PEACE: A Dataset of Pharmaceutical Care for Cancer Pain Analgesia Evaluation and Medication Decision (Dataset Documentation)

1 Data Sheet

1.1 Motivation

Q: For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
A: The PEACE dataset was created to advance cancer pain medication research by addressing existing gaps in available datasets. It aims to improve cancer pain management through comprehensive data collection, including long-term and multiple follow-ups, multidisciplinary treatment (MDT) team assessments, and patient self-perceptions of medication effects and impacts on their lives.

Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
A: The dataset was jointly developed by a collaborative effort from the research groups:

1. Central South University
2. Hunan University
3. The University of Sydney

Q: Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number
A: Funding provided by The Fundamental Research Funds for the Central South University.

Q: Any other comments?
A: None

1.2 Composition

1.2.1 Instances and Representation

Q: What do the instances that comprise the dataset represent?
A: The instances in the dataset represent patient information related to cancer pain management, including demographics, clinical signs, medication details, physiological parameters, pain assessment, treatment outcomes, and follow-up information.

Q: How many instances are there in total?
A: The dataset includes 103 features from more than 38,000 patients.

Q: Does the dataset contain all possible instances or is it a sample?
A: The dataset is a sample of instances selected based on inclusion criteria such as definitive

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

cancer diagnosis with pain and exclusion criteria like incomplete key medical records or significant complications.

Q: What data does each instance consist of?
A: Each instance consists of demographic data, clinical signs, medication details, physiological parameters, treatment outcomes, pain assessment, and follow-up data.

Q: Is there a label or target associated with each instance?
A: Yes, each instance has multiple labels or targets associated with it, including pain assessment scores, treatment decisions, and follow-up outcomes.

Q: Is any information missing from individual instances?
A: Some instances may have missing information due to unavailability at the time of data collection.

Q: Are relationships between individual instances made explicit?
A: Relationships are made explicit through linkages between patient records and treatment outcomes, pain assessments, and follow-up data.

Q: Are there recommended data splits?
A: We recommend that 80% of the dataset be used to build the model via 5-fold cross-validation, and the remaining 20% be used as an independent test set.

Q: Are there any errors, sources of noise, or redundancies in the dataset?
A: The PEACE dataset includes some errors, sources of noise, and redundancies.

- Errors and Noise Sources: Inconsistencies and anomalies due to human errors were addressed through expert consultation and the removal of data points to prevent model bias and improve robustness.

- Redundancies: Useful fields from duplicate records were merged, and some features were categorized to enhance usability in machine learning tasks. These measures ensure the consistency and reliability of the PEACE dataset, providing a high-quality foundation for cancer pain medication therapy research.

Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources?
A: The dataset is self-contained and does not rely on external resources.

Q: Does the dataset contain data that might be considered confidential?
A: Yes, the dataset contains sensitive health information that has been de-identified to comply with privacy regulations.

Q: Does the dataset contain data that might be offensive, insulting, threatening, or cause anxiety?
A: There is no indication that the dataset contains such data. It focuses on clinical and health-related information.

Q: Does the dataset identify any subpopulations?
A: No

Q: Is it possible to identify individuals from the dataset?
A: No, it is not possible to identify individuals directly or indirectly as all protected health information is de-identified.

Q: Does the dataset contain data that might be considered sensitive in any way?
A: Yes, it includes health data which is considered sensitive. Measures are taken to ensure this data is handled according to privacy regulations.

1.3 Collection Process

1.3.1 Data Acquisition

Q: How was the data associated with each instance acquired?
A: The data was acquired through a combination of direct observation and self-reports from patients. Data sources include manually collected hospital records and an online follow-up platform. Data validation involved cross-referencing with hospital records and expert reviews.

Q: What mechanisms or procedures were used to collect the data?
A: Data collection mechanisms included manual curation by clinical staff and automated data entry from the online follow-up platform. Validation procedures involved expert reviews and consistency checks across multiple data points.

Q: If the dataset is a sample from a larger set, what was the sampling strategy?
A: The dataset combines clinical features from Xiangya Hospital with data from our online cancer pain follow-up platform. It encompasses a wide range of patient information, including demographics, clinical signs, medications, physiological parameters, treatment outcomes, and others. This data reflects the complete daily records of doctors and clinical pharmacists, and the data were collected and collated manually without specific sampling.

Q: Who was involved in the data collection process?
A: Data collection involved clinical staff, including doctors, nurses, and pharmacists. No additional crowdworkers or contractors were involved.

Q: Over what timeframe was the data collected?
A: Data collection aligned with hospital record-keeping and follow-up timelines, spanning multiple years from 2016 onwards. This timeframe matches the creation dates of the instance data.

Q: Were any ethical review processes conducted?
A: Yes, ethical review was conducted by the Institutional Review Board of Xiangya Hospital, with Ethics Approval ID: 202109422. Informed consent was obtained from patients, and data was de-identified to protect privacy.

1.3.2 Data Source and Consent

Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?
A: Data was collected directly from the individuals during hospital visits and through the online follow-up platform.

