

Protocol

Extraction of Pain Severity and Functional Interference From Clinical Narratives Using Domain-Informed Large Language Models: Protocol for a Development and Validation Study

Xiaoyi Zhang¹, MS; Christopher R Wilson², MS; Hannah Eyre², MS; David E Reed II^{2,3}, PhD; Travis Y Hee Wai², PhD; Alexander Kloehn², MPH; Ethan W Rosser², PhD; Gang Luo¹, PhD; Steven B Zeliadt^{2,4}, MPH, PhD

¹Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

²Seattle-Denver Center of Innovation for Veteran-Centered and Value-Driven Care, VA Puget Sound Health Care System, Seattle, WA, United States

³Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States

⁴Department of Health Systems and Population Health, School of Public Health, University of Washington, Seattle, WA, United States

Corresponding Author:

Steven B Zeliadt, MPH, PhD

Seattle-Denver Center of Innovation for Veteran-Centered and Value-Driven Care

VA Puget Sound Health Care System

1660 S. Columbian Way

Seattle, WA 98108

United States

Phone: 1 2062774175

Email: steven.zeliadt@va.gov

Abstract

Background: Chronic pain is a leading cause of disability and requires multidimensional assessment of pain intensity and functioning, yet electronic health records rarely capture these measures systematically. By contrast, surveys collecting patient-reported outcomes can assess pain over multiple dimensions but remain resource-intensive and difficult to scale for continuous population-level monitoring.

Objective: The objective of this study is to develop and validate a domain-informed natural language processing framework to derive pain severity and functional interference outcomes from unstructured clinical narratives. We aim to demonstrate that natural language processing-derived outcomes can serve as a reliable, scalable surrogate for resource-intensive patient-reported surveys.

Methods: This study uses a retrospective cohort of 3725 Veterans with chronic musculoskeletal pain initiating complementary and integrative health therapies across 18 Veterans Health Administration Whole Health Flagship sites (2021-2023). The dataset encompasses longitudinal patient-reported outcome surveys serving as the benchmark, linked with unstructured clinical narratives from the Veterans Health Administration electronic health record. Guided by established psychometric instruments and subject matter expert (SME) input, we developed a seed lexicon and annotation guidelines to identify language distinguishing 3 pain domains: pain severity, interference with enjoyment of life, and interference with general activities. Preliminary large language model (LLM) prompting was used to identify 600 candidate encounters (200 per domain) from 6747 notes across 260 patients for SME annotation, forming a ground truth validation sample. Two candidate LLMs will be evaluated on this sample; the best-performing LLM will generate a large library of span-level annotations to train a scalable, lightweight language model. The study uses a 3-stage validation process: (1) documentation completeness of pain interference in clinical narratives against SME-annotated references; (2) inference accuracy of the LLM-as-annotator and the fine-tuned lightweight model against SME annotations across note-level classification and span-level localization; and (3) concordance between the lightweight model's output and patient-reported pain, enjoyment, and general activity scores across a range of temporal windows.

Results: As of July 2026, the cohort of 3725 Veterans has been identified and linked to clinical notes. The seed lexicon and annotation guidelines have been developed. Applying a developmental LLM to screen 6747 text notes in 6642 unique encounters over a 7-month period for 260 patients, at least 1 of the 3 pain domains was identified in 75% of notes and 99%

of patients. SME validation at the encounter level is in progress. Final results from the subsequent knowledge distillation and validation stages are expected in the first half of 2027.

Conclusions: This protocol outlines a framework for identifying severe pain intensity and interference from clinical narratives, addressing a critical gap in health care system surveillance. To our knowledge, this is the first study to validate clinical text-based pain outcome extraction against patient-reported outcomes in a nationwide longitudinal cohort. If successful, this approach will enable health care systems to continuously monitor reports of pain-related functional interference and support more holistic, patient-centered pain management at scale.

International Registered Report Identifier (IRRID): DERR1-10.2196/96346

JMIR Res Protoc 2026;15:e96346; doi: [10.2196/96346](https://doi.org/10.2196/96346)

Keywords: chronic pain; natural language processing; electronic health records; large language models; patient-reported outcomes; knowledge distillation; Veterans; pain assessment

Introduction

Background

Chronic pain is a pervasive health issue that leads to poor patient outcomes and high health care costs. It affects one-fourth of adults in the United States, ranks as a leading cause of disability, and costs the nation over US \$700 billion annually [1-3]. The prevalence and severity of chronic pain among Veterans of the US military are disproportionately high compared with the general population [4]. As such, chronic pain management is a high priority for health care providers and systems, including the Veterans Health Administration (VA) [5].

As pain is a subjective, multidimensional experience, its assessment is complex but central to evaluating treatment efficacy and effectiveness. The IMMPACT (Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials) consensus recommends assessment of multiple core outcome domains for chronic pain treatment trials, including pain (with its intensity, quality, and temporal aspects), physical functioning, and emotional functioning [6, 7]. Assessing these domains, whether capturing the subjective experience of pain severity or outcomes concerning pain's interference with functioning (eg, general activity) and well-being (eg, enjoyment of life), relies heavily on patients' self-reports.

While electronic health records (EHRs) are increasingly integrating self-reported pain data, with VA leading the shift from basic "fifth vital sign" scores [8] to comprehensive measures [9,10], systematic implementation of multidimensional patient-reported pain measures is rare [11]. The commonly used Numerical Rating Scale (NRS) for pain intensity, a single-item numeric measurement, was intended for clinical screening and has limited utility for longitudinally tracking outcomes or capturing pain's broad interference in patient functioning and well-being [6, 9,12-14]. Although multidimensional instruments like the Brief Pain Inventory, PEG (Pain, Enjoyment, and General Activity) scale, and Patient-Reported Outcomes Measurement Information System (PROMIS) measures offer richer insights, they are typically administered in specific research settings or specialty clinics rather than in routine care [13]. As pain is often discussed with providers as part of routine

care, descriptions of pain interference, treatment plans, and associated progress may offer an alternative way to monitor the multidimensional aspects of pain management, supplementing targeted efforts to capture patient-reported pain outcomes.

Prior Work

As part of a prior quality improvement initiative, VA conducted a nationwide survey across 18 VA medical centers participating in the Whole Health Flagship initiative, collecting patient-reported outcomes (PROs) from Veterans initiating complementary and integrative health (CIH) therapies [15,16]. The survey used measures from the PEG scale, a validated and widely used instrument to assess pain severity and its interference with functioning and quality of life [13]. Over the 2-year period spanning from March 2021 to March 2023, patient surveys were used to track longitudinal PROs from 3725 patients. Although providing valuable patient-reported data, surveys are resource-intensive, suffer from low response rates, and are difficult to scale for continuous population-level monitoring.

An alternative approach to obtaining outcomes equivalent to PROs at scale is through clinical notes, which contain narratives of patient reports and patient-provider discussions. Recent advances in machine learning and natural language processing (NLP) have generated techniques to extract pain and self-reported outcomes from narrative text data [17,18]. Prior works have leveraged NLP methods to achieve objectives such as constructing an ontology of pain over diverse attributes [19], identifying attributes of pain interference [20], and using fine-tuned large language models (LLMs) to parse musculoskeletal pain features from unstructured clinical notes [21]. Notably, recent studies using advanced Transformer-based deep learning models [22] found that model-extracted features of pain experience were in agreement with domain experts [20,23].

Despite the growing body of literature demonstrating the feasibility of using NLP to extract outcomes similar to PROs of chronic pain, several questions remain. First, documentation patterns for pain vary significantly across clinics and patient sociodemographic characteristics [24]. How such variation manifests within the heterogeneous VA population and whether it introduces systematic bias into algorithmic models remains largely unknown. Second, while the broad

notion of pain is frequently documented, documentation of concepts related to subjective aspects such as patient-reported functioning is reportedly more sparse compared with aspects like severity or site [25,26]. A quantitative understanding of how completely clinical text data captures pain severity and functional interference domains remains underexamined. With these challenges in mind, we aim to develop a domain-informed NLP framework to operationalize these pain outcomes, leveraging linked longitudinal survey data as the benchmark to systematically quantify documentation completeness and validate the model's clinical utility.

Study Objectives

The primary objective of this study is to develop a domain-informed NLP framework for operationalizing pain severity and functional interference outcomes directly from clinical text data that can be scaled at the population level. Leveraging our survey dataset of 3725 Veterans with longitudinal PEG measures and an encounter with the VA within a year of the survey as the benchmark, we pursue 3 specific aims:

1. Quantify how completely EHR clinical notes capture the 3 PEG domains of pain intensity and functional interference among patients with chronic pain engaging in pain management therapies.
2. Evaluate the accuracy of a best-performing LLM and a scalable, lightweight language model in extracting pain intensity and interference domains against subject matter expert (SME) annotations.
3. Assess the concordance between model-derived pain assessments from EHR encounters and proximal patient-reported PEG scores.

