## Supplementary Methods: Rule-Based Classification of Phenotype

### 1 Overview of the Classification Framework

This document details a multi-layered rule set developed through extensive manual curation combined with AI-assisted recommendations from Google Gemini 2 Pro, designed for the automated classification of clinical phenotype descriptions in the ClinVar database.

The primary goal of this rule set is to systematically identify phenotypes associated with severe, life-threatening genetic disorders, thereby facilitating the inference of clinical severity for related genetic variants.

The classification system is designed to emulate the decision-making process of clinical genetics experts. It adopts a pattern-driven approach that goes beyond simple keyword matching, instead interpreting combinations of medical terms. Phenotypes are hierarchically categorized into three final classes: "Confirmed Lethal," "High Severity," and "Non-severe." This structure yields a high-precision list of definitively fatal conditions while also capturing a broad spectrum of severe diseases with high recall.

### 2 Algorithm and Logic Flow

The classification algorithm processes each unique phenotype string through a sequential, prioritized pipeline:

**Step 1: High-Priority Exclusion (Rule Zero)**: The algorithm first applies a set of high-priority exclusion rules to filter out phenotypes that are unambiguously non-severe, represent risk states, or contain mitigating modifiers. A match at this stage results in a "Non-Severe" classification, and no further rules are evaluated.

**Step 2: General Exclusion**: Phenotypes that pass Rule Zero are then checked against a list of non-lethal, late-onset, or effectively manageable conditions.

**Step 3: Confirmed Lethal Classification**: The remaining phenotypes are evaluated against high-confidence rules designed to identify phenotypes with near-certain perinatal, neonatal, or infantile mortality.

**Step 4: High Severity Classification**: Phenotypes not meeting the Confirmed Lethal criteria are assessed against a broader set of rules that capture conditions associated with significant premature mortality, catastrophic clinical events, or a fatal natural history.

**Step 5: Default Classification**: Any phenotype that does not match an inclusion or exclusion rule is assigned a default classification of "Non-Severe".

### 3. The Rule Set

The complete rule set is detailed in the following tables.

**Table S1: Exclusion Rules** *These rules are applied with the highest priority to filter out non-relevant phenotypes.*

| Rule ID | Category | Keywords / Logic (Complete) | Rationale & (Exemptions) |
|---|---|---|---|
| **E1** | Non-Lethal Modifiers | mild, incomplete, partial, delayed onset, -induced, drug-associated, exercise-induced, in situ, unilateral, equivocal | These terms explicitly indicate a deviation from the classic, most severe form of a disease. **(Exemptions: 'mild' is ignored in the context of 'with mild'; 'atypical' is handled by Rule E2)**. |
| **E2** | Atypical Modifier | atypical | The term "atypical" usually implies a less severe or non-classic disease course. **(Exemptions: Ignored for 'atypical teratoid rhabdoid tumor'. If conflicting with a Confirmed Lethal term, the phenotype is escalated to High Severity for manual review)**. |
| **E3** | Risk / Susceptibility Status | susceptibility to, predisposition, risk, family history of, finding | These terms describe a risk state rather than a diagnosed disease entity. **(Exemption: If associated with high-risk malignancy keywords (H6), the phenotype is classified as High Severity)**. |

| | | | |
|---|---|---|---|
| **E4** | General Non-Lethal Conditions | achondroplasia, Rett syndrome, Angelman syndrome, Prader-Willi syndrome, Huntington disease, Alzheimer disease, Parkinson disease, benign familial, phenylketonuria, galactosemia, biotinidase deficiency, sweat chloride elevation, modifier of, resistance to, protection against, response to, polymorphism | A curated list of well-known non-lethal, primarily disabling, late-onset, or effectively treatable disorders and non-disease terms. |

**Table S2: Core Lexicon for Pattern Matching** *This lexicon provides the semantic components for pattern-based rules C4, C5, and H5.*

| **Component** | **Regular Expression** |
|---|---|
| **SEVERITY** | (severe\|acute\|fulminant\|progressive\|malignant) |
| **EARLY_ONSET** | (neonatal\|infantile\|congenital) |
| **CRITICAL_SYSTEMS** | (encephalopathy\|cardiomyopathy\|failure\|leukodystrophy\|leukoencephalopathy\|atrophy\|atresia\|agenesis) |
| **CATASTROPHIC_FINDINGS** | (brainstem anomalies\|high lactate\|pyloric atresia) |

