# Chapter 2: Transforming Data for Reliable Evaluation with bat29

Evaluating Large Language Models in clinical diagnosis demands clean, consistent data. Raw clinical records, however, are typically messy, inconsistent, and unstructured, making direct use unreliable for rigorous experimentation. As highlighted in Table 5, variations in terminology, format, coding, and representation can obscure genuine model performance. The src/bat29 module serves as the essential pre-processing pipeline within Kernel29, transforming these diverse raw inputs into the standardized, structured datasets required for trustworthy evaluation.

## 1. Consolidating and Structuring Clinical Narratives

A primary challenge is the heterogeneity of input formats. Hospital records often contain clinical details scattered across numerous free-text fields, while research datasets might simply list coded phenotypes without a narrative structure.

**Input Example 1 (Hospital Record - URG\_Torre\_Dic\_2022\_IA\_GEN.csv excerpt):** | Motivo Consulta | Enfermedad Actual | Exploracion | Juicio Diagnóstico | | :————– | :——————————————————————— | :————————————————————————– | :————————————— | | Disnea.- | Paciente que es remitido en ambulancia… episodio de disnea… | Vigil, reactivo, aceptable estado general… Taquipneico en reposo… | Broncospasmo.\nProbable tapòn mucoso. | | Otalgia.- | Segùn refiere el paciente presenta desde hace 15 dìas… otalgia… | Lùcido, reactivo, buen estado general… Otoscopia derecha: CAE hiperémico… | Otitis externa derecha.\nOtitis media… | *(Note: The full URG dataset typically includes columns like Sexo, EDAD, TA Max, TA Min, Frec. Cardiaca, Temperatura, Sat. Oxigeno, Exploracion Compl., etc.)*

**Input Example 2 (Research Data - Conceptual JSONL for Rare Diseases):**

{ "Phenotype": ["HP:0001276", "HP:0001258", "HP:0000407"], "RareDisease": ["OMIM:176270"] } // HPO IDs for Hypotonia, Spasticity, Sensorineural hearing impairment; OMIM ID for Prader-Willi

Feeding such disparate data directly to an LLM or evaluation metric is problematic. bat29 addresses this by: a) Reading various source formats (CSV, JSONL, etc.). b) Extracting relevant clinical information (symptoms from free text or coded lists, history, exam findings, demographics, vitals, test results). c) Synthesizing or restructuring this information into a standardized narrative format using a predefined template.

For hospital records like the URG data, it extracts demographics (Sexo, EDAD) and clinical details (Motivo Consulta, Enfermedad Actual, Exploracion, Antecedentes) and concatenates them into structured Anamnesis, Exploracion, etc., sections. Crucially, it also parses and formats available vital signs and complementary tests (Exploracion Compl.) into a standardized Pruebas clinicas section.

For coded phenotype lists (like the source for test\_PUMCH\_ADM.csv), it looks up the names corresponding to the HPO codes and integrates them into the template, typically listing the symptoms under Anamnesis:. To prevent potential biases from inconsistent original consultation reasons in source datasets, the Motivo de consulta for these synthesized cases is **intentionally standardized** to the neutral phrase: Paciente acude a consulta para ser diagnosticado. Placeholders are used for demographics if unavailable.

**Simulated Output (case column derived from Input Example 1, *with invented plausible data*):**

*Row 1 (Disnea case):*

Motivo de consulta:  
Paciente acude a consulta para ser diagnosticado  
  
Anamnesis:  
  
Paciente Hombre de 54 años. Paciente que es remitido en ambulancia... episodio de disnea...  
  
Antecedentes:  
  
No hay antecedentes  
  
Exploracion:  
  
Vigil, reactivo, aceptable estado general... Taquipneico en reposo...  
  
Pruebas clinicas:  
  
-Rapidas:  
  
Tensión arterial: 144.0 / 75.0  
  
Frecuencia cardiaca: 73.0  
  
Temperatura: 36.6  
  
Saturación de oxígeno: 95.0  
  
-Complementarias:  
Rx torax Portatil: ICT <0.5; no consolidaciones ni derrame pleural. No neumotorax.-

**Output Example (case column in test\_PUMCH\_ADM.csv**

Motivo de consulta:  
Paciente acude a consulta para ser diagnosticado  
  
Anamnesis:  
  
Paciente de sexo desconocido de desconocidos años. El paciente presenta los siguientes síntomas:  
 -Compulsive behaviors  
 -Delayed speech and language development  
 -Neonatal hypotonia  
(... plus other phenotype names corresponding to HPO codes ...)  
  
