**Clinical Studies Exploration and Harmonization Ontology (CHEO): For secondary use of data**

**Ontology for Clinical Study Metadata and Data Analysis**

**Version**

* **Draft**: v2.0
* **Date**: *07/03/2025*

**Authors and Contributors:** *Komal Gilani*

**Table of Content:**

1. **Introduction (CHEO)**
   1. Motivation and Need for a Semantic Framework
   2. Overview of CHEO’s Purpose and Scope
2. **Key Structural Components of CHEO** (New Section)
   1. **Specification Modeling** (Study process, protocol, and study data elements)
   2. **Observation Representation** (Patient-level data and its connection to study specifications)
3. **Ontological Foundations and Conceptual Structure**
   1. BFO-based Classification (Continuants vs. Occurrents)
   2. Data elements, Observations, and their Relationships
4. **Study Metadata and Harmonization Specifications**
5. **Implementation Strategy and Future Development**
6. **Conclusion**
7. **Introduction**

**Data integration, secondary use of data**

The integration of multi-center clinical research data remains a major bottleneck due to inconsistencies in study methodologies, non-standardized data collection protocols, and the fragmented use of clinical terminologies across different electronic health record (EHR) systems. These discrepancies create substantial challenges in data harmonization, leading to misalignment in data element definitions, incompatible data formats, and semantic ambiguities that hinder cross-study comparability. Without a structured approach, efforts to standardize datasets for secondary use—such as meta-analyses, predictive modeling, and regulatory submissions—are often ad hoc, labor-intensive, and prone to bias.

Several existing ontologies provide partial solutions for clinical study modeling but remain **domain-specific and fragmented** in their coverage. The **Ontology for Biomedical Investigations (OBI)** and the **Ontology of Biological and Clinical Statistics (OBCS)** capture aspects of study specification and statistical analysis, respectively, while the **Statistical Methods Ontology (STATO)** and its extension OBCS focuses on statistical methodologies. Other ontologies, such as the **Clinical Trial Ontology** and **Cohort Ontology,** address specific study designs but lack broader integration with clinical study metadata and patient-level data harmonization. Despite these contributions, **no unified ontology exists that comprehensively integrates study design specification, data elements specification, and patient-level observations within a single harmonized framework.**  Addressing these challenges requires an ontology-driven solution, adhering to foundational principles, that provides a unified semantic framework for metadata alignment, cohort integration, and automated data transformation.

To address this gap, we introduce the **Clinical studies exploration and harmonization ontology (CHEO)**, which is designed to provide a structured and interoperable approach to clinical study metadata representation to aid in finding suitable studies and assess potential for harmonization. It is builds on foundational ontology principles to ensure **semantic interoperability, automated reasoning, and cross-study comparability**. CHEO is designed around **two fundamental components: Specification Modeling**and**Observation Representation.** These components ensure a comprehensive and semantically rich representation of both study design and patient-level data, enabling seamless data harmonization and interoperability across studies.

1. **Specification Modeling** captures the **study process, its protocol, and the structured definition of data elements.** This includes the formal specification of study data elements, measurement protocols, and the intended data collection methodology. By defining these elements within a standardized ontological framework, CHEO ensures that study metadata is **semantically structured, interoperable, and reusable** across different research contexts. The study data element design specification serves as a blueprint, detailing **the expected structure, value constraints, measurement units, and categorical classifications of clinical data elements**, thereby ensuring that collected data adheres to standardized research protocols.
2. **Observation Representation**focuses on the **individual patient-level data and its relationship to the defined study specifications.** This component ensures that observed data points are **systematically linked to the study’s conceptual framework**, facilitating semantic alignment between study metadata and actual patient data. Each observation is modeled as a **data item**, representing a specific recorded value associated with a participant. These data items are directly tied to the corresponding study data elements and are further connected to the **participant's role in the study (e.g., participant under investigation, control group, or intervention group).** By maintaining **structured linkages between study metadata and real-world observations**, CHEO supports **data integrity, harmonization, and reusability** in clinical research.

