MA and Kunreuther H (2007) Operations Risk Management: Overview of Paul Kleindorfer's contributions. *Production and Operations Management* 16(5): 525–541.
- Cohen MC (2018) Big data and service operations. *Production and Operations Management* 27(9): 1709–1723.
- Coleman C, Chou E, Katz-Samuels J, et al. (2022) Similarity search for efficient active learning and search of rare concepts. *AAAI* 36(6): 6402–6410.
- Cui R, Gallino S, Moreno A, et al. (2018) The operational value of social media information. *Production and Operations Management* 27(10): 1749–1769.
- De Bock KW, Coussement K, Caigny AD, et al. (2023) Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research* 317(2): 249–272.
- Degenhardt L, Grebely J, Stone J, et al. (2019) Global patterns of opioid use and dependence: Harms to populations, interventions, and future action. *Lancet* 394(10208): 1560–1579.
- Devlin J, Chang M-W, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics, pp. 4171–4186.
- DuBois JM, Anderson EE, Chibnall JT, et al. (2019) Serious ethical violations in medicine: A statistical and ethical analysis of 280 cases in the United States from 2008–2016. *American Journal of Bioethics* 19(1): 16–34.
- DuBois JM, Chibnall JT, Anderson EE, et al. (2016) A mixed-method analysis of reports on 100 cases of improper prescribing of controlled substances. *Journal of Drug Issues* 46(4): 457–472.
- Ekin T, Frigau L and Conversano C (2021) Health care fraud classifiers in practice. *Applied Stochastic Models in Business and Industry* 37(6): 1182–1199.
- Fineschi V, Turillazzi E and Cateni C (1997) The new Italian code of medical ethics. *Journal of Medical Ethics* 23(4): 239–244.
- Gal U, Hansen S and Lee A (2022) Research perspectives: Toward theoretical rigor in ethical analysis: The case of algorithmic decision-making systems. *Journal of the Association for Information Systems* 23(6): 1634–1661.
- Gallup (2023) Nurses retain top ethics rating in U.S., but below 2020 high. *Gallup.com*. Available at <https://news.gallup.com/poll/467804/nurses-retain-top-ethics-rating-below-2020-high.aspx> (accessed 22 November 2023).
- Gao G, Greenwood BN, Agarwal R, et al. (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality. *MIS Quarterly* 39(3): 565–590.
- Garbutt G and Davies P (2011) Should the practice of medicine be a deontological or utilitarian enterprise? *Journal of Medical Ethics* 37(5): 267–270.
- Giles T, Hungerman D and Oostrom T (2023) *Opiates of the Masses? Deaths of Despair and the Decline of American Religion* (No. w30840). Cambridge, MA: National Bureau of Economic Research.
- Gour A, Aggarwal S and Kumar S (2022) Lending ears to unheard voices: An empirical analysis of user-generated content on social media. *Production and Operations Management* 31(6): 2457–2476.
- Ham C and Alberti KGMM (2002) The medical profession, the public, and the government. *The British Medical Journal* 324(7341): 838–842.
- Hao H (2015) The development of online doctor reviews in China: An analysis of the largest online doctor review website in China. *Journal of Medical Internet Research* 17(6): e134.
- Harcourt BE (2007) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.
- Harvey J, Heineke J and Lewis M (2016) Editorial for journal of operations management special issue on “professional service operations management (PSOM)”. *Journal of Operations Management* 42–43(1): 4–8.
- Herland M, Bauder RA and Khoshgoftaar TM (2019) The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *Journal of Big Data* 6(1): 1263. DOI: 10.1186/s40537-019-0181-8.
- Herland M, Khoshgoftaar TM and Bauder RA (2018) Big data fraud detection using multiple medicare data sources. *Journal of Big Data* 5(1): 1–21.
- Husted GL and Husted JH (2005) Ethics: A patient-centered approach. In: Daly J (ed) *Professional Nursing: Concepts, Issues, and Challenges*. New York, NY: Springer Publishing Company, 175–179.
- Janssen A and Zhang X (2023) Retail pharmacies and drug diversion during the opioid epidemic. *American Economic Review* 113(1): 1–33.
- Jiang S, Garnett R and Moseley B (2019) Cost effective active search. *Advances in Neural Information Processing Systems* 32: 3880–3889.