Antecedentes:  
  
No hay antecedentes  
  
Exploracion:  
  
No se realiza  
  
Pruebas clinicas:  
nan

This dual transformation approach results in a consistent case column across all processed datasets (like those in /data/tests/treatment/), regardless of the original source format (scattered text fields or coded lists), providing a uniform input format suitable for LLMs.

## 2. Normalizing and Standardizing Diagnoses

Raw data presents diagnoses inconsistently – as free text, non-standard codes, or ontology identifiers.

**Input Example (Hospital Record - URG\_Torre\_Dic\_2022\_IA\_GEN.csv excerpt):** | DIAG CIE | Juicio Diagnóstico | | :——- | :————————– | | I50.9 | ICC DESCOMPENSADA\nEPOC | | J18.9 | NEUMONIA BILATERAL | | N23 | CRU derecho. |

**Input Example (Research Data - Conceptual JSONL for Rare Diseases):**

{ ... "RareDisease": ["OMIM:277900"] } // OMIM ID for Wilson Disease  
{ ... "RareDisease": ["ORPHA:284979", "ORPHA:558"] } // ORPHA IDs for Neonatal Marfan / Marfan

bat29 leverages curated knowledge bases (built from ICD-10, OMIM, HPO, etc.) to standardize these inputs: a) It maps provided codes (like DIAG CIE) or text diagnoses to canonical codes (primarily ICD-10 for URG, or retaining OMIM/ORPHA). b) It retrieves the official name for the canonical code. c) For research data, it resolves multiple IDs to a primary concept and fetches validated synonyms. d) It often combines the standard name with the original text or synonyms for completeness.

**Output Example (Diagnosis columns in test\_death.csv):** | id | golden\_diagnosis | diagnostic\_code/s | icd10\_diagnosis\_name | | :— | :————————————————————- | :—————- | :——————————- | | 6230 | “Heart failure, unspecified also known as ICC DESCOMPENSADA\nEPOC” | I50.9 | “Heart failure, unspecified” | | 6231 | “Pneumonia, unspecified organism also known as NEUMONIA BILATERAL” | J18.9 | “Pneumonia, unspecified organism” |

**Output Example (Diagnosis columns in test\_PUMCH\_ADM.csv):** | id | case | golden\_diagnosis | diagnostic\_code/s | | :- | :— | :————————————————- | :——————————- | | 0 | … | Prader-Willi syndrome | [OMIM:176270] | | 5 | … | “Wilson disease also known as Hepatolenticular…” | [OMIM:277900] | | 15 | … | “Neonatal Marfan syndrome also known as Marfan…” | “[ORPHA:284979, ORPHA:558]” | This yields consistently formatted golden\_diagnosis and diagnostic\_code/s columns, crucial for unambiguous evaluation against ground truth.

## 3. Extracting Ontological Hierarchy

Raw data typically lacks the context of where a specific diagnosis fits within a broader medical classification. This hierarchical information is vital for analyzing dataset composition (e.g., prevalence of rare vs. common disease categories) and enabling evaluation metrics that understand semantic relationships.

bat29 addresses this by using its knowledge bases to retrieve the full hierarchical path for standardized codes like ICD-10.

**Output Example (Hierarchy columns in test\_death.csv):** | … | icd10\_chapter\_code | icd10\_block\_code | icd10\_category\_code | icd10\_category\_name | icd10\_disease\_group\_code | icd10\_disease\_group\_name | icd10\_disease\_code | icd10\_disease\_name | icd10\_disease\_variant\_code | icd10\_disease\_variant\_name | | :-: | :—————– | :————— | :—————— | :—————————– | :———————– | :——————————— | :—————– | :—————– | :————————- | :————————- | | … | IX | I30-I5A | I50 | Heart failure | I50.9 | Heart failure, unspecified | … | … | … | … | | … | X | J09-J18 | J18 | Pneumonia, unspecified organism| J18.9 | Pneumonia, unspecified organism | … | … | … | … | | … | X | J95-J99 | J96 | Respiratory failure, n.e.c. | J96.9 | Respiratory failure, unspecified | J96.90 | Resp fail, unspec w hypoxia/hypercapnia | … | … |

This explicit hierarchy embedded in the output data allows for powerful analysis of case distribution across different medical specialties or disease groups (see below).