Q: Were the individuals in question notified about the data collection?
A: Yes, individuals were notified through informed consent forms detailing the data collection process, its purpose, and the measures taken to protect their privacy.

Q: Did the individuals in question consent to the collection and use of their data?
A: Yes, individuals provided informed consent for data collection and use. The consent process included detailed information on how the data would be used and measures to ensure confidentiality.

Q: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
A: According to the use agreement, either party may terminate the agreement at any time. However, obligations regarding restricted data from PEACE will persist even after termination.

1.3.3 Impact Analysis

Q: Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?
A: The data was de-identified and ethical approvals were obtained to minimize any potential negative impacts.

1.4 Preprocessing and Cleaning

Q: Was any preprocessing/cleaning/labeling of the data done?
A: Yes, extensive preprocessing, cleaning, and labeling of the data were performed. The raw medication data presented challenges such as noise, complex attribute relationships, and high dimensionality. Issues like disorganization, duplicate records, and missing information were addressed through a comprehensive preprocessing pipeline. This included standardizing synonym variations within pain intensity labels (e.g., "burning pain," "scalding pain," and "burn-like pain" were standardized to "burning-type pain") and merging useful fields from duplicate records to enhance data quality.

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?
A: The raw data is retained in the internal database of Xiangya Hospital.

Q: Is the software that was used to preprocess/clean/label the data available?
A: No.

Q: Any other comments?
A: None

2 Dataset Nutrition Labels

2.1 Metadata Module

- **Filename**: PEACE
- **File format**: CSV
- **URL**: https://github.com/YTYTYD/PEACE
- **Domain**: Pharmaceutical care for cancer pain management
- **Keywords**: Cancer pain, pain management, pharmaceutical care, medication decision, analgesia evaluation
- **Type**: Patient health records
- **Dataset size**: Over 38,000 patients with 103 features each
- **Percentage of missing cells**: Some instances may have missing information due to unavailability at the time of data collection
- **License**: CC-BY
- **Collection range**: From 2016 To 2023
- **Description**: The PEACE dataset was created to advance cancer pain medication research by addressing existing gaps in available datasets. It aims to improve cancer pain management through comprehensive data collection, including long-term and multiple follow-ups, multidisciplinary treatment team assessments, and patient self-perceptions of medication effects and impacts on their lives.

2.2 Provenance Module

- **Source**: From Xiangya Hospital and data from a cancer pain online follow-up platform
- **Version history**: Version 1.0

2.3 Variables Module

Patients in the PEACE dataset have the following variables (for data type, B: Binary, N: Numeric, M: Multiclass, *: Label):

Patient Basic Information(50)

1. **Demographics**

- **ID (N)**: A unique random identification number assigned to each patient.
- **Gender (B)**: The gender of the patient.
- **Age (N)**: The age of the patient.
- **Height (N)**: The height of the patient.
- **Weight (N)**: The weight of the patient.
- **BMI (N)**: A common indicator for assessing body fat, calculated using weight and height.
- **Body Surface Area (BSA) (N)**: The total surface area of the human body.
- **Medical Record Date (N)**: The date on which the doctor makes a decision regarding cancer pain medication treatment based on a comprehensive pain assessment.
- **Length of Hospital Stay (N)**: The duration of the patient's stay during the current hospital visit, measured in days.
- **Number of Hospital Admissions (N)**: The total number of times the patient has been hospitalized due to tumour diseases.
- **Diagnosis (M)**: The diagnosis provided by the doctor at the time of discharge, only including tumour-related diseases.
- **Smoking History (B)**: Whether the patient has a history of smoking continuously for 6 months or more.
- **Drinking History (B)**: Whether the patient has a history of drinking alcohol at least once a week for 6 months or more.
- **Allergy History (B)**: Whether the patient has experienced allergic reactions.
- **Tumour Treatment Methods (M)**: The methods of tumour treatment, including surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy.
- **Gastrointestinal Risk (B)**: The likelihood of the patient developing gastrointestinal diseases (such as gastric ulcers, gastritis, enteritis) or related adverse reactions (such as gastrointestinal bleeding, indigestion) after taking pain medication.
- **Cardiovascular Risk (B)**: The likelihood of the patient developing cardiovascular diseases (such as hypertension, coronary heart disease, myocardial infarction) or related adverse reactions (such as arrhythmia, heart failure) after taking pain medication.
- **PS Score (N)**: The performance status score.

2. **Laboratory Examination Variables**

   (a) **Complete Blood Count:**

   - **White Blood Cell Count (N)**: The number of white blood cells in a unit volume of blood.
   - **Red Blood Cell Count (N)**: The number of red blood cells in a unit volume of blood.
   - **Hemoglobin (N)**: The amount of hemoglobin in a unit volume of blood.
   - **Platelet Count (N)**: The number of platelets in a unit volume of blood.
   - **Hematocrit (N)**: The volume percentage of red blood cells in blood.
   - **Neutrophil Count (N)**: The number of neutrophils in a unit volume of blood.
   - **Lymphocyte Count (N)**: The number of lymphocytes in a unit volume of blood.
   - **Eosinophil Count (N)**: The number of eosinophils in a unit volume of blood.
   - **Basophil Count (N)**: The number of basophils in a unit volume of blood.
   - **Monocyte Percentage (N)**: The proportion of monocytes in the total white blood cell count.
   - **Neutrophil Percentage (N)**: The proportion of neutrophils in the total white blood cell count.
   - **Lymphocyte Percentage (N)**: The proportion of lymphocytes in the total white blood cell count.