Biomedical Investigations (OBI). OBI provides a structured representation of key study components, including study design specifications, measurement protocols, experimental conditions, and research activities. This enables CHEO to capture the directive and procedural aspects of clinical research, ensuring that study metadata is aligned with established investigation models. For statistical and exploratory data analysis, CHEO incorporates the Statistical Methods Ontology (STATO), which provides standardized concepts for statistical methodologies, data analysis techniques, and inferential procedures. Since OBI, STATO and its extension OBCS, all are built upon BFO, their integration within CHEO ensures a coherent ontological structure that supports reasoning across both study design metadata and data analysis processes. Beyond study metadata and statistical modeling, CHEO extends into biomedical domain-specific standardization by aligning study data elements with external biomedical ontologies. Each data element, whether it represents a clinical measurement, laboratory test, or medical condition, is mapped to an appropriate biomedical reference ontology such as SNOMED CT, LOINC, OMOP, or CDISC. This alignment process is performed via an automated mapping mechanism, ensuring that study data elements adhere to recognized terminologies while maintaining interoperability across clinical datasets.

By leveraging BFO, OBI, STATO and OBCS, and external biomedical ontologies, CHEO establishes a harmonized framework for structuring both study metadata and patient-level observations. This integrated approach ensures that CHEO not only provides a robust ontological model for study design and statistical analysis but also enables semantic alignment of data element with biomedical terminologies, supporting scalable and interoperable research data harmonization

### **Need to Clarify Our Objective**

#### **Primary Goal**

The primary goal is to develop a unified ontological framework (hyper-ontology) that integrates multiple existing ontologies, supported by foundational ontologies tailored for healthcare contexts. This framework aims to standardize both clinical study metadata specification and patient-level observational data while providing a unified representation for statistical analyses across studies. Existing ontologies built for one or more similar objectives includes Ontology of Biomedical Investigation, statistics ontology and ontology for biomedical and clinical statistics, all are built upon Basic Formal Ontology (BFO) principles. Thet can be extended for coverage to detailed definitions of data elements and addressing gaps in patient-level data representation. Inspired by real-world clinical applications, CHEO emphasizes data harmonization between clinical studies by representing study design specification and data elements specified and observation values for each data element with exploratory data analysis in unified manner.

### **Objectives**

* **Represent Clinical Study Metadata:** Clearly define protocols, data elements, and statistical roles.
* **Real-World Application Focus:** Emphasize summarized, decision-critical information over granular procedural details.
* **Interoperability:** Align local data elements with standardized terminologies (LOINC, SNOMED) and ensure compatibility with top-level ontologies.
* **Advanced Querying and Reasoning:** Support queries like identifying harmonizable data elements across studies.
* **Reusability:** Enable data reuse in various clinical and analytical contexts.
* **Support Secondary Research Use:** Facilitate interoperability critical for meta-analyses, epidemiological research, outcome comparisons, and predictive analytics.
* **Key Clarifications Needed**:
* Approach to mapping data element types to OMOP classes or granular clinical classification (not available in metadata dictionaries).
* Specification of the focus on prospective versus retrospective data. Can we capture both with same model?
* level of detail required regarding measurement processes versus storage of finalized measurement values.

### **Vision**

* ***Short-Term (6 months)***
  + Draft initial ontology with core classes and properties.
  + Demonstrate feasibility with core clinical variables (e.g., NYHA Classification, Blood Pressure, BMI).
  + Implement initial mappings to standard terminologies (LOINC, SNOMED).
  + Pilot test ontology in real-world and simulated analytical workflows.

#### **Long-Term Vision**

* + Extend ontology coverage to additional data elements and statistical methods.
  + Establish governance, versioning, and metadata management protocols.
  + Establish CHEO as a reference ontology for clinical research metadata and statistical analyses, bridging existing standards such as OBI and OBCS/STATO.
  + Enhance metadata governance through ISO 11179 principles.
  + Enable automated reasoning capabilities for cohort selection, data integration, and advanced analytics.
  + Promote broad adoption across multi-center research networks and internal departmental settings.