## 4. Generating and Analyzing Standardized Test Sets

The standardized and enriched data produced by the bat29 pipeline serves as the foundation for creating targeted test sets for evaluating LLMs. Each test set represents a collection of clinical cases, formatted consistently and sharing a defined characteristic, allowing for focused assessment of model capabilities across different clinical scenarios or data sources.

### 4.1 Test Set Creation Methodology

Specific test sets are generated by filtering the processed data based on criteria relevant to clinical outcomes, patient demographics, or data origin. This is primarily handled by scripts within the src/bat29/ directory:

* **Outcome/Severity-Based Filtering (treatment\_urg.py):** This script processes the standardized URG hospital records. It examines fields such as motivo\_alta\_ingreso (discharge reason, including “Fallecimiento” for death), est\_planta (days in ward), and est\_uci (days in ICU) to categorize cases. For instance, cases ending in death are assigned to the death set, while cases involving ICU stays or extended hospitalization (>= 18 days) form the critical set. Cases with moderate hospitalization (5-17 days, no ICU) constitute the severe set. Pediatric cases (edad <= 15) are also identified. Subsets limited to the first 1000 entries are created for baseline comparisons (e.g., first\_1000, first\_1000\_severe).
* **Source-Based Processing (treatment\_ramebench\_paper.py):** This script handles data derived from research datasets focused on rare diseases (e.g., HMS, LIRICAL, MME, PUMCH\_ADM). It standardizes the narrative structure (often using phenotype names derived from HPO codes) and diagnosis (using OMIM/ORPHA IDs and synonyms), creating separate test sets for each source dataset (e.g., test\_PUMCH\_ADM.csv) and a combined set (test\_ramebench.csv).

The result of these processes is a suite of .csv files located in /data/tests/treatment/, each containing standardized case, golden\_diagnosis, diagnostic\_code/s, and ICD-10 hierarchy columns, ready for use in evaluation protocols.

### 4.2 Overview of Generated Test Sets

The methodology yields diverse test sets, summarized in the table below:

| Test Set Name | Creation Criteria | Number of Cases |
| --- | --- | --- |
| test\_all.csv | All processed URG cases | 6255 |
| test\_1000.csv | First 1000 processed URG cases | 1000 |
| test\_death.csv | URG cases resulting in death | 7 |
| test\_critical.csv | URG cases with death, ICU stay, or ward stay >= 18 days | 43 |
| test\_severe.csv | URG cases with ward stay 5-17 days, no ICU | 82 |
| test\_pediatric.csv | URG cases with age <= 15 | 1653 |
| test\_1000\_pediatric.csv | Pediatric cases from test\_1000.csv | 335 |
| test\_RAMEDIS.csv | Cases from RAMEDIS rare disease dataset | 624 |
| test\_HMS.csv | Cases from HMS rare disease dataset | 87 |
| test\_LIRICAL.csv | Cases from LIRICAL rare disease dataset | 369 |
| test\_MME.csv | Cases from MME rare disease dataset | 40 |
| test\_PUMCH\_ADM.csv | Cases from PUMCH Admission rare disease dataset | 75 |
| test\_RAMEDIS\_SPLIT.csv | Cases from RAMEDIS (Split) rare disease dataset | 200 |
| test\_ramebench.csv | Combination of all processed rare disease datasets (RAMEDIS, HMS, LIRICAL, etc.) | 1395 |

### 4.3 Dataset Composition Analysis

The embedded ICD-10 hierarchy enables detailed analysis of the diagnostic composition for each URG-derived test set using interactive sunburst plots (available in /data/tests/treatment/icd10\_stats/). These plots visualize the distribution of cases across ICD-10 chapters, blocks, categories, and potentially finer levels using the following conceptual mapping:

