Introduction and Rationale

Older adults represent the majority of patients diagnosed with prostate cancer, with over 40% of new cases occurring in individuals aged 75 or older. Despite this demographic reality, elderly patients are historically underrepresented in the pivotal clinical trials that support systemic therapy approvals. Concerns about comorbidities, frailty, and competing risks often lead to the exclusion of older participants, resulting in a limited evidence base to guide treatment decisions in this population.

Regulatory agencies including the FDA and EMA have called for more inclusive trial designs and improved transparency around age-specific outcomes. However, the degree to which elderly-specific evidence informs systemic therapy approvals in prostate cancer remains largely unquantified and unstandardized. As a result, clinicians are frequently forced to extrapolate from younger, fitter trial populations when treating older adults — a practice that may introduce risk or uncertainty, especially for agents with narrow therapeutic windows or substantial toxicity.

To address this gap, **ELDER-PC-Bench** aims to create the first structured, reproducible benchmark evaluating the extent of adult-to-elderly extrapolation in systemic therapy for prostate cancer. By combining regulatory document analysis, large language model–assisted evidence extraction, and expert adjudication, this study will characterize the current landscape of geriatric trial inclusion and define the strength of age-specific evidence for each approved therapy. The resulting dataset and extrapolation scoring system will support more transparent clinical decision-making, inform regulatory and guideline development, and lay the groundwork for expanded inclusion of older adults in future trials.

ELDER-PC-Bench is a structured framework to evaluate the degree to which systemic therapies for prostate cancer have been clinically validated in elderly patients, particularly those aged ≥75. We will assemble a curated dataset of prostate cancer drug approvals and their supporting trials, apply a two-stage natural language processing pipeline to extract age-specific evidence from regulatory and clinical documents, and classify the degree of elderly extrapolation using both categorical labels and a continuous scoring system. All methods rely exclusively on publicly available data and do not involve human subjects.

**1. Scope and Definitions**

**Cancer Type and Indications**

We will focus exclusively on prostate cancer and include therapies used in the following indications:

* Metastatic hormone-sensitive prostate cancer (mHSPC)
* Metastatic castration-resistant prostate cancer (mCRPC)
* Non-metastatic castration-resistant prostate cancer (nmCRPC)
* Biochemical recurrence (BCR) requiring systemic therapy

**Definition of Elderly**

We define “elderly” as age ≥75 years, consistent with prior geriatric oncology literature and FDA guidance on older adult representation in clinical trials.

**2. Drug and Trial Identification**

**Therapy Selection**

We will identify all systemic therapies approved by the U.S. Food and Drug Administration (FDA) and/or European Medicines Agency (EMA) for prostate cancer between 2005 and 2025. This includes:

* Androgen receptor signaling inhibitors (e.g., abiraterone, enzalutamide)
* Chemotherapies (e.g., docetaxel, cabazitaxel)
* Radiopharmaceuticals (e.g., radium-223, lutetium-177-PSMA)
* Immunotherapies and targeted agents (e.g., sipuleucel-T, PARP inhibitors)

Therapy–indication pairs (e.g., abiraterone for mCRPC) will form the unit of analysis.

**Trial Identification**

For each therapy–indication pair, we will retrieve:

* The pivotal trial(s) cited in the FDA label

To standardize trial identification and indication mapping, we restricted our analysis to the **most recent FDA label associated with the original branded product** for each prostate cancer therapy. In cases where multiple labels existed for a given active ingredient—such as those from generic manufacturers or supplemental approvals—we selected the original **new drug application (NDA)** sponsor label (e.g., Zytiga for abiraterone acetate) as the canonical source. This approach is consistent with prior methodology used in PED-X-Bench and reflects the fact that pivotal trial data and detailed subgroup analyses are typically only reported in the initial branded application. Generic drug labels generally replicate the clinical content of the reference NDA and do not contribute additional data relevant to age subgroup extrapolation. All label documents were accessed via the [FDA Drugs@FDA database](https://www.accessdata.fda.gov/scripts/cder/daf/), and only the most current version of each branded label was included for analysis.