  - **Basophil Percentage (N)**: The proportion of basophils in the total white blood cell count.
  - **Eosinophil Percentage (N)**: The proportion of eosinophils in the total white blood cell count.
  - **Mean Corpuscular Volume (N)**: The average volume of a single red blood cell.
  - **Mean Corpuscular Hemoglobin (N)**: The average amount of hemoglobin in a single red blood cell.
  - **Mean Corpuscular Hemoglobin Concentration (N)**: The average concentration of hemoglobin in a single red blood cell.
  - **Red Cell Distribution Width (N)**: The variation in the size of red blood cells.
  - **Plateletcrit (N)**: The volume percentage of platelets in blood.
  - **Mean Platelet Volume (N)**: The average volume of a single platelet.

(b) **Liver Function:**

  - **Total Protein (N)**: The total amount of proteins in a unit volume of blood.
  - **Albumin (N)**: The amount of albumin in a unit volume of blood.
  - **Globulin (N)**: The amount of globulin in a unit volume of blood.
  - **Albumin/Globulin Ratio (N)**: The ratio of albumin to globulin in blood.
  - **Total Bilirubin (N)**: The total amount of bilirubin in a unit volume of blood.
  - **Direct Bilirubin (N)**: The amount of direct (conjugated) bilirubin in a unit volume of blood.
  - **Total Bile Acids (N)**: The total amount of bile acids in a unit volume of blood.
  - **Alanine Aminotransferase (N)**: The amount of alanine aminotransferase (ALT) in a unit volume of blood.
  - **Aspartate Aminotransferase (N)**: The amount of aspartate aminotransferase (AST) in a unit volume of blood.

(c) **Kidney Function:**

  - **Urea (N)**: The amount of urea in a unit volume of blood, reflecting kidney excretory function.
  - **Creatinine (N)**: The amount of creatinine in a unit volume of blood, reflecting kidney filtration function.
  - **Uric Acid (N)**: The amount of uric acid in a unit volume of blood, reflecting kidney excretory function and purine metabolism status.

Comprehensive Pain Assessment (15):

- **Pain Type (M)**: Classification of pain based on the pathological mechanism.
- **Worst Pain (N)**: The highest level of pain experienced in the last 24 hours, assessed using the Numerical Rating Scale (NRS).
- **Mildest Pain (N)**: The lowest level of pain experienced in the last 24 hours, assessed using NRS.
- **Average Pain (N)**: The average level of pain experienced in the last 24 hours, assessed using NRS.
- **Current Pain (N)**: The current level of pain, assessed using NRS.
- **Impact of Pain on Daily Life (N)**: The degree to which daily life was affected by pain in the past week.
- **Impact of Pain on Mood (N)**: The degree to which mood was affected by pain in the past week.
- **Impact of Pain on Walking Ability (N)**: The degree to which walking ability was affected by pain in the past week.
- **Impact of Pain on Daily Work (N)**: The degree to which daily work was affected by pain in the past week.

- **Impact of Pain on Relationships with Others (N)**: The degree to which relationships with others were affected by pain in the past week.
- **Impact of Pain on Sleep (N)**: The degree to which sleep was affected by pain in the past week.
- **Impact of Pain on Interest in Life (N)**: The degree to which interest in life was affected by pain in the past week.
- **Pain Frequency (M)**: The number of times pain occurred in a day for cancer pain patients.
- **Type of Breakthrough Pain (M)**: Classification of breakthrough pain according to the National Comprehensive Cancer Network (NCCN).
- **Frequency of Breakthrough Pain (M)**: The number of times breakthrough pain occurred in a day for cancer pain patients.

Previous Analgesic Treatment(23):

