

Data Dictionary for Kenyan Medical Vignettes

Overview

This data dictionary describes the dataset structure for medical vignettes used in the Kenya LLM project. The dataset contains clinical scenarios presented to various AI language models (GPT-4.0, LLAMA, GEMINI) and human clinicians, along with their responses and additional metadata.

Fields

Field Name	Data Type	Description	Example 1	Example 2
Master_Index	Number	Unique identifier for each vignette	1	2
County	String	Kenyan county where the nurse is working	Kiambu	Kakamega
Health level	String	Type of healthcare facility	Sub-county Hospitals and Nursing Homes	Dispensaries and Private Clinics
Prompt	String	Combined nurse profile and clinical scenario, including the nurse's introduction and the clinical situation with a specific question	"I am a nurse with 10 years of experience in General nursing working in a Sub-county Hospitals and Nursing Homes in Kiambu county in Kenya. A 28yrs old woman came to maternity unit with complains of lower abdominal pain, 38/40 weeks VE os closed..with no fetal heart heard. As a midwife which is the best mode of delivery CS or SVD?"	"I am a nurse with 2 years of experience in General nursing working in a Dispensaries and Private Clinics in Kakamega county in Kenya. A three-month-old baby brought to the facility for immunization, which is due today. On exam, the baby was pale and febrile. The baby was crying a lot. Should I treat the baby first or just give the vaccine and treat later?"
Nursing Competency	String	Area of nursing competency relevant to the scenario	Maternal and Child Health	Child Health
Clinical Panel	String	Medical specialty category for the vignette	OBSTETRICS AND GYNAECOLOGY	PAEDIATRICS

Field Name	Data Type	Description	Example 1	Example 2
Clinician	String	Response from a human clinician to the scenario	"Summary: 28 year old pregnant woman 38/40 weeks with complaints of lower abdominal pain. Vaginal examination - closed os. No fetal heart heard. DIAGNOSIS: Non reassuring fetal status. Q1 - Which is the best mode of delivery: obstetric ultrasound to confirm fetal status. Examine the mother to rule out absolute indications of a Cesarean section such as cephalopelvic disproportion, 2 previous scars. If no indications of a CS, induce labour for spontaneous vertex delivery"	"Summary: A 3 month old baby was brought to the facility for immunization. The baby was crying a lot. On examination, the baby was pale and febrile. So the question is, should I treat the baby first or just give the vaccine and treat later? Treat the baby first, then give the vaccine later."
GPT4.0	String	Response generated by the GPT-4.0 language model	"Given the clinical scenario, where the patient is a 28-year-old woman at 38 weeks gestation presenting with lower abdominal pain, no fetal heart sounds heard, and a closed cervix (as indicated by the vaginal exam), the management approach requires careful consideration... Based on the information provided, cesarean section (CS) would likely be the safest mode of delivery in this situation..."	"In this case, the baby's condition (pale, febrile, and crying a lot) suggests that there may be an underlying illness that requires urgent attention before proceeding with the immunization... You should not give the vaccine immediately without addressing the baby's current health status. Treating the underlying cause of the fever and pallor should be the priority..."
LLAMA	String	Response generated by the LLAMA language model	"Based on the information you've provided, I would recommend a Caesarean Section (CS) as the best mode of delivery for this patient. The fact that the fetal heart rate is not audible is a concerning sign, and in the absence of any other information, it's prudent to assume that the fetus may be in distress..."	"As a professor of primary healthcare in Kenya, I'd be happy to guide you on this. According to the Kenya Expanded Programme on Immunization (KEPI) guidelines, immunization should be delayed in certain situations, including when the child is severely ill or has a high fever..."
GEMINI	String	Response generated by the GEMINI language model	"Given the scenario: a 28-year-old woman at 38-40 weeks gestation presenting with lower abdominal pain, a closed cervix, and absent fetal heart tones, a Cesarean Section (CS) is the best mode of delivery. The absence of fetal heart tones is a critical finding indicating fetal distress or demise..."	"You should absolutely NOT give the vaccine first. A pale, febrile, and crying three-month-old baby is showing signs of illness. Administering a vaccine while the baby is unwell could potentially worsen their condition and even be unsafe..."
DDX SNOMED	String	SNOMED CT diagnostic codes relevant to the scenario	"289909005 Labor, function (observable entity)\n871479001 Uncertain presentation of fetus (finding)\n199112006 Infections of the genital tract in pregnancy (disorder)"	"271749004 Acute rise of fever (finding)\n61462000 Malaria (disorder)\n91302008 Sepsis (disorder)\n434021000124109 At risk for anemia (finding)"