To our knowledge, this is the first large-scale study to validate clinical text-based pain outcome extraction against patient-reported measures in a longitudinal nationwide survey. Ultimately, we aim to establish the feasibility of using this NLP framework as a scalable, automated alternative to resource-intensive surveys, enabling health care systems to longitudinally monitor severe pain intensity and interference at the population level.

Methods

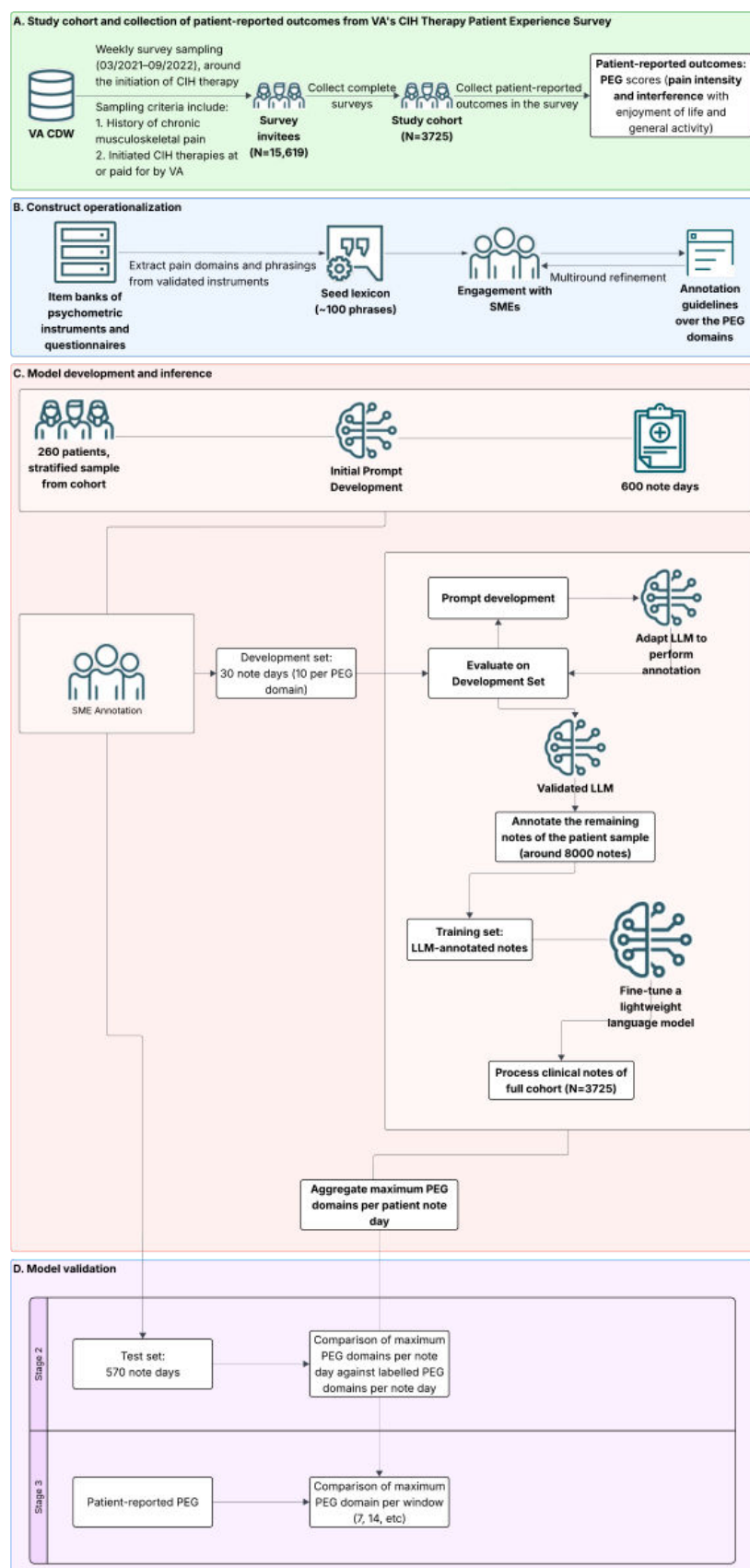
Overview

This protocol outlines an ongoing study to develop and validate an NLP framework for identifying severe pain

intensity and interference from clinical narratives across VA medical centers, using longitudinal PROs from a nationwide survey as the benchmark (Figure 1). At the time of writing, the project has finalized its methodological design and is currently in the model development phase. In the initial phase, we have completed the construct operationalization, translating PEG scale domains into a concrete seed lexicon derived from psychometrically validated instruments for use in developing annotation guidelines for identifying PEG-related concepts in clinical notes and rating each of the 3 PEG domains as either not severe ($<7/10$) or severe ($\geq 7/10$).

Currently, we are in the process of adapting LLMs within VA's secure computing environment through prompt engineering, incorporating annotation guidelines to guide evidence extraction from clinical notes. Early LLM output has identified 600 candidate encounters (200 for each of the 3 pain domains) that are being verified by SMEs. This ground truth validation dataset will be used to assess the performance of 2 candidate LLMs. Although these resource-intensive LLMs are not currently scalable for reviewing EHR notes at the population level, this phase provides initial proof-of-concept validation. To enable scalable processing at the population level, we will adopt a knowledge distillation approach, using the best-performing LLM to generate a large library of span-level annotations that will be used to train a lightweight language model. This step will also allow transparency in reviewing which key language elements the LLM parses to determine severity levels within each of the 3 PEG domains. The lightweight model will then be assessed for performance relative to the best-performing LLM. Finally, we will compare the model-generated PEG domain ratings from EHR encounters to temporally adjacent PEG scores reported by patients to assess how closely EHR reports agree with PROs. Throughout these steps, we will conduct a 3-stage validation process to characterize the completeness of pain intensity and interference documentation in clinical narratives, evaluate the inference accuracy of the LLM-as-annotator and the lightweight model, and assess concordance between the lightweight model's output and PROs captured independently from the EHR. The completed TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and Artificial Intelligence) reporting checklist [27] is provided in Checklist 1.

Figure 1. Overview of the NLP framework for identifying severe pain intensity and interference from clinical narratives. (A) Formation of the study cohort and collection of PEG scores as the benchmark. (B) Development of the seed lexicon and annotation guidelines. (C) Model development and inference, including LLM prompt refinement on SME-annotated notes, LLM annotation of the remaining patient sample, and fine-tuning of a lightweight language model applied to the full cohort. (D) Validation pipeline for both the LLM and lightweight language models. CDW: Corporate Data Warehouse; CIH: complementary and integrative health; LLM: large language model; NLP: natural language processing; PEG: pain, enjoyment, and general activity; SME: subject matter expert; VA: Veterans Health Administration.



Ethical Considerations

This study constitutes a secondary analysis of PROs collected from the VA CIH Therapy Patient Experience Survey quality improvement evaluation [15,16]. This evaluation was deemed a nonresearch activity under the Federal Policy for the Protection of Human Subjects and was initiated and executed as an internal quality improvement project for VHA operations in accordance with VA Program Guide 1200.21 [28]. The present secondary analysis is conducted under an existing Memorandum of Understanding between the study team and the VA Office of Patient-Centered Care and Cultural Transformation, which governs data access for this work. No additional institutional review board review or waiver was required for this secondary analysis, as both the survey data from the initial evaluation project and the development of NLP evaluation tools to monitor pain interference in clinical narratives were deemed nonresearch.

Data Governance and Privacy Safeguards

All analyses are conducted within VA's Health Insurance Portability and Accountability Act (HIPAA)-compliant computing environment via the VA Informatics and Computing Infrastructure. All study data, including clinical notes, structured EHR fields, and survey responses, remain within VA firewalls at all times. Open-weight LLMs and the lightweight model are hosted on local VA graphics processing unit infrastructure, with no protected health information (PHI) transmitted to external API end points or third-party services; proprietary commercial LLMs accessed through external APIs are excluded from this study's pipeline. Access to the underlying Corporate Data Warehouse is role-based and limited to personnel credentialed by the VA to conduct internal quality improvement data analytics, with all access events logged through VA Informatics and Computing Infrastructure's standard auditing infrastructure.