- **Prev_Extended Release Strong Opiates (ERSO) (N)**: The number of types of extended-release strong opiates used by the patient in the past week.
- **Prev_Immediate Release Strong Opiates (IRSO) (N)**: The number of types of immediate-release strong opiates used by the patient in the past week.
- **Prev_Extended Release Weak Opiates (ERWO) (N)**: The number of types of extended-release weak opiates used by the patient in the past week.
- **Prev_Immediate Release Weak Opiates (IRWO) (N)**: The number of types of immediate-release weak opiates used by the patient in the past week.
- **Prev_Nonsteroidal Anti-inflammatory Drugs (NSAID) (N)**: The number of types of nonsteroidal anti-inflammatory drugs used by the patient in the past week.
- **Prev_Anticonvulsants/Antidepressants (A/A) (N)**: The number of types of anticonvulsants/antidepressants used by the patient in the past week.
- **Prev_Others (N)**: The number of other analgesics used by the patient in the past week, excluding ERSO, IRSO, ERWO, IRWO, NSAIDs, and A/A.
- **Opiate Tolerance (B)**: Whether the patient has developed a decreased effect or reduced duration of action when using opiates for pain treatment.
- **Days of Medication Use (N)**: The number of days the patient used opiates (calculated based on the highest level of opiates used if multiple types were used simultaneously).
- The following 9 items are from the Morisky Medication Adherence Scale (MMAS-8), including 8 questions and a total score:
  - **M1 (N)**: Do you sometimes forget to take your medications?
  - **M2 (N)**: People sometimes miss taking their medications for reasons other than forgetting. Thinking over the past two weeks, were there any days when you did not take your medications?
  - **M3 (N)**: Have you ever cut back or stopped taking your medications without telling your doctor because you felt worse when you took them?
  - **M4 (N)**: When you travel or leave home, do you sometimes forget to bring along your medications?
  - **M5 (N)**: Did you take all your medications yesterday?
  - **M6 (N)**: When you feel like your symptoms are under control, do you sometimes stop taking your medications?
  - **M7 (N)**: Taking medication every day is a real inconvenience for some people. Do you ever feel hassled about sticking to your treatment plan?
  - **M8 (N)**: Do you have difficulty remembering to take all your medications?

  - **MMAS-8 Total Score (N)**: The total score ranges from M1 to M8, with higher scores indicating better adherence to medication.

- **Duration of Analgesic Control (N)**: The duration of pain control after taking analgesics.
- **Constipation (B)**: Whether the patient experienced constipation as an adverse reaction after taking analgesics.
- **Nausea/Vomiting (B)**: Whether the patient experienced nausea or vomiting as an adverse reaction after taking analgesics.
- **Other Adverse Reactions (B)**: Whether the patient experienced other adverse reactions besides constipation and nausea/vomiting after taking analgesics.
- **Medication for Adverse Reactions (B)**: Whether the patient used medications to manage adverse reactions.

Evaluation of Previous Analgesic Treatment(5):

1. The following 5 features are classified according to the Pharmaceutical Care Network Europe (PCNE) V8.0 classification of drug-related problems (DRPs):
    - **Drug-Related Problems (DRPs) (M)**: Any undesirable outcome or potential issue arising during the patient's drug therapy. This includes aspects of treatment effectiveness and safety.
    - **Causes (M)**: The underlying causes or factors leading to drug therapy problems.
    - **Interventions (M)**: Specific actions or measures taken to address drug therapy problems. These interventions can be implemented by pharmacists, doctors, or other healthcare professionals.
    - **Acceptance of Interventions (M)**: The patient's acceptance of the intervention plans proposed by healthcare professionals.
    - **Status of DRPs (M)**: The resolution status of DRPs after healthcare professionals' intervention.

Cancer Pain Medication Decision(9):

- **ERSO_Recommended (N*)**: The number of extended-release strong opiates recommended by the doctor.
- **IRSO_Recommended (N*)**: The number of immediate-release strong opiates recommended by the doctor.
- **ERWO_Recommended (N*)**: The number of extended-release weak opiates recommended by the doctor.
- **IRWO_Recommended (N*)**: The number of immediate-release weak opiates recommended by the doctor.
- **NSAIDs_Recommended (N*)**: The number of nonsteroidal anti-inflammatory drugs recommended by the doctor.
- **A/A_Recommended (N*)**: The number of anticonvulsants/antidepressants recommended by the doctor.
- **Others_Recommended (N*)**: The number of other analgesics recommended by the doctor, excluding ERSO, IRSO, ERWO, IRWO, NSAIDs, and A/A.
- **Constipation Medication Recommended (M)**: The types of medication recommended by the doctor for managing constipation.
- **Nausea/Vomiting Medication Recommended (M)**: The types of medication recommended by the doctor for managing nausea and vomiting.

Follow-up(1):

- **Pain Relief Status (M*)**: The degree of pain relief experienced by the patient after using the analgesic regimen recommended by the doctor.

2.4 Label Statistics

Table 1: MR task: numerical labels

Column	Max	Min	Mean	Missing Rate (%)
ERSO_Recom	3	0	0.715502	0.0
IRSO_Recom	2	0	0.232956	0.0
ERWO_Recom	1	0	0.031155	0.0
IRWO_Recom	1	0	0.095974	0.0
NSAIDs_Recom	3	0	0.293342	0.0
A/A_Recom	2	0	0.043307	0.0
Others_Recom	2	0	0.001079	0.0

Table 2: TEA task: multi-classification labels

Column	Count	Missing rate	Label 1 count	Label 2 count	Label 3 count	Label 4 count
Pain Relief Status	30,932	27.43%	14,470	11,523	3,327	1,612

3 Data Statement for PEACE

Curation Rationale

Description: The PEACE dataset was curated to facilitate research in pharmaceutical care for cancer pain management, with a specific focus on pain analgesia evaluation and medication decision-making. The dataset includes detailed patient information related to cancer pain management, encompassing demographics, clinical signs, medication details, physiological parameters, pain assessments, treatment outcomes, and follow-up information.

Language Variety

Description: The dataset includes medical records and patient reports primarily in English. The original data was collected in Mandarin and has been professionally translated into English to facilitate broader accessibility and usability in research and clinical settings.