**3. Document Retrieval and Corpus Creation**

We will collect the following documents for each drug–indication pair:

* **FDA drug labels** (via Drugs@FDA)
* **EMA European Public Assessment Reports (EPARs)**
* **ClinicalTrials.gov registry entries** (including tabular results when available)
* **Peer-reviewed trial publications**

Documents will be stored as structured text using pdfplumber or equivalent tools.

**4. Metadata Extraction**

For each trial, we will extract structured metadata including:

* Indication (e.g., mCRPC)
* Trial name and phase
* Year of FDA approval
* Total sample size
* Number and proportion of patients aged ≥65, ≥70, ≥75, and ≥80
* Age-related inclusion/exclusion criteria
* Reporting of age-stratified outcomes (efficacy or safety)
* Use of geriatric-specific tools (e.g., G8, CARG)
* Whether dose modifications or PK differences by age are discussed

**5. Extrapolation Evidence Extraction (LLM Pipeline)**

**Stage 1: Sentence Extraction**

A large language model (e.g., GPT-4 via OpenAI API) will be prompted to extract all sentences from each document that mention:

* Older adults or specific age groups (e.g., ≥75)
* Age-related efficacy or safety data
* Pharmacokinetic or pharmacodynamic differences by age
* Dosing modifications based on age or comorbidity

**Stage 2: Labeling and Summarization**

Extracted sentences will be passed into a second LLM prompt that:

* Assigns one of four extrapolation categories:
  + **None**: Elderly-specific trial or analysis (no extrapolation)
  + **Partial**: Adult trial data with some age-stratified safety/PK/efficacy evidence
  + **Full**: Approval based solely on younger adult data with no elderly-specific validation
  + **Unlabeled**: No discussion of age-related considerations
* Generates a 100–150 word summary describing the age-related evidence base, including:
  + Number of elderly participants
  + Reported outcomes or tolerability in older adults
  + Regulatory notes on dosing, PK, or extrapolation rationale

**6. Elderly Extrapolation Scoring System**

We will also develop a 0–5 point **Elderly Extrapolation Score (EES)** to quantify evidence depth:

| **Criterion** | **Points** |
| --- | --- |
| ≥10% of participants aged ≥75 | 1 |
| Efficacy results stratified by age | 1 |
| Safety/adverse events stratified by age | 1 |
| Age-specific PK or dose adjustment data | 1 |
| Use of geriatric assessment tools | 1 |

ach drug–indication pair will receive an EES (0–5), enabling ordinal comparisons and model training.

**7. Expert Adjudication (Validation Set)**

To ensure classification reliability, a stratified random sample of 50 therapy–indication pairs will be reviewed by an expert panel including:

* A urologic oncologist
* A geriatric oncologist
* A regulatory pharmacologist

Each expert will independently assign a categorical label and EES. Inter-rater agreement (Cohen’s κ) will be calculated. Disagreements will be resolved by consensus.

**8. Baseline Model Development**

We will train two baseline models to predict extrapolation category and EES score:

**Model 1: Logistic Regression**

* Inputs: Extracted metadata (e.g., trial size, % ≥75, cancer stage)
* Output: 4-class extrapolation label
* Evaluation: Accuracy, macro-F1, ROC-AUC

**Model 2: Transformer-based Classifier (e.g., BigBird or Longformer)**

* Input: Full regulatory text + trial summary
* Output: Extrapolation label and EES
* Evaluation: F1 score, calibration curve, attention map interpretability

**9. Output and Dissemination**

All data, model weights, prompts, and annotations will be:

* Released as a public benchmark dataset on GitHub or Hugging Face
* Submitted for peer-reviewed publication
* Shared with relevant stakeholders (e.g., NCCN, FDA, ASCO) to inform geriatric inclusion policies