- Jimenez MD and Foster D (1998) The importance of compliance programs for the health care industry. *University of Miami Business Law Review* 7: 503.
- Joglekar NR, Davies J and Anderson EG (2016) The role of industry studies and public policies in production and operations management. *Production and Operations Management* 25(12): 1977–2001.
- Justpoint (2021) US Medical Malpractice Case Statistics | Justpoint. Available at <https://justpoint.com/knowledge-base/us-medical-malpractice-case-statistics> (accessed 28 December 2022).
- Kaptein M (2008) Developing a measure of unethical behavior in the workplace: A stakeholder perspective. *Journal of Management* 34(5): 978–1008.
- Kc D, Kim TT and Liu J (2022) Electronic prescription monitoring and the opioid epidemic. *Production and Operations Management* 31(11): 4057–4074.
- Kc DS, Scholtes S and Terwiesch C (2020) Empirical research in healthcare operations: Past research, present understanding, and future opportunities. *Manufacturing & Service Operations Management* 22(1): 73–83.
- Ko D, Mai F, Shan Z, et al. (2019) Operational efficiency and patient-centered health care: A view from online physician reviews. *Journal of Operations Management* 65(4): 353–379.
- Kornfield M, Higham S and Rich S (2022) Inside the sales machine of the ‘kingpin’ of opioid makers. *Washington Post*. Available at <https://www.washingtonpost.com/investigations/interactive/2022/mallinckrodt-documents-doctors-sales/> (accessed 20 October 2022).
- Krauss C, Do XA and Huck N (2017) Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on

- the S&P 500. *European Journal of Operational Research* 259(2): 689–702.
- Kumar A, Raghunathan A, Jones RM, et al. (2021) Fine-tuning can distort pretrained features and underperform out-of-distribution. *International Conference on Learning Representations*.
- Kumaraswamy N, Markey MK, Ekin T, et al. (2022) Healthcare fraud data mining methods: A look back and look ahead. *Perspectives in Health Information Management* 19(1): 1i.
- Lantzy S. and Anderson D.. 2020. Can consumers use online reviews to avoid unsuitable doctors? Evidence from RateMDs.com and the Federation of State Medical Boards. *Decision Science* 51(4): 962–984.
- Li R, Tobey M, Mayorga ME, et al. (2023) Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management* 25(3): 1051–1065.
- Li X, Guo P and Lian Z (2016) Quality-speed competition in customer-intensive services with boundedly rational customers. *Production Operation Management* 25(11): 1885–1901.
- Liu R, Huang J and Zhang Z (2023) Tracking disclosure change trajectories for financial fraud detection. *Production and Operation Management* 32(2): 584–602.
- Lu SF and Rui H (2018) Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. *Management Science* 64(6): 2557–2573.
- Lucey C and Souba W (2010) Perspective: The problem with the problem of professionalism. *Academy Medical* 85(6): 1018–1024.
- Lundberg SM and Lee S-I (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30: 4768–4777.
- Luo F. 2021. State-level economic costs of opioid use disorder and fatal opioid overdose — United States, 2017. *Morbidity and Mortality Weekly Report* 70.
- Macy B (2018) *Dopesick: Dealers, Doctors and the Drug Company that Addicted America*. New York, NY: Bloomsbury Publishing.
- Manary MP, Boulding W, Staelin R, et al. (2013) The patient experience and health outcomes. *New England Journal of Medicine* 368(3): 201–203.
- Mandal J, Ponnambath DK and Parija SC (2016) Utilitarian and deontological ethics in medicine. *Tropical Parasitology* 6(1): 5–7.
- Markou P and Corsten D (2021) Financial and operational risk management: Inventory effects in the gold mining industry. *Production and Operation Management* 30(12): 4635–4655.
- Medscape (2013) Malpractice: When to settle a suit and when to fight. *Medscape*. Available at <https://www.medscape.com/viewarticle/811323> (accessed 28 December 2022).
- Mejia J, Mankad S and Gopal A (2021) Service quality using text mining: Measurement and consequences. *Manufacturing & Service Operations Management* 23(6): 1354–1372.
- Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26: 3111–3119.
- Mohri M, Rostamizadeh A and Talwalkar A (2018) *Foundations of Machine Learning*, 2nd edn. Cambridge, MA: The MIT Press.