Study Setting

The Veterans Health Administration is the integrated health care system within the US Department of Veterans Affairs, providing care to enrolled Veterans through a national network of medical centers and outpatient sites. Enrollment is open to individuals with qualifying military service histories, although not all eligible Veterans enroll. Veterans receiving VA care carry a particularly heavy burden of chronic pain, with prevalence substantially higher than that of the general US adult population. This makes the cohort clinically important and represents a high-impact setting in which to develop scalable pain measurement tools. The Whole Health Flagship sites are 18 VA medical centers selected to lead implementation of VA's Whole Health System of Care, a patient-centered health care model emphasizing what matters to the Veteran rather than what is the matter with the Veteran.

The 6 priority CIH therapies expanded under this initiative are acupuncture, chiropractic care, massage, Tai Chi, yoga, and meditation.

Data for this study were derived from a secondary analysis of longitudinal PROs collected by Office of Patient-Centered Care and Cultural Transformation's CIH Therapy Patient Experience Survey and corresponding EHRs from the VA Corporate Data Warehouse. The survey study protocol has been published [29,30]. VA's Whole Health Flagship initiative, initially funded by the Comprehensive Addiction and Recovery Act of 2016, led to the expansion of CIH therapies across VA. The CIH Therapy Patient Experience Survey was conducted to assess the impact of this expansion on outcomes including pain severity and functional interference. Veterans invited to participate in the survey were aged 18 to 89 years, had been diagnosed with chronic musculoskeletal pain, and had newly initiated 1 of the 6 priority CIH therapies at any of the 18 VA medical centers participating in the Whole Health Flagship initiative.

Baseline surveys were sent on a weekly basis following identification of CIH therapy initiation between March 2021 and September 2022, with 6-month follow-up survey data collected through March 2023. Veterans included in this quality improvement evaluation were participants who had EHR-documented or self-reported use of a priority CIH therapy during the 6-month study period and who had completed self-reported pain assessment surveys at baseline and 6-month time points. This ensured that patients included in our study population had chronic pain and provided complete self-reported pain outcome data for analysis.

Primary Pain Outcomes

The analysis focuses on replicating the individual domains of the PEG scale, a 3-item psychometric instrument consisting of 1 item for pain severity ("P") and 2 items for pain interference with Enjoyment of life ("E") and General activity ("G") [13] (Figure 2). For each domain, the goal is to determine whether there is information in the clinical narrative deemed to represent a patient's score of severe ($\geq 7/10$) or not severe ($< 7/10$). The threshold of 7 for severe was determined from prior literature, which observed that this threshold explained the highest proportion of variance in patient-reported interference, as introduced by Serlin et al [31] and replicated in many different patient populations [32-34]. Future efforts may address methods for combining the 3 domains into a summary score and identifying finer-level gradients of pain interference; because the 3 PEG domains are certain to be documented at different frequencies in clinical narratives relative to survey collection, where all domains are structurally asked at the same time, combining the individual domains extracted from clinical narratives is a substantially different analytic task that will depend on the intermediate findings of this analysis.

Figure 2. Pain, enjoyment, and general activity scale items administered in the complementary and integrative health Therapy Patient Experience Survey.

1. What number best describes your pain on average in the past week.
012345678910
No painPain as bad as you can imagine
2. What number best describes how, during the past week, pain has interfered with your enjoyment of life ?
012345678910
Does not interfereCompletely interferes
3. What number best describes how, during the past week, pain has interfered with your general activity ?
012345678910
Does not interfereCompletely interferes

Developing an NLP Pipeline to Identify Pain Severity and Interference From Clinical Notes

Overview and Rationale

Processing clinical notes at the population level requires a scalable approach. LLMs demonstrate strong performance in understanding clinical text, but their high memory and inference time requirements make direct application to hundreds of thousands of notes impractical within current local infrastructure constraints. To address this issue, we adopt a practical design inspired by knowledge distillation, with the goal of developing a scalable, lightweight language model that can process large volumes of notes. To ensure the lightweight language model can accurately accomplish the nuanced task of evidence extraction, we will develop a comprehensive library of reference annotations using a combination of human review and expansion using an LLM applied to a sample larger than what would be feasible for efficient human review.

Construct Operationalization

Pain interference is a multifaceted and abstract construct. Operationalizing this construct and extracting the related concrete information from clinical text present significant challenges. For instance, the PEG domain “Enjoyment of life” is a high-level concept that can be too abstract for direct extraction. To extract information relevant to pain intensity and interference as outlined in the PEG scale in a consistent and grounded manner, we developed an annotation guideline derived from multiple psychometric instruments, including item banks from the 41-item Patient-Reported Outcomes Measurement Information System-Pain Interference (PROMIS-PI) scale [35], the 52-item West Haven-Yale Multidimensional Pain Inventory (WHYMPI) [36], and the 10-item Oswestry Low Back Disability Questionnaire [37]. These instruments were selected to provide specific and comprehensive coverage across the PEG domains. For example, PEG’s “Enjoyment of life” (“E”) domain can be mapped to WHYMPI’s concepts of “satisfaction or enjoyment you get from participating in social and recreational

activities.” Similarly, for PEG’s “General activity” (“G”) domain, the Oswestry Low Back Disability Questionnaire provides specific activity-based descriptions that align with documentation in clinical notes, such as “I can only walk using crutches or a cane” and “Pain prevents me from standing for more than 10 mins.” To capture PEG’s pain intensity domain (“P”), we also included explicit severity descriptors (eg, “severe pain,” “NRS 8/10”) commonly used in clinical narratives. This guideline serves as the foundation for SME annotation and as the supervision signal for our language models. We curated concepts and verbatims from these instruments to form a comprehensive lexicon of 101 pain descriptors, as listed in [Multimedia Appendix 1](#).

Validation Sample Development

We are developing a validation sample of 600 encounters, targeting 200 encounters for each of the 3 P, E, and G domains, with at least 75 severe (≥7) encounters within each domain. To efficiently generate this validation corpus, we sampled 260 patients from the survey population and identified all unique notes occurring between 1 month prior to completion of the baseline survey and 6 months following completion. To ensure relevance, all notes were required to have at least 1 instance of the word “pain” to be sampled. After this restriction, these patients had 6747 notes across 6642 unique encounters during this period. An early LLM applied to the 6747 text notes preliminarily identified notes with severe and not severe descriptions of interference across the 3 pain domains (Table 1).

Using eHost software [38], annotation is being conducted by identifying sentence spans in clinical narratives that semantically map to concepts in the annotation guideline. Three SMEs with expertise in pain research are independently annotating clinical notes for a subset of patients sampled from the study population. Each SME reviews notes, identifies phrases representing PEG domains, and assigns severity to each domain using the annotation guide, grouped into “Not severe (0-6)” and “Severe (7-10)” when appropriate descriptions are present in the text. Discrepancies are resolved through consensus adjudication among the SMEs.

Table 1. Preliminary counts of pain, enjoyment, and general activity domains and severity levels from the developmental large language model applied to the validation sample.

PEG ^a domain and severity level	Unique patients (N=260), n (%)	Total notes over 7-month period (N=6747) ^b , n (%)
Any of 3 PEG domains identified	257 (99)	5092 (75)
P (pain severity)—any mention	257 (99)	4760 (71)
P (pain severity)—severe	165 (63)	848 (13)
E (interference with enjoyment of life)—any mention	187 (72)	960 (14)
E (interference with enjoyment of life)—severe	50 (19)	88 (1)
G (interference with general activities)—any mention	237 (91)	1938 (29)
G (interference with general activities)—severe	98 (38)	287 (4)

^aPEG: pain, enjoyment, and general activity.^bNotes occurring on 6642 unique encounter dates.

Task Definition

Our pipeline involves 2 models, an LLM and a lightweight language model, each with distinct roles. After establishing proof of concept that the resource-intensive LLM can reliably identify elements of pain intensity and interference in clinical narratives, it will serve as an annotator (“LLM-as-Annotator”). Given the same information as provided to the SMEs, the LLM will receive free-text clinical narratives from our cohort. The expected output is extracted evidence mapped to the PEG domains, including relevant structured data when present within the note (eg, NRS scores) and specific sentence spans indicating pain intensity and interference, including

markers of duration that represent severe interference such as “most days.” This will efficiently expand annotation capacity beyond what SMEs can achieve. An example input and annotation output is shown in Figure 3. These annotations will then serve as the supervision signal for fine-tuning the second model, a lightweight language model—such as a sentence transformer, a Bidirectional Encoder Representations from Transformers (BERT)—based architecture, or a lightweight LLM (<4B parameters)—that performs both evidence extraction and severe pain classification for the entire cohort. We plan to use ClinicalBERT for this task.