#### **Competency Questions (CQs):**

The ontology design will address questions related to:

* **Clinical Studies:** Identification of clinical study protocol, endpoints and documentation of specifications including variables specification and their extensive metadata representation. It includes
  + How many common data elements all studies have?
  + Which studies are about disease X?
  + Which studies follows study design A with inclusion criteria Y?
  + How many common data elements have different contextual factor and time points?
  + Return all cohorts of x disease and age range y for which I have the permission for research use-case Z? (permission specification module not included yet)
* **Data Harmonization:** Differentiating measured from derived elements, mapping standard terminologies, assessing necessary transformations.
  + How many data element of Study A align with Study B data elements?
    - Can we use standardized Code for checking harmonization?
    - can use other technique on populated knowledge graph that predict infer matches other than existing codes?
* **Statistical Analysis:** Identification of data elements involved in statistical modeling; specification of statistical tests applied.
  + Give me all study with cohort study design and age range X with diagnosis Y of grade Z?

### **Background**

The harmonization and integration of clinical data originating from diverse sources necessitate the application of standardized ontological frameworks to achieve semantic interoperability and precise representation. To facilitate integration across clinical research and studies, it is essential to understand the roles and limitations of various ontology types. Foundational (upper-level) ontologies provide a general, abstract framework supporting cross-domain alignment, serving as the backbone for structured representation in specialized contexts. Domain ontologies focus deeply on specific subject areas, offering detailed conceptualizations but typically lacking broader contextual elements such as study designs or methodological frameworks. Application ontologies bridge this gap by integrating foundational and domain-specific concepts, offering tailored solutions for specific use cases, including clinical study design and execution. Metadata registries support these ontologies by systematically managing and governing data elements, ensuring semantic clarity and consistency across implementations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type | Focus/Scope | Strengths | Limitations | Examples |
| Domain Ontologies | Specific to a particular domain (e.g., clinical terms, lab tests) | Extensive coverage and detailed modeling of domain-specific concepts | Limited coverage on aspects like study design or statistical methodologies | SNOMED CT, LOINC |
| Upper (Foundational) Ontologies | Very general, abstract frameworks that cover high-level concepts (e.g., entities, processes) | Provide top-level classes that facilitate integration across domains | Lack detailed domain-specific content | BFO, DOLCE |
| Application Ontologies | Built upon domain and upper ontologies to address specific applications (e.g., clinical studies) | Tailored to specific use cases; integrate multiple sources and provide context | May require alignment with several standards and can become complex | OBI (for biomedical investigations), STATO (for statistical methods) |
| Hyperontology | Integrates multiple ontologies across different abstraction levels to enhance interoperability | Enables semantic interoperability across diverse domains | Increased complexity; requires robust mapping and maintenance strategies | Meta-ontology frameworks (conceptual; less standardized in practice) |
| Metadata Registries | Structured approaches for defining, managing, and governing data elements | Enhance semantic clarity, provide versioning, and stewardship of data elements | Not fully designed for reasoning or inferencing like formal ontologies | ISO 11179, CDISC |

1. **Methodology**

The development of the Clinical Harmonization & Exploration Ontology (CHEO) followed a structured methodology to ensure a consistent, interoperable, and semantically rich representation of study specifications, variables, and statistical analyses. The ontology design was guided by three key objectives:

1. ***Selection of Top-Level Ontology***

To ensure maximum semantic interoperability, a comparative evaluation was conducted between the most widely used upper ontologies in the biomedical domain: DOLCE and BFO. The evaluation was guided by three key criteria: Community activity metrics, Number of citations in the biomedical field, Number of biomedical ontologies that reuse each framework as indicated in Table 2. A survey of ontological reuse from BioPortal, publication counts from PubMed, and citation metrics from Google Scholar were used to assess the adoption and impact of each top ontology. The analysis, aligned with standard ontology selection practices, confirmed that BFO is more widely adopted in biomedical research than DOLCE. While DOLCE adopts a cognitive bias, BFO follows a realism-based approach, making it preferable for ensuring consistent representation of biomedical entities. However, since the focus of CHEO is pragmatic semantic interoperability, the realism vs. cognitive bias debate was considered secondary.