#### 1) **test\_all (N=6,255):**

This dataset represents the most comprehensive collection of processed URG Emergency Department encounters (N=6,255) and serves as the primary baseline for assessing the general diagnostic capabilities of Large Language Models. Its composition reflects the broad clinical heterogeneity typical of unsorted emergency presentations before applying outcome or demographic filters. The Analysis of the diagnostic distribution using the embedded ICD-10 hierarchy reveals a distinct profile heavily weighted towards common, lower-to-moderate acuity conditions. The table below summarizes the prevalence of the three most frequent Primary Medical Specialties (ICD-10 Chapters) and other specialties particularly relevant to the dataset’s characterization, including dominant or illustrative sub-categories:

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **(Chapter Total)** | **~36.1%** |
| ↳ Sub Specialty: Resp. URI | J00-J06 | Acute Upper Respiratory Infections | ~22.6% |
| ↳ Disease Group: Pneumonia | J18 | Pneumonia, unspecified organism | ~1.3% |
| **Medical Specialty: Injury/Poison** | **XIX** | **(Chapter Total)** | **~12.9%** |
| ↳ Sub Specialty: Ankle/Foot Inj. | S90-S99 | Injuries to Ankle/Foot | ~3.1% |
| ↳ Sub Specialty: Wrist/Hand Inj. | S60-S69 | Injuries to Wrist/Hand | ~2.9% |
| ↳ Disease Group: Sprain/Disloc | S93/S63 | Sprains/Dislocations (Ankle/Hand) | ~1.8% / ~0.7% |
| **Medical Specialty: Musculoskeletal** | **XIII** | **(Chapter Total)** | **~9.7%** |
| ↳ Disease Group: Dorsalgia | M54 | Dorsalgia (Back pain) | ~3.4% |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **(Chapter Total)** | **~8.1%** |
| ↳ Disease Group: Abd/Pelv Pain | R10 | Abdominal/Pelvic Pain | ~1.8% |
| ↳ Disease Group: Headache | R51 | Headache | ~0.6% |
| ↳ Disease Group: Fever | R50 | Fever of unknown origin | ~0.5% |
| **Medical Specialty: Circulatory** | **IX** | **(Chapter Total)** | **~1.3%** |
| ↳ Disease Group: Heart Failure | I50 | Heart Failure | ~0.2% |
| ↳ Sub Specialty: Ischemic Heart Dz | I20-I25 | Ischemic Heart Diseases | ~0.1% |

*(Note: Percentages are approximate. Sub-categories shown are illustrative.)*

**Conclusion & Implications for Evaluation:** The quantitative profile, detailed above, highlights that test\_all primarily tests an LLM’s diagnostic breadth across common conditions encountered in primary or urgent care settings. The distribution confirms a focus on common respiratory infections, minor injuries, back pain, and frequent non-specific symptoms like pain and fever. This pattern contrasts sharply with the significant underrepresentation of high-acuity conditions, particularly within Circulatory Diseases, where specific critical categories like Heart Failure and Ischemic Heart Disease have very low prevalence. The notable proportion of cases falling under Symptoms, Signs, and Abnormal Findings (Chapter XVIII) underscores the presence of diagnostic ambiguity. Evaluating performance on these less specific presentations is crucial for assessing an LLM’s ability to handle the uncertainty inherent in real-world clinical encounters. Therefore, performance on test\_all reflects a model’s ability to manage a high volume of varied, moderate-acuity cases involving common infections, minor injuries, and non-specific symptoms, rather than its proficiency in diagnosing critical or complex conditions prevalent in specialized or intensive care.

#### 2) **test\_1000 (N=1000):**

This dataset consists of the first 1,000 processed URG encounters, serving as a computationally less demanding, yet substantial, random subsample of test\_all. Its primary utility lies in providing a representative benchmark for initial or comparative evaluations where using the full test\_all dataset may be prohibitive. As expected from a large subsample, its diagnostic composition closely mirrors that of the parent dataset.

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **Respiratory Diseases** | **~39.1%** |
| **Medical Specialty: Injury/Poison** | **XIX** | **Injury, Poisoning, External** | **~10.1%** |
| **Medical Specialty: Musculoskeletal** | **XIII** | **Musculoskeletal Diseases** | **~10.0%** |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **Symptoms, Signs, Abnormal Findings** | **~7.6%** |
| **Medical Specialty: Circulatory** | **IX** | **Circulatory Diseases** | **~1.2%** |

*(Note: Percentages are approximate.)*

The quantitative profile confirms the similarity to test\_all, dominated by respiratory, injury/poisoning, musculoskeletal, and symptom-based presentations, with a similarly low prevalence of circulatory diseases. Therefore, its implications for LLM evaluation are largely identical to test\_all: it primarily assesses diagnostic breadth across common, moderate-acuity conditions and the ability to handle diagnostic ambiguity, rather than performance on critical care or highly specialized cases.