Field Details

Master_Index

- **Definition:** Unique numeric identifier for each vignette in the dataset
- **Format:** string
- **Usage:** Used to reference specific clinical scenarios
- **Constraints:** Must be unique within the dataset

County

- **Definition:** Kenyan county where the nurse in the scenario is working
- **Format:** Text string
- **Examples:** Kiambu, Kakamega
- **Usage:** Provides geographic context for the clinical scenario

Health level

- **Definition:** Type of healthcare facility where the scenario takes place
- **Format:** Text string
- **Examples:**
 - "Sub-county Hospitals and Nursing Homes"
 - "Dispensaries and Private Clinics"
 - "Health Centres"
- **Usage:** Indicates the level of care and resources available

Prompt

- **Definition:** Combined nurse profile and clinical scenario, including the nurse's introduction and patient case
- **Format:** Text string beginning with a standardized introduction ("I am a nurse with [X] years of experience...") followed by the clinical scenario and a specific question
- **Contents:** Nurse background, patient demographics, presenting symptoms, relevant clinical findings, and a specific clinical question
- **Usage:** The complete prompt that models and clinicians respond to

Nursing Competency

- **Definition:** Area of the Kenya nursing practice relevant to the scenario
- **Format:** Text string
- **Examples:**
 - "Maternal and Child Health"
 - "Child Health"
 - "Adult Health"
- **Usage:** Categorizes the scenario by nursing specialty area

Clinical Panel

- **Definition:** Medical specialty category for the expert panel evaluating the Clinician and LLM responses
- **Format:** Text string
- **Examples:**
 - "OBSTETRICS AND GYNAECOLOGY"
 - "PAEDIATRICS"
 - "CRITICAL CARE"
- **Usage:** Categorizes the scenario by medical specialty for evaluation purposes

Clinician

- **Definition:** Response from a human healthcare professional to the clinical scenario
- **Format:** Text string, often starting with summary followed by assessment and recommendations
- **Contents:** Clinical reasoning, differential diagnosis, management plan, and specific answers to the scenario questions
- **Usage:** Provides the human expert response baseline for comparison with AI models

GPT4.0, LLAMA, GEMINI

- **Definition:** Responses generated by the GPT-4.0, LLAMA, and GEMINI large language models
- **Format:** Detailed text response
- **Contents:** Analysis of the scenario, differential diagnosis, management recommendations, and answers to the clinical questions
- **Usage:** Used to evaluate GPT-4.0's clinical reasoning and recommendations

DDX SNOMED

- **Definition:** SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) codes for possible diagnoses
- **Format:** Multiple SNOMED codes with descriptions, separated by newlines
- **Example:** "289909005 | Labor, function (observable entity)"
- **Usage:** Provides standardized medical terminology for the diagnostic considerations

Relationships Between Fields

- **Prompt** combines the nurse's introduction (containing **County** and **Health level** information) with the clinical scenario
- **Prompt** contains the clinical scenario that all response fields (**Clinician**, **GPT4.0**, **LLAMA**, **GEMINI**) address
- **DDX SNOMED** provides standardized diagnostic codes that should correspond to conditions mentioned in the responses
- **Clinical Panel** and **Nursing Competency** categorize the scenario by medical specialty and nursing domain

Notes on Usage

1. The **Prompt** field contains the complete information that should be sufficient for clinical decision-making.
2. **DDX SNOMED** codes can be used to verify if models identified the key diagnostic considerations.
3. **County** and **Health level** provide important context about resource availability that may impact appropriate management decisions.