Based on the evaluation, BFO and its corresponding Open Biological and Biomedical Ontology (OBO) Foundry consortium were chosen as the top ontology framework for CHEO, for the following reasons:

* BFO is reused in more biomedical ontologies compared to DOLCE, ensuring greater interoperability across clinical and biomedical knowledge domains.
* BFO is cited in more research articles—both in the biomedical field and in general—indicating stronger adoption within scientific literature.
* The BFO community demonstrates higher activity in the biomedical domain, fostering ongoing development and support.

1. **Importing Foundational Ontologies**

To **maximize interoperability and avoid redundancy**, CHEO was built upon **widely adopted biomedical ontologies,** ensuring that study variables, study specifications, and statistical concepts were aligned with existing standards. Instead of developing a new framework from scratch, we selected **established ontologies** that have been extensively validated and used in the biomedical research community. By searching with keywords [“study design” , “protocol”, “variable”, “participant under investigation role”, eligibility critieria” , … ] on bioportal and OLS lookup, we found more than 100 ontologies in which these terms have been redefined. We filter out all ontologies that don’t adhere to BFO completely as well those that have few given keywords.With this process , we choose following ontologies

* OBI (Ontology for Biomedical Investigations) was chosen as the primary ontology, as it provides a structured representation of study designs, study processes, and information content entities (ICEs) relevant to CHEO’s scope.

**Table 2. The choice of a top ontology: results of the comparison between DOLCE and BFO top ontologies**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criterion** | **BFO** | **DOLCE** | **Source** |
| Projects using BFO | 23 (bioportal) | 4 (verified) | bioportal search 8/03/2025, For DOLCE: [https://www.loa.istc.cnr.it/old/Projects.html 7/03/2025](https://www.loa.istc.cnr.it/old/Projects.html%207/03/2025), <https://www.loa.istc.cnr.it/old/dolcus.html> |
| Number of biomedical ontologies reusing the top ontology | 125 (OBO Foundary ) | 4 (google scholar) | OBO Foundary https://obofoundry.org and google scholar on 7/03/2025 |
| Number of articles mentioned ontology in Pubmed | 62 (pubmed) | 14(pubmed) | Pubmed search with keywords "BFO ontology" and "DOLCE ontology" on 8/03/2025 |
| Community activity | Maintained | Maintained | Visited https://fairsharing.org on 7/03/2025 |
| Number of citations | 4850 (google scholar) | 1080 (google scholar) | google scholar search with keywords "Basic Formal Ontology" and "DOLCE ontology" on 8/03/2025 |
| Number of bioportal ontologies reusing the top ontology | 439 | None | BioPortal, calculating the number of ontologies via "Mapping" TAB . 7/03/2025 |

* STATO (Statistical Methods Ontology) was selectively integrated to represent statistical concepts and methods essential for harmonizing study datasets and data transformations. Only the Statistics class and its child concepts were imported, ensuring that CHEO remains focused and avoids unnecessary complexity.
* Both ontologies were downloaded from BioPortal and imported into Protégé, forming the foundational structure of CHEO.

1. **Ensuring Ontological Consistency**

To ensure logical consistency by adhering to the principles of Basic Formal Ontology (BFO) v2.0 and domain-specific constraints., we set following criteria in place:

* BFO v2.0 relationships are prioritized, ensuring that CHEO remains consistent with upper-level ontological constraints.
* Relationships from OBI were adopted where necessary, particularly for defining data transformations and study process relationships (e.g., "is\_specified\_input\_of" and "is\_specified\_output\_of").

### **CHEO Model**

Ontologies provide a structured, formalized representation of knowledge by defining concepts, their relationships, and the constraints governing them. In clinical research, ontologies facilitate semantic interoperability, enabling the integration of diverse datasets by providing machine-readable definitions of real-world entities. They support reasoning over data, allowing for automated inference, validation, and harmonization across studies.