Demographic

Description: The dataset contains demographic information of the patients, including their age and gender. The patients are primarily adults over the age of 18, with a balanced representation of genders.

Annotator Demographic

Description: The annotation was performed by a team of medical professionals and data scientists, including both men and women from diverse ethnic backgrounds. The team comprised individuals aged 25 to 50, with representation from multiple ethnicities, ensuring a broad perspective in the annotation process. The annotators received extensive training in medical terminology and the specific requirements of the PEACE dataset, including guidelines on consistency and accuracy in annotation.

Text Characteristics

Description: The dataset includes patient clinical information and follow-up records focusing on cancer pain management. The text covers a variety of topics, including patient demographics, clinical information, medication details, physiological parameters, pain assessment, treatment outcomes, and follow-up information. This comprehensive data collection allows for a thorough analysis of various aspects related to cancer pain and its management.

Speech Situation

Description: Not applicable.

Recording Quality (if applicable)
Description: Not applicable.

4 Data Card

4.1 Dataset Overview

Data Subject(s): Cancer patients experiencing pain.

Dataset Snapshot: The PEACE dataset includes detailed pharmacological care records for over 38,000 patients, covering demographics, clinical examination, treatment outcomes, medication plans, and patient self-perceptions.

Content Description: Records long-term and multiple follow-ups both inside and outside hospitals, includes patients' self-assessments of medication effects and the impact on their lives.

Descriptive Statistics: The dataset contains 103 features related to diverse pathologies, symptoms, and etiologies, with multi-visit, long-term observations for 2,600 patients.

Dataset snapshot : The snapshot of the dataset is shown in Table 3.

Table 3: Dataset snapshot

Category	**Data**
Size of Dataset	10.9 MB
Number of Instances	38,766
Number of Features	103
Number of Fields	7
Labeled Classes (Classification)	4

4.2 Sensitivity of Data

Sensitivity Type(s): Medical information

Field(s) with Sensitive Data: Patient demographics, medical history, medication details, and treatment outcomes.

Security and Privacy Handling: Data anonymization and privacy protection protocols are in place. Patient identifiers are removed, and dates are shifted to ensure privacy.

Risk Type(s): Data breaches, misuse of sensitive information.

Supplemental Link(s): PEACE dataset

Risk(s) and Mitigation(s): De-identification of patient data, secure data storage, and controlled access.

4.3 Dataset Version and Maintenance

Version Details: Version 1.0

Maintenance Plan: Updates based on new data and feedback.

Expected Change(s): Addition of new patient records and features.

4.4 Example of Data Points

Primary Data Modality: Structured tabular data.

Sampling of Data Points: Data points include demographics, clinical signs, medication details, physiological parameters, pain assessment, treatment outcomes, and follow-up information.

Data Fields: 103 features categorized into six groups.

4.5 Motivations & Intentions

Motivations: To improve cancer pain management by providing comprehensive data for TEA and MR systems.

Purpose(s): Research in pharmaceutical care, machine learning model training for treatment assessment, and medication recommendation.

Domain(s) of Application: Cancer pain management, pharmacotherapy, clinical research.

Motivating Factor(s): Enhancing treatment effectiveness and patient quality of life.

Intended Use: Development of machine learning models for pain assessment and medication recommendations.

Dataset Use(s): Research, model training, clinical decision support.

Suitable Use Case(s): Studies on treatment efficacy, medication optimization, personalized medicine.

Unsuitable Use Case(s): Applications not related to cancer pain or pharmacological research.

Research and Problem Space(s): Pharmacotherapy for cancer pain, machine learning in healthcare.

Citation Guidelines: Please cite the PEACE dataset as follows: Dataset available at `https://github.com/YTYTYD/PEACE`.

4.6 Provenance

Collection: **Method(s) Used**: Clinical data collection, patient follow-ups.

Methodology Detail(s): Data anonymized, dates shifted, Delphi consensus method for feature selection.

Source Description(s): Hospital records, patient self-reports.

Collection Cadence: Continuous.

Data Integration: Combined hospital and follow-up data.

Data Processing: Standardization, imputation, and categorization of features.

Collection Criteria: **Data Selection**: Patients with definitive cancer diagnoses and associated pain.

Data Inclusion: Comprehensive medical records and follow-up reports.

Data Exclusion: Incomplete records or patients under 18 years.

Relationship to Source: **Use & Utility(ies)**: Enables detailed analysis of cancer pain management and medication efficacy.

Benefit and Value(s): Supports research in personalized medicine and treatment optimization.

Limitation(s) and Trade-Off(s): Limited to cancer pain, potential biases from a single regional source.

4.7 Annotations & Labeling

Annotation Workforce Type: Medical professionals.

Annotation Characteristic(s): Detailed annotations by clinical experts.

Annotation Description(s): Includes medical diagnoses, treatment outcomes, and patient-reported symptoms.

Annotation Distribution(s): Across all patient records.

Annotation Task(s): Annotation of clinical features and patient follow-up reports.