Figure 3. Output of the LLM-as-annotator for development of the annotation training library. AUD: alcohol use disorder; BP: blood pressure; CC: chief complaint; EHR: electronic health record; HPI: history of present illness; LBP: low back pain; LLM: large language model; MDD: major depressive disorder; NRS: numeric rating scale; PDMP: prescription drug monitoring program; PEG: pain, enjoyment of life, and general activity; Pt: patient; PTSD: posttraumatic stress disorder; sx: symptoms; UDS: urine drug screen.

Example input multimodal EHR data	Desired output of the LLM
Episode 1: Primary Care Visit Date: 2024-08-15 Provider: Dr. Jane Doe, MD [Structured Data] - BP: 145/92 - Pain NRS: 8/10 - Order Entry: Hydrocodone-Apap 5-325 mg (Refill); Urine Drug Screen (Completed) [Clinical Note] CC: “Back is killing me” HPI: 68yoM w/ hx chronic pain presents requesting early refill. States pain is “unbearable” again this past week after trying to do yard work. A/P: 1. LBP exacerbation. UDS negative. Checked PDMP - consistent. Refill sent today. Discussed risks of increasing dose esp w/ hx of AUD. Pt agrees to stay on current dose.	<pre>{ "extracted_evidence_peg_domains": { "P_pain_severity": [{ "date": "2024-08-15", "source": "Structured Data: Vital Sign", "value": "NRS 8/10" }, { "date": "2024-08-15", "source": "Clinical Note (Primary Care)", "span": "States pain is 'unbearable'" }], "E_interference_enjoyment": [{ "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "missed granddaughter's birthday party", }], "G_interference_general_activity": [{ "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "Spends most days in recliner" }, { "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "couldn't make the drive" }] } }</pre>
Episode 2: Mental Health Clinic Date: 2024-09-10 Provider: Dr. John Smith, PsyD [Structured Data] - BP: 138/85 - Pain NRS: -/10 (Note: Default entry, patient not asked) - Diagnosis Code: F33.1 (Major Depressive Disorder, Recurrent, Moderate) [Clinical Note] S: Pt appears downcast. Reports increasing social isolation. Says he missed his granddaughter's birthday party last weekend. “I just couldn't make the drive, it's too much for me right now.” Spends most days in recliner. O: Affect blunted. A: MDD vs PTSD sx worsening. P: Review coping skills. Behavioral activation discussed.	<pre>{ "extracted_evidence_peg_domains": { "P_pain_severity": [{ "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "missed granddaughter's birthday party", }], "E_interference_enjoyment": [{ "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "Spends most days in recliner" }, { "date": "2024-09-10", "source": "Clinical Note (Mental Health)", "span": "couldn't make the drive" }] } }</pre>

Data Preprocessing

We will retrieve clinical notes anchored to the same temporal windows surrounding each survey time point. The data representation will be concatenated with the system prompt to the LLM, along with subsequent instructions for formatting the output.

Model Selection

We will leverage state-of-the-art open-weight (ie, models whose weights are publicly released, enabling local deployment) LLMs to perform this annotation task. Recent advances have demonstrated that LLMs pretrained on internet-scale corpora acquire extensive medical knowledge, with performance surpassing passing scores on professional medical examinations [39-42] and possessing robust medical reasoning capabilities [43]. For our application scenario, we will compare 2 open-weight models that are available within the secure VA computing environment and demonstrate strong performance on a variety of benchmarks: Gemma 4 E4B [44] and MedGemma 1.5 [45]. Comparison between the 2 candidate LLMs will be based on their agreement with SME annotations, as described in the Three-Stage Validation section below. These models will be hosted on local machines, ensuring that all computation remains within VA's secure, HIPAA-compliant environment without transmitting PHI to external servers. For the downstream fine-tuned lightweight language model, we will evaluate lightweight architectures suitable for processing the entire cohort, as discussed in the *Knowledge Distillation for Scalable Inference* section below.

Adapting LLMs to Generate Annotations

The LLMs will be adapted through prompt engineering. Core definitions from our annotation guidelines will be used to develop the prompt, with a structured JSON output schema aligned with the PEG domains. As a baseline, we will evaluate zero-shot performance using only prompt instructions. To expose the model to real-world documentation patterns, we will implement few-shot prompting with up to 5 representative examples selected from our development annotation set. These examples will cover diverse clinical scenarios and documentation styles, each pairing raw clinical notes with reference JSON outputs containing extracted evidence.

To mitigate hallucination risk, we will deploy a chain-of-thought (CoT) output schema requiring a sequential reasoning method [46] where the model first identifies and quotes specific text spans indicating pain severity or functional interference, then maps this evidence to the corresponding PEG domains. By grounding outputs in cited evidence, the approach will enable systematic error analysis and iterative prompt refinement. This CoT reasoning is also important for complicated cases involving comorbidities. For example, social withdrawal documented in a mental health note may initially appear to stem from depression; however, if pain is documented elsewhere in the note as impacting the patient's mood, this functional limitation should also be considered as pain attributable. By requiring the model to first extract all

relevant evidence before making attributions, the CoT schema ensures such contextual information is accounted for. The transparency granted by CoT reasoning also facilitates SME validation of LLM outputs, as human experts can directly audit the extracted evidence and reasoning process.

Knowledge Distillation for Scalable Inference

To support scalable inference, we will adopt a knowledge distillation approach wherein LLM-generated annotations serve as the supervision signal for training a lightweight language model capable of processing the entire patient cohort. Our primary focus is ClinicalBERT [47], a domain-specific BERT-based architecture [48] pretrained on clinical and biomedical text, selected for its computational efficiency and its ability to be fully fine-tuned on local infrastructure to scale to population-level monitoring. Standard BERT-based models have a 512-token context limit, which may be insufficient for lengthy clinical notes. To address this, we will evaluate strategies including truncation to retain the most informative sections using rule-based sectionization found in medspaCy [49], as well as long-context encoder architectures such as Clinical ModernBERT [50], which supports up to 8192 tokens while maintaining the computational efficiency required for full-cohort processing.

Fine-tuning will be implemented via the Hugging Face Transformers library in an information extraction framework. Hyperparameters will be selected through k-fold cross-validation on the LLM-generated data, using F_1 -score as the focus metric, with early stopping to mitigate overfitting. The highest-performing hyperparameter configuration, as measured by F_1 -score, and the highest-performing overall configuration will be used to process all notes for the cohort, generating phrases containing PEG concepts and interference labels for all patients.

Computational Feasibility and Optimization

Our local graphics processing unit infrastructure (4 NVIDIA L40S, 48 GB) supports inference of models up to 70B parameters at 16-bit floating-point precision. For further memory optimization, we may additionally implement 4-bit quantization [51], which significantly reduces the memory footprint compared with full precision. This allows us to deploy all open-weight models on local machines, guaranteeing that all computation is contained entirely within the VA's secure, HIPAA-compliant environment without transmitting PHI to external servers.

Three-Stage Validation for Our NLP Pipeline

Overview of 3-Stage Validation Process

To assess the reliability and clinical utility of our proposed approach, we will employ a 3-stage validation process. First, we will characterize the completeness of pain documentation in clinical narratives against SME-annotated references, quantifying how the 3 PEG domains are covered across notes, patients, and clinical settings. Second, we will evaluate the inference accuracy of the LLM-as-annotator, comparing 2 candidate LLMs (Gemma 4 E4B and MedGemma 1.5), and

of the fine-tuned lightweight model against SME annotations across both note-level classification of pain domain and severity and span-level localization of pain-relevant text. Third, we will test the population-level utility of the fine-tuned lightweight model by assessing concordance between PROs captured contemporaneously and lightweight model classifications of proximal clinical narratives.

Stage 1 and Stage 2 validation activities will be conducted with the sample of 260 patients, along with their 600 sampled notes, who participated in the CIH Therapy Patient Experience Survey; Stage 3 will be conducted among the full sample of 3725 survey participants. The validation sample of 260 patients includes 6747 unique text notes that include the keyword “pain” over a 7-month period across 6642 unique encounters.