**Table S3: Inclusion Rules for Confirmed Lethal and High Severity Classification** *These rules are applied sequentially after a phenotype has passed all exclusion checks. All keywords are listed completely.*

| **Rule ID** | **Severity Level** | **Clinical Category** | **Keywords / Patterns (Complete)** |
|---|---|---|---|
| **C1** | **Confirmed Lethal** | Absolute Lethality | lethal, fatal, thanatophoric, hydrops fetalis |
| **C2** | **Confirmed Lethal** | Infantile-Lethal Syndrome | achondrogenesis, zellweger syndrome, walker-warburg, meckel syndrome, meckel-gruber, neu-laxova, pena-shokeir, fowler syndrome, alpers-huttenlocher, alpers syndrome, menkes disease, krabbe disease, tay-sachs, sandhoff disease, canavan disease, wolman disease, niemann-pick disease, type a, herlitz, restrictive dermopathy, hydrolethalus syndrome, gracile syndrome, rhabdoid tumor, leigh syndrome, leigh-like, atypical teratoid rhabdoid tumor, glycine encephalopathy, nonketotic hyperglycinemia, alveolar capillary dysplasia, pontocerebellar hypoplasia, type 2, infantile neuroaxonal dystrophy, inad, ethylmalonic encephalopathy, arc syndrome |
| **C3** | **Confirmed Lethal** | Lethal Subtype | (Osteogenesis imperfecta + type II/perinatal), (Hypophosphatasia + infantile/perinatal), (Spinal muscular atrophy / sma + type 0/I/Werdnig-Hoffmann), (Pompe disease / Glycogen storage disease, type II + infantile), (Marfan syndrome + neonatal), (Osteopetrosis + infantile/malignant), (Epidermolysis bullosa + pyloric atresia), (Urea cycle + neonatal) |
| **C4** | **Confirmed Lethal** | Lethal Clinical Pattern | SEVERITY + EARLY_ONSET + CRITICAL_SYSTEMS |
| **C5** | **Confirmed Lethal** | Catastrophic Pattern | CRITICAL_SYSTEMS + CATASTROPHIC_FINDINGS |
| **H1** | **High Severity** | Potentially Treatable but Lethal | severe combined immunodeficiency, scid, hemophagocytic lymphohistiocytosis, fhl, urea cycle, maple syrup urine disease, msud, propionic acidemia, methylmalonic acidemia, glutaric aciduria type 1, cystic fibrosis, duchenne muscular dystrophy |
| **H2** | **High Severity** | High Risk of Sudden Death | arrhythmogenic, long qt syndrome, brugada syndrome, catecholaminergic polymorphic ventricular tachycardia, cpvt, pulmonary arterial hypertension, marfan syndrome, loeys-dietz syndrome, ehlers-danlos syndrome, vascular type, aortic dissection |
| **H3** | **High** | Variable / | niemann-pick disease, type c, gaucher disease, type 3, ataxia- |

| | | | |
|---|---|---|---|
| | **Severity** | Later Onset Lethality | telangiectasia, fanconi anemia, dyskeratosis congenita, adams-oliver syndrome, diaphragmatic hernia, hypoplastic left heart syndrome, renal agenesis, renal aplasia |
| **H4** | **High Severity** | High-Risk Metabolic Crisis | mcad, vlcad, lchad deficiency, mitochondrial trifunctional protein deficiency, salt-wasting |
| **H5** | **High Severity** | Progressive Critical System Disease | (progressive) + CRITICAL_SYSTEMS |
| **H6** | **High Severity** | Heritable Malignancy | cancer, carcinoma, sarcoma, lethal_pheno_confirmed_v11.txt, lymphoma, blastoma, glioma, melanoma, malignant |

## 4 Final Output Structure

The script generates a single tab-separated values (TSV) file containing all phenotypes classified as either Confirmed Lethal or High Severity. The file includes the following columns to ensure full transparency and facilitate downstream analysis:

**Phenotype**: The original, unmodified phenotype string.

**Severity_Level**: The final classification: Confirmed Lethal or High Severity.

**Clinical_Category**: A descriptive clinical grouping (e.g., Infantile-Lethal Syndrome, High Risk of Sudden Death).

**Matched_Rule_Or_Pattern**: The specific keyword or pattern that triggered the classification.

**Reasoning_Note**: A brief, human-readable explanation of the classification rationale.

This structured output allows for robust, reproducible analysis and clear reporting of the phenotype classification methodology.