A crucial distinction in data modeling is between metadata and instance data. Metadata describes the **structure, definition, and semantics** of study data elements, offering contextual information about data elements. A data element is a more concrete unit of data specification describing how a variable is operationalized, implemented, or recorded within a specific database, electronic health record (EHR), form, or dataset. It represents a specific instantiation or operational representation of a variable within a particular context or data collection instrument. Includes additional metadata, such as data type, permissible values, coding scheme, units of measurement, and collection instructions. For examples: "Systolic Blood Pressure (mmHg) at Baseline Visit" (data element) corresponds to the clinical variable "Blood Pressure.", "BMI calculated at 12-month follow-up using measured height and weight".

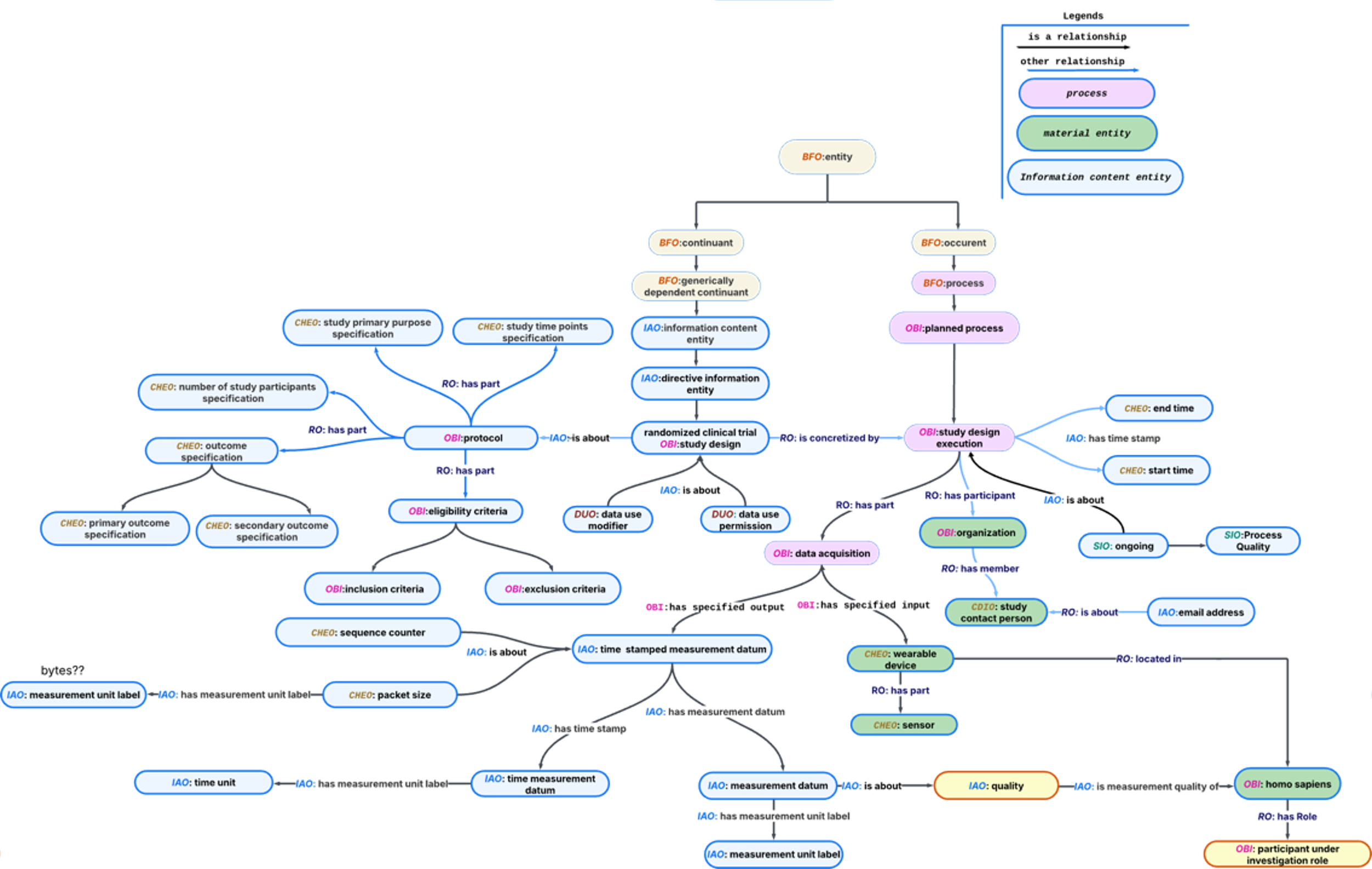
CHEO is built upon the **BFO,** which classifies entities into continuants and occurrents. A **continuant** is any entity that persists over time, even as its properties may change. Within continuants, there are independent continuants, such as physical objects like medical devices or human subjects, and dependent continuants, such as qualities, roles, or dispositions that rely on another entity for existence. Generically dependent continuants, such as data specifications, can be instantiated multiple times in different contexts while maintaining their conceptual identity. In contrast, occurrents represent entities that unfold over time, such as clinical procedures, patient visits, or data collection processes. Occurrents such as planned processes to represent study design execution process is used in CHEO’s core model. The temporal region, an occurrent, is important concept for capturing **temporal aspects of study execution and patient interactions.**