Human Annotators: **Annotator Description(s)**: Clinical pharmacists, anesthetists, oncologists, and nurses.

Annotator Task(s): Assessing treatment outcomes and medication plans.

Language(s): Mandarin (translated to English for documentation).

Location(s): Xiangya Hospital and affiliated follow-up platform.

Gender(s): Both male and female annotators.

4.8 Human and Other Sensitive Attributes

Sensitive Human Attribute(s): Health status, pain levels, medication details.

Intentionality: Required for accurate pain and treatment assessment.

Rationale: Critical for understanding treatment efficacy and patient well-being.

Source(s): Clinical and self-reported data.

Known Correlations: Pain levels and medication efficacy, patient demographics and treatment outcomes.

Risk(s) and Mitigation(s): Privacy risks mitigated through data anonymization and secure storage.

4.9 Validation Types

Method(s): Expert consensus and Delphi method.

Breakdown(s): Multiple rounds of expert surveys to refine features.

Description(s): Features validated through structured communication and agreement among experts.

Description of Human Validators: **Characteristic(s)**: Experienced clinicians and pharmacists.

Description(s): Experts in cancer pain management.

Language(s): Mandarin.

Gender(s): Both male and female validators.

4.10 Sampling Methods

Method(s) Used: Judgmental sampling and expert consensus.

Characteristic(s): Targeted selection of experts and comprehensive patient data.

Sampling Criteria: Inclusion of patients with complete medical records and definitive cancer diagnoses.

4.11 Known Applications & Benchmarks

ML Application(s): TEA (classification) and MR (regression) systems.

Evaluation Result(s): Validated the efficacy of 13 machine learning models.

Evaluation Process(es): Experiments with 5-fold cross-validation and independent test sets.

Description(s) and Statistic(s): Detailed performance metrics in the experiment section.

Expected Performance and Known Caveats: Tree-based models perform effectively, while neural network models demand specialized tuning.

4.12 Use in ML or AI Systems

Dataset Use(s): Training models for cancer pain treatment assessment and medication recommendation.

Notable Feature(s): Long-term follow-up, comprehensive patient assessments.

Usage Guideline(s): We release the PEACE dataset under a CC-BY license. The access process for the dataset involves three steps:

1. Complete some training and provide certification (such as the CITI or GCP certification).
2. Carefully read the terms of the Data Use Agreement and if you agree and wish to proceed, send your application to the manager. Please use an official email address (such as .edu).
3. Final approval of data access is required by Xiangya Hospital

Once an application has been approved, the researcher will receive emails containing instructions for downloading the dataset. Any model trained on this dataset should not be deployed in real-world systems until its performance has been rigorously evaluated and the system's scope and representativeness in relation to real-world applications have been validated. The use of data must strictly comply with relevant regulations in China. Access to the PEACE dataset can be found at the following address:[https://github.com/YTYTYD/PEACE].

Distribution(s): Controlled access via GitHub repository.

Known Correlation(s): Patient demographics and treatment outcomes.

Split Statistics: 80% training, 20% testing.

4.13 Terms of Art

Concepts and Definitions: TEA: Treatment Effectiveness Assessment.
MR: Medication Recommendation.
MDT: Multidisciplinary Treatment.
NRS: Numerical Rating Scale.

4.14 Reflections on Data

Any additional information not captured by the data card: The dataset aims to fill gaps in existing cancer pain datasets by including multidisciplinary assessments and long-term follow-ups.

4.15 Access Retention & Wipeout

Access: **Access Type**: Restricted, for research purposes.

Documentation Link(s): PEACE dataset.

Prerequisite(s): 1) Complete relevant training, 2) Agree to the data usage agreement, 3) Obtain approval from Xiangya Hospital.

Policy Link(s): Included in the data use agreement.

Access Control List(s): Managed by the dataset maintainers and Xiangya hospital.

Retention: **Duration**: Long-term.

Policy Summary: Data retained for ongoing research and updates.

Process Guide: Data use agreement outlines retention policies.

Exception(s) and Exemption(s): As specified in the data use agreement.

Wipeout and Deletion: **Duration**: As needed based on ethical guidelines.

Deletion Event Summary: Deletion upon request or end of research period.

Acceptable Means of Deletion: Secure deletion protocols.

Post-Deletion Obligations: Ensure no residual data remains.

Operational Requirement(s): Compliance with institutional guidelines.

Exceptions and Exemptions: None specified.

5 Accountability Frameworks

5.1 Dataset Requirements Specification

Name of Dataset:
PEACE: Pharmaceuticals for easing cancer pain with care

Owner:
Developed by a collaborative effort from multiple research groups.
Funding provided by The Fundamental Research Funds for the Central South University.

Vision:
Vision 1.0

Motivation:
The dataset was created to improve cancer pain management by filling gaps in existing data. It focuses on long-term follow-ups, MDT assessments, and patient self-reports on medication effects and impacts.

Intended uses:
Specific uses include advancing cancer pain medication research, improving cancer pain management, and facilitating comprehensive data collection for various related analyses.

Non-intended uses:
The dataset should not be used for purposes outside the scope of cancer pain management and related research without proper context and understanding of its limitations.