Stage 1: Documentation of Pain Intensity and Interference in Clinical Narratives

The first validation measures the extent to which EHR documentation can be assessed for meaningful elements of pain intensity and pain interference, using the targeted set of 600 SME-annotated notes identified from the preliminary LLM prompt as the basis for this analysis. As noted in [Table 1](#), the frequency of PEG domains may vary considerably, with the E domain reported less frequently. The goal of this validation stage is to confirm that the core elements defined in the annotation guidelines and used by the LLMs are reliably identified in clinical narratives. The analysis unit for this stage is the full encounter day, with SMEs reviewing all notes from an encounter date to ensure all contextual information is assessed. The most severe intensity or interference level on the day will be assessed by the SME. This activity will provide additional training for both refinement of the LLM prompts and span-level annotation for fine-tuning of the lightweight model, as SMEs will be asked to highlight specific text spans they used in their judgment to assess each of the PEG domains. We define the following metrics:

- **Interpretability:** we report the proportion of notes in which pain was mentioned but SMEs could not determine domain or severity (“indeterminate”), and the proportion of spans requiring consensus adjudication.
- **Interrater reliability:** prior to any adjudication, agreement among the 3 SMEs will be quantified using Fleiss κ for the binary severity classification and for per-domain presence/absence. Because the sample is enriched for pain-relevant content by design, we note that these completeness metrics describe interpretable content within candidate notes rather than population-level documentation prevalence. Population-level documentation metrics will be assessed in subsequent validation stages using ClinicalBERT.

While preliminary examination of notes and our early LLM prompting have suggested there are extensive pain-related descriptions documented in clinical notes, this validation will shed light on potential gaps in EHR-documented clinical narratives. Potential limitations may arise due to limited provider documentation or insufficient descriptions by patients, leading to mentions of pain interference but

uncertainty surrounding the extent of severity. To identify sources of bias prior to model deployment, we will examine how these metrics vary by clinical characteristics including age and sex.

In addition, we will record the proportion of notes in which a specific PEG domain is not mentioned as documentation gaps. Documentation gaps may arise from multiple sources, including limited symptom presence during the temporal window, provider documentation practices, and patient self-reporting patterns during clinical encounters. Our framework does not distinguish among these sources at the level of individual observations, as they collectively contribute to the observable evidence density from routine clinical documentation. Documentation gaps are characterized as observations of this evidence density rather than treated as missing data.

Stage 2: Inference Accuracy of the LLM-as-Annotator and Lightweight Model Against SME Annotation

LLM-as-Annotator Evaluation

Conditional upon documentation completeness established in stage 1, this stage evaluates whether an LLM can serve as a reliable annotator, extending SME annotation capacity to a scale suitable for training a downstream lightweight model. Two candidate LLMs, Gemma 4 E4B [44] and MedGemma 1.5 [45], will be evaluated against the 600 SME-annotated notes across 2 complementary sub-analyses: note-level classification and span-level localization.

Note-Level Classification of Domain and Severity

The purpose of this subanalysis is to demonstrate that an LLM can correctly identify descriptions of pain from text records and correctly classify its domain and severity. The LLM-as-annotator will be evaluated against the reference human annotations from the 600-encounter validation set to assess agreement with SME judgments ([Multimedia Appendix 2](#)). Each LLM’s performance will be calculated for each PEG domain and for binary level of severity if considered to contain elements suggestive of a level ≥ 7 , to validate how it performs against each of the individual domains and overall intensity. Precision, recall, and F_1 -scores will be computed for each, with 95% CIs constructed by bootstrapping individuals and their corresponding notes, with replacement, to account for repeated measures within individuals (1000 replicates). An overall model-specific F_1 measure will also be reported to compare the 2 candidate LLMs.

Span-Level Localization

Conditional on the note/visit-level analysis identifying an LLM that can accurately identify the key PEG domains and levels of severity from patient notes, this subanalysis evaluates whether the LLM-as-annotator can identify the specific span of text and assign the correct pain domain and severity label. The goal is to demonstrate that using the LLM

to develop a large dataset of span annotations maintains a high quality of accuracy. This large training dataset supports development of the lightweight language model, which can then scale to the entire patient cohort (Figure 3).

We first evaluate whether the LLM extracts the same pain-relevant text spans as our human SMEs. Given the generative nature of LLMs, outputs may not be verbatim copies of source text. We will employ 2 matching criteria:

- Relaxed string match: A match is recorded if the extracted span shares tokens with the annotated span, relaxing exact boundary requirements.
- Semantic equivalence: We will use an LLM-as-a-judge approach, prompting a separate LLM to determine whether extracted content preserves the same meaning as the annotation. A recent study has shown that LLM-as-a-judge provides a scalable way to identify accurate and safe LLM-generated clinical summaries [52].

For both criteria, we will report precision, recall, and F_1 -score. Precision measures the proportion of LLM-extracted spans that overlap with reference annotations. Any nonoverlapping extractions that do not exist in the source text constitute hallucinations. Recall measures the proportion of reference annotations successfully retrieved by the model. CIs for each performance metric will be constructed by bootstrapping individuals in the validation set, with replacement, to account for repeated observations within individuals (1000 replicates). For each of the 3 PEG domains, we will compare spans identified by the LLM to the SME annotations and report the full confusion matrix along with per-class precision, recall, and F_1 with bootstrapped 95% CIs. We will separately report the same metrics for the binary severity label (severe vs not severe) for each of the 3 domains.

Once the fine-tuned lightweight model is completed, we will repeat both note-level classification and span-level localization subanalyses using the lightweight model against the same 600 SME-annotated encounters. This provides direct evaluation of the lightweight model's inference accuracy against the SME reference, alongside the LLM comparison.

Stage 3: Concordance Between the Scalable Lightweight Model and PROs

The lightweight model is designed for scalable, population-level deployment rather than individual-level diagnostic classification. Accordingly, the evaluation of concordance is framed around whether the model's output can serve as a reliable surrogate for PROs at population scale, complementing rather than replacing direct patient self-report. As such, we will assess the concordance between severe pain and interference identified by the lightweight model in clinical narratives and patient-reported PEG scores in the full 3725-patient cohort, with patient-reported scores serving as the benchmark (Multimedia Appendix 3). This analysis will be performed on the baseline data for each individual.

The primary discrimination metric is area under the receiver operating characteristic curve, computed individually for each PEG domain using the proportion of severe pain

and interference notes within a survey-anchored window as the continuous ranking score against the patient's binary per-domain survey label (severe $\geq 7/10$ vs not severe $< 7/10$).

Patients without any pain-related clinical narrative within a given temporal window are excluded from that window's analysis, as no narrative observation exists. This exclusion is reflected in the reduced sample size at narrower windows (Multimedia Appendix 3). Patients with pain-related narratives but without documentation of a specific PEG domain are retained as observations of low evidence density for that domain.

In addition to discrimination, we will assess calibration to evaluate how closely model-predicted rates of severe pain interference align with patient-reported rates at the aggregate level. Calibration will be reported through 3 complementary measures. Calibration plots will bin patients by decile of predicted probability and compare predicted rates against observed patient-reported rates within each bin. Brier scores will summarize overall probabilistic performance across the full cohort, with lower values indicating closer agreement between predicted and observed outcomes. Calibration-in-the-large will be computed as the ratio of predicted to observed severe pain interference prevalence, providing a single-number summary of systematic overprediction or underprediction.

Each analysis will be repeated for a range of temporal windows around the survey date (± 7 , ± 30 , ± 90 , and ± 180 d) to assess how the breadth of the aggregation window affects concordance. The 7-day window reflects the anchoring period of the PEG instrument, which asks patients to consider the past week. All point estimates will be reported with 95% CIs derived from a nonparametric bootstrap resampling patients as the independent unit over 1000 replicates.

Given the model's intended role as a population-scale complement to represent a similar concept as patient self-report of pain severity and interference, the following minimum performance criterion is calibrated to surrogate utility rather than individual-level diagnostic accuracy. The prespecified minimum performance criterion goal for Stage 3 is area under the receiver operating characteristic curve ≥ 0.75 for each of the 3 domains, requiring discrimination meaningfully above chance while accommodating the attenuation expected when validating an EHR-derived surrogate against patients' self-reports. We anticipate performance will deteriorate as encounters/clinical narratives become less temporally related to the survey date.

Performance will additionally be evaluated within prespecified subgroups of interest (age, sex, primary pain diagnosis, and clinic type) to determine whether the model systematically overestimates or underestimates severe pain interference for these subgroups. Subgroup differences will be assessed by comparing bootstrap-derived confidence intervals across subgroups.

Together, the 3 validation stages assess the full pipeline from documentation completeness to model inference accuracy and population-level utility. These results will determine the feasibility of monitoring severe pain and

interference at population scale and whether our NLP framework can serve as a scalable surrogate for resource-intensive patient-reported surveys.

Results

This study is ongoing, with final results expected in 2027. As of July 2026, the study cohort of 3725 Veterans with chronic musculoskeletal pain who were new users of CIH therapies by survey administration date has been identified. The longitudinal PEG survey data serving as the benchmark have been fully aggregated. Demographic and clinical characteristics of the cohort have been previously published [53].