* 1. Study Design

The CHEO model formalizes the representation of study design specification, ensuring a structured and semantically rich framework for modeling clinical study execution as shown in Figure 1. The ontology extends concepts from the Ontology for Biomedical Investigations (OBI) to capture essential components of study design, execution, and their relationships to study variables. By incorporating detailed representations of protocol specifications, study participants, temporal aspects, and execution processes, CHEO facilitates interoperability and harmonization across studies. A study design execution is conceptualized as a process that realizes a study plan, which itself concretizes a study design. In alignment with OBI, the study design is modeled as a plan specification, providing a structured framework for defining study objectives, methodologies, and expected outcomes. The execution of a study design follows a temporal trajectory, ensuring that all aspects of the study are formally linked to a defined timeline, participants, and protocol requirements.

CHEO captures the core components of a study design by structuring it as a composite entity composed of multiple interrelated parts. At its foundation, a study protocol is an essential component of the study design, encapsulating key directives that guide study execution. The protocol includes elements such as eligibility criteria, which define the conditions for participant inclusion or exclusion, and outcome specifications, which outline the primary and secondary objectives of the study. The primary purpose specification further refines the overarching goal of the study, distinguishing between observational, interventional, and exploratory research paradigms. All these elements are classified as directive information entities, emphasizing their role in guiding the structured execution of the study.

Figure Study specification and data element modelling in CHEO



To ensure the temporal integrity of study execution, CHEO models the temporal region associated with each study design execution process. This representation includes specific temporal instances that define the start and end points of the study. The first temporal instance marks the study's initiation, while the last temporal instance signifies its completion. These time points are modeled as time-based temporal entities, providing a structured means to capture the chronological scope of study execution.The execution of a study involves multiple stakeholders, including research institutions, coordinating centers, and individual study personnel. CHEO defines these relationships to ensure a precise representation of participant roles and organizational hierarchies. The study design execution process includes a ***data coordination center role***, which is assigned to an ***organization*** responsible for overseeing the study's data management and execution. This organization, in turn, has designated ***members***, including a ***study contact person*** who serves as a primary liaison for inquiries related to the study. The study contact person is further characterized by specific attributes, such as an ***email address*** and institutional affiliation, ensuring traceability and structured communication within the study framework.

In addition to its structural and temporal attributes, the study execution process possesses intrinsic characteristics that define its operational status. A ***process status*** is assigned to the study execution, representing its current stage—whether ***planned***, ***ongoing***, ***completed***, or terminated. This characteristic is modeled as a ***quality***, ensuring that the execution of a study is dynamically tracked and contextualized within the broader research lifecycle. The