Related documents:

- All_Data.csv: a .CSV file containing all patients in the dataset, with patient ID.
- All_data.json: a .JSON file describing all the data in the dataset.
- D_ Numerical.csv: A .csv file containing the units of the numerical features.
- D_ Multiclass.csv: A .csv file containing the meaning of multiclass features.
- D_ Diagnosis.csv: A .csv file containing the meaning of diagnosis.
- Train data: a .CSV file containing the training set of patients.
- Test data: a .CSV file containing the test set of patients.

Stakeholders consulted:
Clinical staff, including doctors, nurses, and pharmacists, were involved in the data collection process.

Creation requirements:
Data was collected from hospital and an online follow-up platform.
Manual curation by clinical staff.
The data spans multiple years and includes comprehensive patient information.

Instance requirements:
The dataset includes demographic data, clinical signs, medication details, physiological parameters, treatment outcomes, pain assessments, and follow-up data.

There are 103 features from more than 38,000 patients.
Data is a sample selected based on inclusion and exclusion criteria.

Distributional requirements:
The dataset should represent a comprehensive range of patient information and ensure the inclusion of diverse patient demographics and conditions.

Data processing requirements:
Deep de-identification and privacy protection processing.
Extensive preprocessing, cleaning, and labeling were performed.
Issues like disorganization, duplicate records, and missing information were addressed.
Standardization and merging of useful fields from duplicate records were done to enhance data quality.

Performance requirements:
Medical related users can expect high-quality data suitable for advancing cancer pain medication research and improving cancer pain management practices. Users of machine learning can look forward to new data sources and baselines for recommended treatment effect assessment and medication recommendation systems.

Maintenance requirements:
Data should be regularly updated to maintain its relevance and accuracy. The frequency and duration of updates depend on ongoing research and clinical needs.

Sharing requirements:
We release the PEACE dataset under a CC-BY license. We ask that users proactively complete human subjects research training and adhere to a data use agreement that requires responsible data handling and compliance with collaborative research principles. Data access is required to be reviewed by Xiangya Hospital.

Caveats and risks:
The dataset includes de-identified sensitive health information, and measures were taken to ensure data privacy and security. Users should handle the data responsibly to avoid misuse. Any model trained on this dataset should not be deployed in real-world systems until its performance has been rigorously evaluated and the system's scope and representativeness in relation to real-world applications have been validated.

Data ethics:
Ethical review and approvals were conducted. Informed consent was obtained from patients, and data was de-identified to comply with privacy regulations.

5.2 Design Document

Name of Dataset:
PEACE: Pharmaceuticals for easing cancer pain with care

Owner:
Developed by a collaborative effort from multiple research groups.
Funding provided by The Fundamental Research Funds for the Central South University.

Primary Data Type(s):
Numerical data, Binary Data, Multiclass data

Data Content:
It includes basic patient information, comprehensive pain assessment, previous analgesic treatment and evaluation, cancer pain medication decision-making, monitoring and management of adverse reactions, and pain relief assessment.

Objective:
This research aims to improve cancer pain medication and management by addressing data gaps. The

dataset offers a comprehensive view, including patient perspectives and multidisciplinary treatment evaluations.

Background:

Cancer pain is a common symptom among cancer patients, with an incidence rate of up to 53%. This greatly affects patients' quality of life and may impede effective cancer treatment. Pharmacotherapy, the mainstay of cancer pain management, often involves long-term medication use. Physicians must continually assess the efficacy of the current analgesic regimen by considering factors such as the patient's physical condition, pain intensity, type of pain, and prior medications. This enables targeted adjustments to the treatment plan to improve therapeutic outcomes.

Sources:

The data is sourced from hospital records and an online follow-up platform. It includes a broad range of patient information such as demographics, clinical signs, medication details, physiological parameters, treatment outcomes, pain assessments, adverse drug reactions, and dynamic adjustments to medication.

Annotations:

Our data construction process resulted in a comprehensive dataset encompassing 103 features, broadly categorized into six groups. The Patient Baseline Information group (50 features) captures demographic and clinical characteristics of the patients, potentially including age, gender, co-morbidities, and disease stage. The Comprehensive Pain Assessment group (15 features) details the extent and characteristics of the patients' pain experience, potentially including pain intensity scores, pain quality descriptors (e.g., visceral pain, somatic pain), and functional limitations. The Previous Analgesic Treatment group (23 features) details the medications and interventions previously used to manage the patients' pain, potentially including medication names, dosages, durations, and routes of administration. The Evaluation of Previous Analgesic Treatment group (5 features) captures the effectiveness and tolerability of prior pain management strategies, potentially including patient-reported outcomes or physician assessments. The Cancer Pain Medication Decision group (9 features) details the rationale behind the selection of specific pain medications for the study participants, potentially including factors like pain type, treatment history, and co-morbidities. The Follow-Up group (1 feature) captures information on patient outcomes after the intervention of interest, potentially including pain response or adverse events. The labels of the dataset include 7 categories of medication recommendations (regression) and treatment effect evaluation (classification).