- Nair A, Nicolae M and Narasimhan R (2013) Examining the impact of clinical quality and clinical flexibility on cardiology unit performance—Does experiential quality act as a specialized complementary asset? *Journal of Operation Management* 31(7–8): 505–522.
- NEJM Catalyst (2018) What Is Risk Management in Healthcare? *Catalyst Carryover* 4(2).
- Oliver J (2023, October 13) ‘I felt trapped’: Caroline Ellison steps out of Sam Bankman-Fried’s shadow in star witness turn. *Financial Times*.
- Papadakis MA, Teherani A, Banach MA, et al. (2005) Disciplinary action by medical boards and prior behavior in medical school. *New England Journal of Medicine* 353(25): 2673–2682.
- Parsons CA, Sulaeman J and Titman S (2018) The geography of financial misconduct: The geography of financial misconduct. *Journal of Finance* 73(5): 2087–2137.
- Peng DX, Ye Y, Feng B, et al. (2020) Impacts of hospital complexity on experiential quality: Mitigating roles of information technology. *Decision Science* 51(3): 500–541.
- Rehurek R and Sojka P (2011) Gensim—Python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3(2): 2.
- Reimers N and Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 3982–3992.
- Robertson SM, Stanley MA, Cully JA, et al. (2012) Positive emotional health and diabetes care: Concepts, measurement, and clinical implications. *Psychosomatics* 53(1): 1–12.
- Roland M, Rao SR, Sibbald B, et al. (2011) Professional values and reported behaviours of doctors in the USA and UK: Quantitative survey. *BMJ Quality & Safety* 20(6): 515–521.
- Rosenstein AH and O’Daniel M (2008) A survey of the impact of disruptive behaviors and communication defects on patient safety. *Joint Commission Journal on Quality and Patient Safety* 34(8): 464–471.
- Rutkow L, Chang H-Y, Daubresse M, et al. (2015) Effect of Florida’s prescription drug monitoring program and pill mill laws on opioid prescribing and use. *JAMA Internal Medicine* 175(10): 1642–1649.
- Saar-Tschanskysky M and Provost F (2007) Decision-centric active learning of binary-outcome models. *Information Systems Research* 18(1): 4–22.
- Saifee DH, (Eric) Zheng Z, Bardhan IR, et al. (2020) Are online reviews of physicians reliable indicators of clinical outcomes? A focus on chronic disease management. *Information Systems Research* 31(4): 1282–1300.
- Song K, Tan X, Qin T, et al. (2020, November 2) MPNet: Masked and permuted pre-training for language understanding. arXiv preprint arXiv:2004.09297.
- Tamine L and Goeuriot L (2021) Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Computing Surveys* 54(7): 146:1–146:38.
- Tunstall L, Reimers N, Jo UES, et al. 2022. Efficient few-shot learning without prompts. In arXiv preprint arXiv:2209.11055.
- Wang X, Wu Q, Lai G, et al. (2019) Offering discretionary healthcare services with medical consumption. *Production and Operations Management* 28(9): 2291–2304.
- Whitaker BF and Laura Beil H (2023, September 12). How Columbia ignored women, undermined prosecutors and protected a predator for more than 20 years. *ProPublica*.

- Wu D (Andrew) (2023). Text-based measure of supply chain risk exposure. *Management Science* 70(7): 4781–4801.
- Xie J, Zhang Z, Liu X, et al. (2021) Unveiling the hidden truth of drug addiction: A social media approach using similarity network-based deep learning. *Journal of Management Information System* 38(1): 166–195.
- Yan L and Pedraza-Martinez AJ (2019) Social media for disaster management: Operational value of the social conversation. *Production and Operations Management* 28(10): 2514–2532.
- Yang CC, Yang H and Jiang L (2014) Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Transaction Management Information System* 5(1): 1–21.
- Zhang X, Du Q and Zhang Z (2022) A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management* 31(8): 3160–3179.
- Zhao W, Liu QB, Guo X, et al. (2022) Quid pro quo in online medical consultation? Investigating the effects of small monetary gifts from patients. *Production and Operations Management* 31(4): 1698–1718.

**How to cite this article**

Wang K, Mai F, Shan Z, Zhang D and Peng X (2025) Analyzing Professional Ethics of Physicians Using Online Patient Reviews: A Machine Learning Approach. *Production and Operations Management* xx(x): 1–22.