* 1. Study Design Variable Specification: Concept

Within CHEO, key entities align with BFO’s classification to ensure a structured representation of study specifications, protocols, and patient-level observations. Rather than classifying clinical conditions as qualities or dispositions, we focus on representing data elements that define how clinical data elements, study data elements, or classifications are recorded within a study. A data element in CHEO represents a standardized data element specified within a study design and protocol, defining how a particular measurement, classification, or categorical assignment is structured. A data element denotes a clinical variable. The clinical variable ("Blood Pressure") is a single conceptual entity. Each data element provides specific context (timing, method, device, etc.) for collecting or operationalizing this single variable, resulting in multiple, distinct data elements. Data elements are part of a study's variable specification, explicitly defining how each clinical variable is operationalized, measured, recorded, and represented within a particular study design. For example, the clinical variable "Blood Pressure" could be operationalized into distinct data elements such as "Baseline Systolic Blood Pressure measured in clinic (mmHg)", "24-hour Ambulatory Blood Pressure Monitoring (ABPM) mean systolic BP (mmHg)", and " systolic blood pressure follow-up 3 months” Each data element provides specific implementation details, including measurement technique, timing, and units, ensuring clarity and consistency within the study’s data collection and analysis protocols. The conceptual Model

Observed values collected from participants are represented as data items, which serve as instances of their respective data elements. For instance, if a study records "Patient X has a Baseline Systolic Blood Pressure of 130 mmHg measured in clinic," this data item is an instantiation of the "Baseline Systolic Blood Pressure measured in clinic" data element. While this observed value could, in principle, be connected to a quality or disposition (e.g., hypertension severity as a disposition of the patient), CHEO does not explicitly model such connections (possible via bfo:isAbout). Instead, CHEO’s primary role is to structure study data elements and observed values, ensuring that data elements are consistently defined and harmonized across studies.

Observational: In order to connect study participants with their respective observations, each **participant** (obi:homo\_sapiens) is uniquely identified by a **participant identifier** (iao:symbol). The participant is assigned a **role** (CHEO:Role), representing their involvement in the study or investigation. This role is then linked to **data item** (iao:ICE), which represent actual observation recorded for specific data element at some point in time.

While data elements serve as conceptual definitions**, data items represent the recorded values associated with specific patients or study participants**. Both data element and data items are modeled as generically dependent continuants, meaning they can be duplicated and stored in multiple systems while remaining tied to their corresponding data elements. Within CHEO, data element represents structured study data elements that define how clinical measurements, classifications, and patient attributes are recorded. Unlike some ontologies that explicitly distinguish between measured vs. derived data items, CHEO does not categorize data elements based on whether they are directly measured or algorithmically calculated, as this distinction is not necessary for its core purpose of study specification and harmonization. Instead, CHEO categories data elements into 13 granular clinical domains, via categorization process, ensuring that data elements are systematically grouped based on their relevance to clinical research. These categories facilitate semantic structuring, cross-study harmonization, and efficient querying of study data elements.

Each data element in CHEO consists of multiple components, reflecting the structure commonly used in clinical data dictionaries and standardized data formats. These components include categorical value specifications, which define permissible choices or coded values, as well as temporal references, modeled as subclasses of Information Content Entities, to capture the timing of data collection such as baseline, follow-up 1 month etc. Additionally, data elements may include measurement units, ensuring consistency in quantitative values, and contextual references, which specify conditions under which the measurement was taken (e.g., blood pressure recorded in a recumbent position). If applicable, calculation formulas used to derive certain values are also captured, along with missing value specifications to account for data completeness and quality assurance. By structuring variables in this way, CHEO ensures that study data is semantically rich, interoperable, and well-defined for harmonization and secondary use.

* 1. Study Design Variable Specification: Relationships

To structure semantic relationships among entities, CHEO defines a set of formal relationships that establish connections between study data elements, observations, and metadata. Hierarchical relationships, such as **is\_a**, indicate subclass relationships, ensuring that study data elements and measurement concepts follow an organized classification. For instance, "Cardiologist is a Physician" reflects a hierarchical relationship. Part-whole relationships, such as **part\_of**, specify compositional dependencies, ensuring that entities are correctly nested within larger structures. An example of this is "Heart is part of the Human Body." These relationships follow **BFO’s constraints**, where processes can only be part of other processes, and continuants can only be part of other continuants. Associative relationships capture complex interactions between clinical concepts, such as causality, dependencies, or role-based participation. For example, the **has\_participant** property links an occurrent (a study process) to a continuant (a patient or investigator involved in the study). Since only Realizable entities can have ***roles*** in BFO, defined ‘***is specified output’/is specified input’*** in order to use information content entity as input and output, OBI. The ***is\_about*** relationship connects a data element or data item to the another entitie which can be another information content entity or underlying clinical phenomenon it describes.