Data Quality:

Data quality is ensured through expert consultation, the removal of data points to prevent model bias, merging of useful fields from duplicate records, and categorizing features to enhance usability in machine learning tasks.

Characteristics:

Expected Characteristics: The dataset includes 103 features from over 38,000 patients. Relationships are made explicit through linkages between patient records and treatment outcomes, pain assessments, and follow-up data.

Population: The population represented includes patients with a definitive cancer diagnosis and pain.

Privacy Handling:

Privacy is handled by de-identifying all sensitive health information to comply with privacy regulations.

Maintenance:

The dataset is maintained internally at Xiangya Hospital. Issues are addressed as they arise.

Sharing:

The dataset will be shared for research purposes, with access controlled and data de-identified to ensure privacy.

Caveats:
Known caveats include some errors, sources of noise, and redundancies due to human errors. These have been addressed through expert consultation and preprocessing steps.

Data Ethics:
Ethical considerations include obtaining informed consent from patients, ethical review and approval by the Institutional Review Board of Xiangya Hospital, and de-identification of data to protect patient privacy.

Related Datasets:
To build reliable TEA and MR systems, it is crucial to gather comprehensive data on both inpatients and outpatients. This includes medication details, treatment outcomes, adverse events and their etiologies, treatment adjustments, and impact on patients' quality of life. However, no public dataset currently meets all these requirements comprehensively. Widely used datasets such as MIMIC-III [2] and MIMIC-IV [1], while detailed in recording medication specifics, lack pharmacist assessments of treatment outcomes. These datasets primarily focus on single hospitalization events rather than the long-term health status of patients, which is particularly disadvantageous for managing chronic conditions like cancer pain. Similarly, the eICU Collaborative Research Database [5] documents essential medication usage information but fails to provide clear explanations of medication effects and lacks long-term patient follow-up. Additionally, these datasets lack patient feedback on their treatment plans. SEER [6] is a representative large-scale cancer registry databases in the United States, compiling extensive retrospective clinical data. It primarily focuses on the treatment processes of cancer patients but does not include assessments of medication plans following hospital discharge. For medication effect assessment, the SIDER [3] database lists adverse reactions for marketed drugs, while the FAERS [8] and TwoSIDES [7] datasets record potential drug interactions. Although these datasets are useful in some aspects, they generally lack detailed records of patients' conditions and necessary clinical features, limiting their practical utility. ISS[4] is a cancer pain assessment dataset that includes videos of 29 patients, along with their self-reported pain scale scores, used to predict the patients' pain levels. A common shortfall of these datasets is their inability to continuously observe and assess patient conditions. They often describe data from a single perspective and fail to integrate the diverse characteristics needed for making MDT decisions. The following section details the PEACE dataset and the steps taken to construct it, aiming to address the deficiencies of existing datasets.

5.3 Dataset Testing Report

Name of Dataset: PEACE: Pharmaceuticals for easing cancer pain with care

5.3.1 Summary

What is being tested?
We developed the PEACE dataset, a comprehensive resource specifically designed for the construction of treatment effectiveness assessment (TEA) and medication recommendation (MR) systems for cancer pain. The testing focuses on the following aspects:

- TEA, which is a multi-label classification (levels 1-4) using patient characteristics with time series data to quantify levels of treatment efficacy.
- MR, which involves regression analyses utilizing time series data to predict the quantity of various analgesics required by patients following adjustments in their treatment plans based on their medication history.

5.3.2 Testing Metrics

We used the following metrics to evaluate the performance. For TEA (classification tasks), we used the metrics of accuracy (ACC), area under the receiver operating characteristic curve (AUROC), F1 score, recall, and precision. For MR (regression tasks), we used mean squared error (MSE) and mean absolute error (MAE).

5.3.3 Meta-Testing

Is the data still needed?

Yes, the data is essential for ongoing cancer pain management research, improving treatment outcomes, and developing new pharmaceutical care strategies.

Are the data requirements still relevant and up-to-date?

Yes, the data requirements remain relevant and are up-to-date, addressing current research needs.

5.3.4 Requirements Testing

Table 4: Requirements Testing

Requirement from requirements specification	Score or Results	Justification of the results or a link to artifact
Multi-label classification for TEA	Met	Time series data is used to classify treatment efficacy levels (1-4) based on patient characteristics
Regression analysis for MR	Met	Time series data is utilized to predict the quantity of analgesics required, considering medication history

Table 5: Untested Requirements

Untested Requirement	Reason for not testing
None	All requirements have been tested

References

[1] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[2] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[3] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

[4] C. Ordun. Intelligent sight and sound: A chronic cancer facial pain dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[5] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[6] Surveillance, Epidemiology, and End Results (SEER) Program. Seer*stat database: Incidence - seer 18 regs research data, nov 2021 sub (1975-2019) - linked to county attributes - time dependent (1990-2019) income/rurality, 1969-2019 counties, 2022. National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 submission. Available at: `https://seer.cancer.gov/`.

[7] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.

[8] U.S. Food and Drug Administration. Fda adverse event reporting system (faers) public dashboard. `https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html`, 2024. Accessed on 2024-05-09.