The domain-specific seed lexicon has been developed, curating 101 pain descriptors from established psychometric instruments including PROMIS-PI, WHYMPI, and the Oswestry Low Back Disability Questionnaire. Annotation guidelines mapping these concepts to each of the 3 PEG domains have been developed and are being iteratively refined through engagement with SMEs. We have applied a developmental LLM to screen 6747 text notes among 6642 unique encounters over a 7-month period for 260 patients, with the preliminary LLM identifying at least 1 of the 3 PEG domains in 75% of notes and 99% of patients (Table 1). Validation at the note level, targeting 200 notes for each of the 3 PEG domains, is in progress as of July 2026. Upon completion of the validation-sample annotation, the study will proceed to the subsequent knowledge distillation and validation stages.

Discussion

Principal Implications

Chronic pain is a complex health issue. Current screening tools embedded in structured EHR data cannot fully capture this complexity. Our framework extracts multidimensional pain information, including pain severity and interference with physical and emotional functioning, directly from clinical narratives at scale. If validated, this approach could offer a scalable alternative to resource-intensive surveys for monitoring severe pain and interference. To the best of our knowledge, this is the first study to investigate the feasibility of using EHR data, validated against large-scale PROs, for routine monitoring of severe pain and interference.

Methodological Significance

This study adopts a domain-informed approach, grounding the LLM's inference in a seed lexicon derived from established psychometric instruments. This enables rigorous mapping of abstract pain-related PEG domains into concrete terms found in clinical documentation. To enable scalable processing, our approach employs knowledge distillation: an LLM first generates high-quality annotations on a subset of notes, which then serve as training data for a lightweight language model capable of processing the entire cohort. This design leverages LLM capabilities for nuanced evidence extraction while enabling scalable inference through a lightweight model. Our validation strategy extends beyond

standard benchmarking against expert annotations. By first examining documentation completeness, we will assess whether disparities in clinical documentation practices, which have been identified in prior work [24-26] and may introduce systemic bias, exist in our patient population. By validating against PROs, we will evaluate whether the framework's inferences align with how patients themselves report their pain and functional interference.

Clinical and Operational Utility

Subject to successful validation, this framework could improve how pain is monitored in health care systems. Surveys collecting PROs often suffer from low response rates and nonresponse bias, as prior research found that nonresponders had higher pain-catastrophizing scores and more pain at enrollment [12]. A validated NLP tool could serve as a passive approach to identify severe pain and functional interference, complementing active survey collection and reaching patients who are engaged in care but unreached by surveys. A validated severe pain phenotype would also enable longitudinal monitoring of treatment outcomes. While randomized controlled trials establish efficacy, they are limited in tracking long-term functional improvements at scale. By extracting pain and functional interference directly from EHR data, researchers could evaluate the efficacy of high-cost or invasive interventions such as spinal cord stimulation in reducing severe pain and interference in broader, real-world populations.

Limitations

There are several limitations to this protocol. First, the LLMs' inference relies on EHR documentation, which may be incomplete in recording the full picture of patients' experience of pain. We address this limitation by quantifying documentation completeness in our validation. Second, pain fluctuates over time, and temporal gaps between clinical visits and survey dates may reduce concordance between EHR-derived and survey-reported outcomes. Third, significant heterogeneity in clinical documentation practices and reports of interference from pain will likely lead to underidentification of interference. Fourth, documentation biases in the EHR may propagate into model inference. If pain is systematically underdocumented for certain patient populations, the language model's inference for those groups may inherit this bias. We will conduct subgroup analyses by demographic and socioeconomic characteristics, comparing both documentation patterns (eg, the frequency and coverage of the PEG domains in annotations) and model inference accuracy metrics (eg, sensitivity, concordance with PROs) across groups to identify such disparities.

Fifth, patient representatives were not directly consulted during the development of the seed lexicon, annotation guidelines, or interpretation of model outputs in this secondary analysis, as the parent CIH Therapy Patient Experience Survey was conducted between 2021 and 2023 and has since concluded. This gap is partially mitigated by the fact that patient perspectives are embedded indirectly through the psychometric instruments used to construct the seed lexicon (PROMIS-PI, WHYMPI, Oswestry), each of which

was developed with patient-engaged validation. Nonetheless, future deployment work would benefit from direct patient input on interpretability and face validity.

Finally, the direct comparison between aggregated NLP-derived domain evidence and PEG survey scores is an imperfect proxy for true concordance. Clinical documentation of pain is episodic and contextually driven, whereas the PEG reflects a patient's deliberate summary of their experience over time; simple aggregation of sparse span-level extractions may not fully bridge this representational gap. A natural extension of this work, contingent on demonstrating adequate NLP accuracy against SME annotations, would be to train a supervised model that learns to map heterogeneous, potentially sparse clinical documentation

patterns onto PEG-equivalent scores, rather than relying on direct aggregation.

Conclusion

This protocol outlines a framework for identifying severe pain intensity and interference from clinical narratives, addressing a gap in health care system surveillance. By grounding inference in established psychometric instruments and validating against PROs, we aim to develop an NLP-empowered tool that can reliably extract pain severity and functional interference from EHR data. If successful, this approach could provide a scalable complement to resource-intensive surveys, enabling longitudinal monitoring of severe pain interference at the population level.

Acknowledgments

This work was conducted using resources and facilities of the VA Puget Sound Health Care System. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States Government.

The authors used generative AI tools (Claude Opus 4; Anthropic) for initial language editing and polishing of the manuscript text. All scientific content and final text were reviewed and verified by the authors, who take full responsibility for the publication.

Funding

This work was supported by the VA Health Systems Research Center of Innovation for Veteran-Centered and Value-Driven Care (COIN 13-402) and the VA Office of Patient-Centered Care and Cultural Transformation (OPCC&CT), Quality Enhancement Research Initiative (QUERI) award number PEC 13-001. The funders provided support for investigator time, computing infrastructure, and, in the case of OPCC&CT, access to the underlying CIH Therapy Patient Experience Survey data and linked electronic health records via the Memorandum of Understanding described in the Ethics Approval section. The funders had no role in the design of this secondary analysis, in the statistical analysis plan, in the interpretation of results, or in the decision to submit the manuscript for publication.

Data Availability

The data used in this study were obtained from the VA Corporate Data Warehouse and contain protected health information. In accordance with VA data governance policies, these data cannot be made publicly available. Requests for access to VA data may be directed to the VA Informatics and Computing Infrastructure [54].

Authors' Contributions

All authors contributed to the intellectual development of this study. Individual contributions are as follows:

Conceptualization: XZ, CRW, HE, SBZ

Data curation: CRW, HE, AK

Funding acquisition: SBZ

Methodology: XZ, CRW, HE, DER, TYHW, AK, GL, SBZ

Project administration: AK, EWR

Software: CRW, HE

Supervision: GL, SBZ

Writing – original draft: XZ

Writing – review & editing: CRW, HE, DER, TYHW, AK, EWR, GL, SBZ

Conflicts of interest

None declared.

Multimedia Appendix 1

Seed lexicon of 101 terms derived from validated pain assessment instruments and mapped to the 3 PEG domains (Pain severity, Enjoyment of life, General activity).

[\[PDF File \(Adobe File\), 147 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Shell table—encounter-level classification accuracy of candidate large language models against subject matter expert annotation for each pain severity, enjoyment of life, general activity domain and for binary severity level ≥ 7 .

[\[DOCX File \(Microsoft Word File\), 16 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Shell table (partial)—final validation reporting of Lightweight Model and survey responses.

[\[DOCX File \(Microsoft Word File\), 16 KB-Multimedia Appendix 3\]](#)

Checklist 1

TRIPOD+AI checklist.

