Data Quality Guidelines for Clinical Trial Data

1. Purpose

These guidelines define the minimum data quality standards for life sciences datasets, ensuring accuracy, completeness, consistency, and interoperability for downstream analysis and regulatory compliance.

2. Required Data Fields

- study_id: Unique trial identifier (e.g., NCT number)
- intervention_name: Official drug or treatment name
- condition: Medical condition studied (MeSH preferred)
- sponsor: Responsible organization
- start_date, end_date: ISO 8601 format (YYYY-MM-DD)
- phase: One of {Phase 1, Phase 2, Phase 3, Phase 4}
- status: One of {Recruiting, Completed, Terminated, Withdrawn}

3. Data Entry Standards

- Use title case for text fields
- Use controlled vocabulary for intervention_name and condition
- Avoid abbreviations unless standardized
- Dates must be valid and chronological (end_date ≥ start_date)

4. Ontology Mapping Rules

- Intervention names must be mapped to DrugBank or RxNorm terms
- Mapping confidence score ≥ 90% is required for acceptance
- Low-score or unmapped entries must be reviewed manually

5. Data Quality Checks

- Null value rate < 5% per column
- No duplicate study_id entries
- Phase and status values must match allowed list
- Completeness score must be ≥ 0.95 for final datasets

6. Review and Governance

- Perform monthly data audits
- Update controlled vocabularies quarterly
- Maintain change log for data corrections
- Document root causes for data quality issues and remediation steps