#### 3) **test\_severe (N=82):**

This dataset focuses on cases requiring moderate hospitalization (5-17 days ward stay, no ICU), aiming to represent conditions more serious than typical ED discharges but less critical than those requiring intensive care. It is filtered from the broader URG dataset (N=82).

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **(Chapter Total)** | **~41.5%** |
| ↳ Disease Group: Pneumonia | J18 | Pneumonia, unspecified organism | ~15.9% |
| ↳ Disease Group: Resp.Fail/Oth | J96/J98 | Resp. Failure/Other Disorders | ~7.3% / ~11.0% |
| **Medical Specialty: Genitourinary** | **XIV** | **(Chapter Total)** | **~12.2%** |
| ↳ Disease Group: Renal Colic | N23 | Renal Colic, unspecified | ~2.4% |
| **Medical Specialty: Digestive** | **XI** | **(Chapter Total)** | **~11.0%** |
| ↳ Disease Group: Diverticular Dz | K57 | Diverticular Disease | ~3.7% |
| ↳ Disease Group: Pancreatitis | K85 | Acute Pancreatitis | ~1.2% |
| **Medical Specialty: Special Purpose** | **XXII** | **(Chapter Total)** | **~8.5%** |
| ↳ Disease Group: COVID-19 | U07 | Emergency Use (e.g., COVID-19) | ~8.5% |
| **Medical Specialty: Injury/Poison** | **XIX** | **(Chapter Total)** | **~8.5%** |
| **Medical Specialty: Circulatory** | **IX** | **(Chapter Total)** | **~4.9%** |
| ↳ Disease Group: CVA | I63 | Cerebral Infarction | ~3.7% |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **(Chapter Total)** | **~4.9%** |

*(Note: Percentages are approximate. N=82 is relatively small. Sub-categories shown are illustrative.)*

The quantitative profile of test\_severe, detailed in the table, reflects the filtering criteria for moderate illness severity requiring inpatient care. The distribution shifts significantly compared to test\_all, with increased prevalence of conditions often necessitating admission. While **Respiratory Diseases (X)** remain dominant, the driving categories are now more severe conditions like Pneumonia and Respiratory Failure/Disorders. Similarly, **Genitourinary (XIV)** and **Digestive Diseases (XI)** feature more prominently, represented by conditions such as renal colic, diverticular disease, and pancreatitis. The significant presence of **Codes for Special Purposes (XXII)** likely reflects specific events (e.g., COVID-19). Compared to test\_all, the proportion of ambulatory complaints and non-specific **Symptoms/Signs (XVIII)** is reduced. Although **Circulatory Diseases (IX)** show only a modest increase overall, the presence of conditions like Cerebral Infarction points to higher acuity within this subset. This dataset thus challenges LLMs on diagnoses associated with moderate illness severity requiring hospital management, representing a distinct step up in complexity from the broad baseline but stopping short of critical care scenarios.

#### 4) **test\_critical (N=43):**

This dataset specifically targets high-acuity cases, including those resulting in death, requiring an ICU stay, or involving prolonged hospitalization (>= 18 days). It represents a small (N=43) but clinically significant subset focused on life-threatening conditions.

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **(Chapter Total)** | **~37.2%** |
| ↳ Disease Group: Pneumonia | J18 | Pneumonia, unspecified organism | ~16.3% |
| ↳ Disease Group: Resp.Fail/Oth | J96/J98 | Resp. Failure/Other Disorders | ~7.0% / ~7.0% |
| **Medical Specialty: Circulatory** | **IX** | **(Chapter Total)** | **~18.6%** |
| ↳ Disease Group: Heart Failure | I50 | Heart Failure | ~9.3% |
| ↳ Disease Group: CVA | I63 | Cerebral Infarction | ~2.3% |
| **Medical Specialty: Digestive** | **XI** | **(Chapter Total)** | **~14.0%** |
| ↳ Disease Group: Cholecystitis | K81 | Cholecystitis | ~4.7% |
| ↳ Disease Group: Ileus | K56 | Paralytic Ileus/Obstruction | ~4.7% |
| **Medical Specialty: Genitourinary** | **XIV** | **(Chapter Total)** | **~7.0%** |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **(Chapter Total)** | **~7.0%** |
| **Medical Specialty: Injury/Poison** | **XIX** | **(Chapter Total)** | **~7.0%** |