[\[PDF File \(Adobe File\), 235 KB-Checklist 1\]](#)

References

1. Lucas JW, Sohi I. Chronic pain and high-impact chronic pain in U.S. adults, 2023. NCHS Data Brief. Oct 2024;(518):CS355235. [doi: [10.15620/cdc/169630](https://doi.org/10.15620/cdc/169630)] [Medline: [39751180](https://pubmed.ncbi.nlm.nih.gov/39751180/)]
2. Guy GP Jr, Miller GF, Legha JK, et al. Economic costs of chronic pain-United States, 2021. Med Care. Sep 1, 2025;63(9):679-685. [doi: [10.1097/MLR.0000000000002181](https://doi.org/10.1097/MLR.0000000000002181)] [Medline: [40730349](https://pubmed.ncbi.nlm.nih.gov/40730349/)]
3. Mills SEE, Nicolson KP, Smith BH. Chronic pain: a review of its epidemiology and associated factors in population-based studies. Br J Anaesth. Aug 2019;123(2):e273-e283. [doi: [10.1016/j.bja.2019.03.023](https://doi.org/10.1016/j.bja.2019.03.023)] [Medline: [31079836](https://pubmed.ncbi.nlm.nih.gov/31079836/)]
4. Taylor KA, Kapos FP, Sharpe JA, Kosinski AS, Rhon DI, Goode AP. Seventeen-year national pain prevalence trends among U.S. military veterans. J Pain. May 2024;25(5):104420. [doi: [10.1016/j.jpain.2023.11.003](https://doi.org/10.1016/j.jpain.2023.11.003)] [Medline: [37952861](https://pubmed.ncbi.nlm.nih.gov/37952861/)]
5. Luther SL, Finch DK, Bouayad L, et al. Measuring pain care quality in the Veterans Health Administration primary care setting. Pain. Jun 1, 2022;163(6):e715-e724. [doi: [10.1097/j.pain.0000000000002477](https://doi.org/10.1097/j.pain.0000000000002477)] [Medline: [34724683](https://pubmed.ncbi.nlm.nih.gov/34724683/)]
6. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. Jan 2005;113(1-2):9-19. [doi: [10.1016/j.pain.2004.09.012](https://doi.org/10.1016/j.pain.2004.09.012)] [Medline: [15621359](https://pubmed.ncbi.nlm.nih.gov/15621359/)]
7. Wandner LD, Domenichiello AF, Beierlein J, et al. NIH's Helping to End Addiction Long-termSM Initiative (NIH HEAL Initiative) Clinical Pain Management Common Data Element Program. J Pain. Mar 2022;23(3):370-378. [doi: [10.1016/j.jpain.2021.08.005](https://doi.org/10.1016/j.jpain.2021.08.005)] [Medline: [34508905](https://pubmed.ncbi.nlm.nih.gov/34508905/)]
8. Pain as the 5th Vital Sign Toolkit. Geriatrics and Extended Care Strategic Healthcare Group, National Pain Management Coordinating Committee, Veterans Health Administration; 2000. URL: <https://www.va.gov/painmanagement/docs/toolkit.pdf> [Accessed 2025-11-12]
9. Mularski RA, White-Chu F, Overbay D, Miller L, Asch SM, Ganzini L. Measuring pain as the 5th vital sign does not improve quality of pain management. J Gen Intern Med. Jun 2006;21(6):607-612. [doi: [10.1111/j.1525-1497.2006.00415.x](https://doi.org/10.1111/j.1525-1497.2006.00415.x)] [Medline: [16808744](https://pubmed.ncbi.nlm.nih.gov/16808744/)]
10. Scher C, Meador L, Van Cleave JH, Reid MC. Moving beyond pain as the fifth vital sign and patient satisfaction scores to improve pain care in the 21st century. Pain Manag Nurs. Apr 2018;19(2):125-129. [doi: [10.1016/j.pmn.2017.10.010](https://doi.org/10.1016/j.pmn.2017.10.010)] [Medline: [29249620](https://pubmed.ncbi.nlm.nih.gov/29249620/)]
11. Owen-Smith A, Mayhew M, Leo MC, et al. Automating collection of pain-related patient-reported outcomes to enhance clinical care and research. J Gen Intern Med. May 2018;33(Suppl 1):31-37. [doi: [10.1007/s11606-018-4326-9](https://doi.org/10.1007/s11606-018-4326-9)] [Medline: [29633139](https://pubmed.ncbi.nlm.nih.gov/29633139/)]
12. Nugent SM, Lovejoy TI, Shull S, Dobscha SK, Morasco BJ. Associations of pain numeric rating scale scores collected during usual care with research administered patient reported pain outcomes. Pain Med. Oct 8, 2021;22(10):2235-2241. [doi: [10.1093/pm/pnab110](https://doi.org/10.1093/pm/pnab110)] [Medline: [33749760](https://pubmed.ncbi.nlm.nih.gov/33749760/)]
13. Krebs EE, Lorenz KA, Bair MJ, et al. Development and initial validation of the PEG, a three-item scale assessing pain intensity and interference. J Gen Intern Med. Jun 2009;24(6):733-738. [doi: [10.1007/s11606-009-0981-1](https://doi.org/10.1007/s11606-009-0981-1)] [Medline: [19418100](https://pubmed.ncbi.nlm.nih.gov/19418100/)]
14. Krebs EE, Carey TS, Weinberger M. Accuracy of the pain numeric rating scale as a screening test in primary care. J Gen Intern Med. Oct 2007;22(10):1453-1458. [doi: [10.1007/s11606-007-0321-2](https://doi.org/10.1007/s11606-007-0321-2)] [Medline: [17668269](https://pubmed.ncbi.nlm.nih.gov/17668269/)]
15. Taylor SL, Elwy AR, Bokhour BG, et al. Measuring patient-reported use and outcomes from complementary and integrative health therapies: development of the Complementary and Integrative Health Therapy Patient Experience Survey. Glob Adv Integr Med Health. 2024;13:27536130241241259. [doi: [10.1177/27536130241241259](https://doi.org/10.1177/27536130241241259)] [Medline: [38585239](https://pubmed.ncbi.nlm.nih.gov/38585239/)]
16. Calvert C, Taylor SL, Olson J, et al. Complementary and integrative health therapies and pain: delivery through Veterans Affairs and community care. Glob Adv Integr Med Health. 2025;14:27536130251358757. [doi: [10.1177/27536130251358757](https://doi.org/10.1177/27536130251358757)] [Medline: [40657238](https://pubmed.ncbi.nlm.nih.gov/40657238/)]
17. Zhang M, Zhu L, Lin SY, et al. Using artificial intelligence to improve pain assessment and pain management: a scoping review. J Am Med Inform Assoc. Feb 16, 2023;30(3):570-587. [doi: [10.1093/jamia/ocac231](https://doi.org/10.1093/jamia/ocac231)] [Medline: [36458955](https://pubmed.ncbi.nlm.nih.gov/36458955/)]

18. Sim JA, Huang X, Horan MR, Baker JN, Huang IC. Using natural language processing to analyze unstructured patient-reported outcomes data derived from electronic health records for cancer populations: a systematic review. *Expert Rev Pharmacoecon Outcomes Res.* Apr 2024;24(4):467-475. [doi: [10.1080/14737167.2024.2322664](https://doi.org/10.1080/14737167.2024.2322664)] [Medline: [38383308](#)]
19. Dave AD, Ruano G, Kost J, Wang X. Automated extraction of pain symptoms: a natural language approach using electronic health records. *Pain Physician.* Mar 2022;25(2):E245-E254. [Medline: [35322976](#)]
20. Lu Z, Sim JA, Wang JX, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. *J Med Internet Res.* Nov 3, 2021;23(11):e26777. [doi: [10.2196/26777](https://doi.org/10.2196/26777)] [Medline: [34730546](#)]
21. Vaid A, Landi I, Nadkarni G, Nabeel I. Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders. *Lancet Digit Health.* Oct 26, 2023;5(12):e855-e858. [doi: [10.1016/S2589-7500\(23\)00202-9](https://doi.org/10.1016/S2589-7500(23)00202-9)] [Medline: [39492289](#)]
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan SVN, Garnett R, editors. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates Inc; 2017:5998-6008. ISBN: 9781510860964
23. Amidei J, Nieto R, Kaltenbrunner A, Ferreira De Sá JG, Serrat M, Albajes K. Exploring the capacity of large language models to assess the chronic pain experience: algorithm development and validation. *J Med Internet Res.* Mar 31, 2025;27:e65903. [doi: [10.2196/65903](https://doi.org/10.2196/65903)] [Medline: [40163858](#)]
24. Chaturvedi J, Stewart R, Ashworth M, Roberts A. Distributions of recorded pain in mental health records: a natural language processing based study. *BMJ Open.* Apr 19, 2024;14(4):e079923. [doi: [10.1136/bmjopen-2023-079923](https://doi.org/10.1136/bmjopen-2023-079923)] [Medline: [38642997](#)]
25. Carlson LA, Jeffery MM, Fu S, et al. Characterizing chronic pain episodes in clinical text at two health care systems: comprehensive annotation and corpus analysis. *JMIR Med Inform.* Nov 16, 2020;8(11):e18659. [doi: [10.2196/18659](https://doi.org/10.2196/18659)] [Medline: [33108311](#)]
26. Fodeh SJ, Finch D, Bouayad L, et al. Classifying clinical notes with pain assessment using machine learning. *Med Biol Eng Comput.* Jul 2018;56(7):1285-1292. [doi: [10.1007/s11517-017-1772-1](https://doi.org/10.1007/s11517-017-1772-1)] [Medline: [29280092](#)]
27. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* Apr 16, 2024;385:e078378. [doi: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378)] [Medline: [38626948](#)]
28. Office of Research and Development. VHA operations activities that may constitute research. Department of Veterans Affairs; Jan 9, 2019. URL: <https://www.research.va.gov/resources/policies/ProgramGuide-1200-21-VHA-Operations-Activities.pdf> [Accessed 2026-09-12]
29. Zeliadt SB, Coggeshall S, Thomas E, Gelman H, Taylor SL. The APPROACH trial: assessing pain, patient-reported outcomes, and complementary and integrative health. *Clin Trials.* Aug 2020;17(4):351-359. [doi: [10.1177/1740774520928399](https://doi.org/10.1177/1740774520928399)] [Medline: [32522024](#)]
30. Zeliadt SB, Coggeshall S, Gelman H, et al. Assessing the relative effectiveness of combining self-care with practitioner-delivered complementary and integrative health therapies to improve pain in a pragmatic trial. *Pain Med.* Dec 12, 2020;21(Suppl 2):S100-S109. [doi: [10.1093/pm/pnaa349](https://doi.org/10.1093/pm/pnaa349)] [Medline: [33313736](#)]
31. Serlin RC, Mendoza TR, Nakamura Y, Edwards KR, Cleeland CS. When is cancer pain mild, moderate or severe? Grading pain severity by its interference with function. *Pain.* May 1995;61(2):277-284. [doi: [10.1016/0304-3959\(94\)00178-H](https://doi.org/10.1016/0304-3959(94)00178-H)] [Medline: [7659438](#)]
32. Boonstra AM, Stewart RE, Köke AJA, et al. Cut-off points for mild, moderate, and severe pain on the Numeric Rating Scale for pain in patients with chronic musculoskeletal pain: variability and influence of sex and catastrophizing. *Front Psychol.* 2016;7:1466. [doi: [10.3389/fpsyg.2016.01466](https://doi.org/10.3389/fpsyg.2016.01466)] [Medline: [27746750](#)]
33. Hanley MA, Masedo A, Jensen MP, Cardenas D, Turner JA. Pain interference in persons with spinal cord injury: classification of mild, moderate, and severe pain. *J Pain.* Feb 2006;7(2):129-133. [doi: [10.1016/j.jpain.2005.09.011](https://doi.org/10.1016/j.jpain.2005.09.011)] [Medline: [16459278](#)]
34. Jensen MP, Smith DG, Ehde DM, Robinsin LR. Pain site and the effects of amputation pain: further clarification of the meaning of mild, moderate, and severe pain. *Pain.* Apr 2001;91(3):317-322. [doi: [10.1016/S0304-3959\(00\)00459-0](https://doi.org/10.1016/S0304-3959(00)00459-0)] [Medline: [11275389](#)]
35. Amtmann D, Cook KF, Jensen MP, et al. Development of a PROMIS item bank to measure pain interference. *Pain.* Jul 2010;150(1):173-182. [doi: [10.1016/j.pain.2010.04.025](https://doi.org/10.1016/j.pain.2010.04.025)] [Medline: [20554116](#)]
36. Kerns RD, Turk DC, Rudy TE. The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). *Pain.* Dec 1985;23(4):345-356. [doi: [10.1016/0304-3959\(85\)90004-1](https://doi.org/10.1016/0304-3959(85)90004-1)] [Medline: [4088697](#)]
37. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy.* Aug 1980;66(8):271-273. [Medline: [6450426](#)]

38. South B, Shen S, Leng J, Forbush T, DuVall S, Chapman W. A prototype tool set to support machine-assisted annotation. In: Cohen KB, Demner-Fushman D, Ananiadou S, Webber B, Tsujii J, Pestian J, editors. *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics; 2012:130-139. URL: <https://aclanthology.org/W12-2416/> [Accessed 2026-09-12]
39. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
40. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. Mar 2025;31(3):943-950. [doi: [10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7)] [Medline: [39779926](https://pubmed.ncbi.nlm.nih.gov/39779926/)]
41. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*. Preprint posted online on Mar 20, 2023. [doi: [10.48550/arXiv.2303.13375](https://doi.org/10.48550/arXiv.2303.13375)]
42. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*. Jul 12, 2021;11(14):6421. [doi: [10.3390/app11146421](https://doi.org/10.3390/app11146421)]
43. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. *arXiv*. Preprint posted online on Jan 8, 2024. [doi: [10.48550/arXiv.2401.04088](https://doi.org/10.48550/arXiv.2401.04088)]
44. google/gemma-4-E4B. Hugging Face. 2026. URL: <https://huggingface.co/google/gemma-4-E4B> [Accessed 2026-05-28]
45. Sellergren A, Gao C, Mahvar F, et al. MedGemma 1.5 technical report. *arXiv*. Preprint posted online on Apr 6, 2026. [doi: [10.48550/arXiv.2604.05081](https://doi.org/10.48550/arXiv.2604.05081)]
46. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proc Int Conf Neural Inf Process Syst*. 2022:24824-24837. [doi: [10.52202/068431-1800](https://doi.org/10.52202/068431-1800)]
47. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78. [doi: [10.18653/v1/W19-1909](https://doi.org/10.18653/v1/W19-1909)]
48. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *Proc Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol*. 2019:4171-4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
49. Eyre H, Chapman AB, Peterson KS, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc*. 2022;2021:438-447. [Medline: [35308962](https://pubmed.ncbi.nlm.nih.gov/35308962/)]
50. Lee SA, Wu A, Chiang JN. Clinical ModernBERT: an efficient and long context encoder for biomedical text. *arXiv*. Preprint posted online on Apr 4, 2025. [doi: [10.48550/arXiv.2504.03964](https://doi.org/10.48550/arXiv.2504.03964)]
51. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. *Adv Neural Inf Process Syst*. 2023:10088-10115. [doi: [10.52202/075280-0441](https://doi.org/10.52202/075280-0441)]
52. Croxford E, Gao Y, First E, et al. Automating evaluation of AI text generation in healthcare with a large language model (LLM)-as-a-judge. *medRxiv*. May 6, 2025:2025.04.22.25326219. [doi: [10.1101/2025.04.22.25326219](https://doi.org/10.1101/2025.04.22.25326219)] [Medline: [40313300](https://pubmed.ncbi.nlm.nih.gov/40313300/)]
53. Zeliadt SB, Coggeshall SS, Bokhour B, et al. Adding self-care complementary and integrative health therapies to care for chronic pain: the Assessing Pain, Patient Reported Outcomes and Complementary Health (APPROACH) study. *Med Care*. May 1, 2026;64(5):283-292. [doi: [10.1097/MLR.0000000000002295](https://doi.org/10.1097/MLR.0000000000002295)] [Medline: [41771006](https://pubmed.ncbi.nlm.nih.gov/41771006/)]
54. VA Informatics and Computing Infrastructure (VINCI). US Department of Veterans Affairs. URL: <https://www.research.va.gov/programs/vinci/> [Accessed 2026-09-12]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CDW: Corporate Data Warehouse
CIH: complementary and integrative health
CoT: chain-of-thought
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
IMPACT: Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
LLM: large language model
NLP: natural language processing
NRS: Numerical Rating Scale
PEG: pain, enjoyment, and general activity
PHI: protected health information
PRO: patient-reported outcome
PROMIS-PI: Patient-Reported Outcomes Measurement Information System-Pain Interference

SME: subject matter expert

TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and Artificial Intelligence

VA: Veterans Health Administration

WHYMPI: West Haven-Yale Multidimensional Pain Inventory

Edited by Javad Sarvestan; peer-reviewed by Bethany Fordham, Lucky Ilodigwe; submitted 01.Apr.2026; final revised version received 21.Jul.2026; accepted 22.Jul.2026; published 07.Oct.2026

Please cite as:

Zhang X, Wilson CR, Eyre H, Reed II DE, Hee Wai TY, Kloehn A, Rosser EW, Luo G, Zeliadt SB

Extraction of Pain Severity and Functional Interference From Clinical Narratives Using Domain-Informed Large Language Models: Protocol for a Development and Validation Study

JMIR Res Protoc 2026;15:e96346

URL: <https://www.researchprotocols.org/2026/1/e96346>

doi: [10.2196/96346](https://doi.org/10.2196/96346)

© Xiaoyi Zhang, Christopher R Wilson, Hannah Eyre, David E Reed II, Travis Y Hee Wai, Alexander Kloehn, Ethan W Rosser, Gang Luo, Steven B Zeliadt. Originally published in JMIR Research Protocols (<https://www.researchprotocols.org>), 07.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.researchprotocols.org>, as well as this copyright and license information must be included.