*(Note: Percentages are approximate and based on a small N=43. Sub-categories shown are illustrative.)*

The diagnostic profile of test\_critical, detailed above, clearly reflects its focus on severe illness. The high prevalence of **Respiratory Diseases (X)** is characterized by conditions like Pneumonia and Respiratory Failure. Crucially, **Circulatory Diseases (IX)** constitute a much larger proportion compared to less severe subsets, driven primarily by critical conditions such as Heart Failure, along with other serious vascular events. **Digestive Diseases (XI)** also feature prominently with severe conditions like cholecystitis and ileus. The relative contributions from less specific **Symptoms/Signs (XVIII)** and **Injury (XIX)** are lower than in the baseline, suggesting a concentration on more clearly defined, life-threatening systemic diseases. Evaluating LLMs on this dataset specifically probes their ability to recognize and diagnose high-acuity conditions where timely and accurate assessment is paramount, offering a distinct challenge focused on critical care knowledge (particularly related to respiratory and cardiac failure) rather than diagnostic breadth.

#### 5) **test\_death (N=7):**

This dataset represents an extremely small (N=7) and highly specific subset containing only cases from the URG dataset that resulted in patient death. Its purpose is to isolate the diagnostic patterns associated with fatal outcomes within this cohort.

| Concept Level | ICD-10 | Description | Count (N=7) |
| --- | --- | --- | --- |
| **Medical Specialty: Circulatory** | **IX** | **Circulatory Diseases** | **2** |
| **Medical Specialty: Respiratory** | **X** | **Respiratory Diseases** | **2** |
| **Medical Specialty: Digestive** | **XI** | **Digestive Diseases** | **2** |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **Symptoms, Signs, Abnormal Findings** | **1** |

*(Note: Due to N=7, counts are shown instead of percentages.)*

Given the extremely small sample size, the diagnostic profile is highly concentrated among conditions directly associated with mortality in this specific cohort. **Circulatory (IX)**, **Respiratory (X)**, and **Digestive Diseases (XI)** each account for two cases, with one case coded under **Symptoms/Signs (XVIII)**. While drawing broad conclusions is impossible due to the low N, this distribution aligns with common causes of in-hospital mortality (e.g., severe cardiac events, respiratory failure, complications of digestive system diseases). Evaluating an LLM on this dataset tests its ability to recognize potentially fatal conditions, although the statistical significance is minimal. It primarily serves as a qualitative indicator of model performance on the most severe end of the clinical spectrum represented in the source data.

#### 6) **test\_pediatric (N=1653):**

This large dataset (N=1653) includes all processed URG encounters for patients aged 15 years or younger. It aims to specifically evaluate LLM performance on common pediatric presentations seen in an emergency setting.

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **(Chapter Total)** | **~55.8%** |
| ↳ Sub Specialty: Resp. URI | J00-J06 | Acute Upper Respiratory Infections | ~37.1% |
| ↳ Disease Group: Bronchiolitis | J21 | Acute Bronchiolitis | ~3.0% |
| ↳ Disease Group: Pneumonia | J18 | Pneumonia, unspecified organism | ~1.8% |
| **Medical Specialty: Injury/Poison** | **XIX** | **(Chapter Total)** | **~12.2%** |
| ↳ Disease Group: Sprain/Disloc | S93/S63 | Sprains/Dislocations (Ankle/Hand) | ~1.8% / ~1.1% |
| ↳ Disease Group: Superficial Inj | S60/S90 | Superficial Injuries (Hand/Foot) | ~1.5% / ~1.0% |
| **Medical Specialty: Ear** | **VIII** | **(Chapter Total)** | **~8.7%** |
| ↳ Disease Group: Otitis Media | H66 | Otitis Media, supp./unspec. | ~7.2% |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **(Chapter Total)** | **~6.0%** |
| ↳ Disease Group: Fever | R50 | Fever of unknown origin | ~1.3% |
| ↳ Disease Group: Abd/Pelv Pain | R10 | Abdominal/Pelvic Pain | ~1.8% |
| **Medical Specialty: Infectious** | **I** | **(Chapter Total)** | **~4.3%** |
| ↳ Disease Group: Viral Inf | B34 | Viral infection, unspecified | ~2.3% |

*(Note: Percentages are approximate. Sub-categories shown are illustrative.)*

The diagnostic profile of test\_pediatric, detailed in the table, clearly reflects common childhood ailments presenting to emergency care. **Respiratory Diseases (X)** are overwhelmingly dominant, largely attributable to high rates of various Acute Upper Respiratory Infections and specific lower respiratory conditions like Acute Bronchiolitis. **Injury/Poisoning (XIX)** primarily involves minor extremity trauma. **Diseases of the Ear (VIII)** are notably prevalent, driven overwhelmingly by Otitis Media. **Symptoms/Signs (XVIII)**, particularly Fever and Abdominal Pain, and non-specific **Infectious Diseases (I)** are also significant contributors. Conversely, specialties common in adults, such as Musculoskeletal (XIII) and especially Circulatory (IX), are much less represented. This dataset provides a valuable benchmark for assessing LLM diagnostic accuracy specifically within the pediatric domain, testing knowledge of common childhood conditions like URIs, bronchiolitis, otitis media, minor trauma, and undifferentiated febrile presentations.

#### 7) **test\_1000\_pediatric (N=335):**

This dataset is a subset of test\_1000, containing only the pediatric cases (age <= 15) from those first 1,000 encounters (N=335). It allows for analysis of pediatric presentations within the specific context of the initial data sample.

| Concept Level | ICD-10 | Description | Prevalence (%) |
| --- | --- | --- | --- |
| **Medical Specialty: Respiratory** | **X** | **Respiratory Diseases** | **~64.2%** |
| **Medical Specialty: Ear** | **VIII** | **Ear Diseases** | **~9.6%** |
| **Medical Specialty: Injury/Poison** | **XIX** | **Injury, Poisoning, External** | **~5.4%** |
| **Medical Specialty: Symptoms/Signs** | **XVIII** | **Symptoms, Signs, Abnormal Findings** | **~4.2%** |
| **Medical Specialty: Infectious** | **I** | **Infectious Diseases** | **~3.9%** |

*(Note: Percentages are approximate.)*

As a pediatric subset of test\_1000, this dataset’s profile strongly resembles test\_pediatric but reflects the specific case mix within the initial 1000 URG records. **Respiratory Diseases (X)** are even more pronounced here (~64.2%) than in the overall pediatric cohort. **Ear Diseases (VIII)** (~9.6%) remain highly prevalent. **Injury/Poisoning (XIX)** (~5.4%) is present but represents a smaller fraction compared to the full test\_pediatric set. **Symptoms/Signs (XVIII)** (~4.2%) and **Infectious Diseases (I)** (~3.9%) contribute moderately. The implications for evaluation are similar to test\_pediatric, focusing on common childhood conditions, particularly respiratory and ear-related ailments, but derived from a smaller, specific cross-section of the overall URG data.

Examining the specific sunburst plots for each set provides quantitative confirmation of these expected distributions and highlights the specific diagnostic challenges inherent in each subset, allowing researchers to select appropriate tests for targeted LLM evaluation.

### 4.4 Future Directions for Test Set Creation

The standardized framework facilitates the creation of further specialized test sets to probe LLM performance more deeply:

* **Domain-Specific:** Filtering by icd10\_chapter\_code to create sets for cardiology (Chapter IX), pulmonology (Chapter X), etc.
* **Demographically Stratified:** Creating subsets balanced by age or sex to investigate potential biases.
* **Symptom-Oriented:** Developing sets based on primary presenting symptoms, if reliably extractable.
* **Hybrid Sets:** Combining cases from different sources (e.g., URG pediatric and Ramebench pediatric) for broader scope.
* **Refined Outcome Sets:** Creating sets excluding symptom-based codes (Chapter XVIII) to focus on definitive diagnoses.

Such targeted datasets will be invaluable for nuanced and robust evaluation of clinical diagnostic LLMs.