In addition to hierarchical and associative relationships, CHEO captures ***spatial and temporal relationships,*** which define how entities exist in space and time. Spatial relationships such as ***exists at*** specify the physical location of an entity, while temporal relationships such as ***occurs\_at*** define the occurrent existence in temporal region These properties are essential for modeling **study execution timelines, patient follow-up schedules, and event-based cohort studies,** ensuring that both retrospective and prospective study designs are accurately represented within the ontology.

By integrating fundamental ontological principles, CHEO establishes a comprehensive semantic framework for structuring study specifications, clinical observations, and harmonized data elements. Each data element serves as an input to the standardization process, where it is assigned a standardized representation in the form of a code from biomedical terminologies, accompanied by its respective label. To ensure consistency and interoperability, these standardized codes are linked to OMOP IDs, providing a unique identifier for each data element. This structured approach facilitates semantic integrity, interoperability, and reusability of clinical datasets, allowing for seamless cross-study harmonization and secondary data use while maintaining alignment with external ontologies such as SNOMED CT, LOINC, OMOP, and CDISC.

### **Exploratory Data Analysis (EDA)**

CHEO formalizes exploratory data analysis as a data transformation process, providing a structured approach to analyzing datasets by producing statistical outputs. The ontology incorporates concepts from existing biomedical and statistical ontologies, including OBI, OBCS and STATO, while introducing extensions to support exploratory data analysis (EDA) workflows. The conceptual workflow is defined as follows:

* **Exploratory** Data Analysis (subclass of Data Transformation Process): A process that accepts a dataset as input and generates statistical outputs.
* Statistical Variable (CHEO:DIE): A study-defined variable analyzed within a statistical data analysis process, which may be categorical or continuous.
* Data item (CHEO:ICE): an observed value for specific patient
* Data set (CHEO:ICE): An aggregated of all data items
* Data Transformation Process: A parent class encompassing statistical data analysis, data distribution, and data visualization, ensuring that each statistical output is logically associated with the dataset and its variables.

EDA is a critical component of clinical data analysis, providing descriptive statistical insights into study variables. CHEO supports EDA profiling for categorical and continuous variables, extending existing OBCS and STATO ontology classes to enhance representation. For categorical variables, CHEO extends OBCS ontology classes to capture statistical measures including the frequency of each unique category within the dataset., Most Frequent Categorical Value (Mode) – The most commonly occurring category, Missing Value Count and Percentage, The number and proportion of missing values, Chi-Square Test Statistic, bar chart (x-axis, y-axis) as shown in Fig,

As depicted in Fig , For continuous variables, CHEO extends OBCS ontology classes to capture key statistical measures, including: Total Data Count – The number of observations for a given variable, Measures of Central Tendency, Mean (Arithmetic Average), Median (Middle Value of the Distribution), Mode (Most Frequently Occurring Value), Dispersion Measures, Standard Deviation, Variance, Range (Difference Between Minimum and Maximum Values), Interquartile Range (IQR) [Q1, Q3], Skewness and Kurtosis – Measures of the asymmetry and peaked-ness of the distribution, Outlier Identification (z-score, IQR).

* 1. **Mapping Data Element Categories to OMOP Domains**

|  |  |  |
| --- | --- | --- |
| **Category** | **OMOP Domain** | **Comments** |
| **Etiology** | **Observation** | Etiology” often represents the cause or contributing factor of a condition. It is not itself a disease entity but rather a characteristic or fact about a condition. Hence it is typically stored in the **Observation** table |
| **Measurement**   * Stages and scales * Functional and Behavioral Assessment Score * Blood measurement * Cardiac Measurement * Anthropometric Measurement * Consumption Measurement * Time to Event Measurement | **Measurement or Visit** | All of these represent quantifiable or semi-quantifiable clinical measurements, test results, or assessment instruments, which align with the **Measurement** domain in OMOP. If representing an actual inpatient encounter detail in the data system, it should be recorded in **Visit** |
| **Clinical Finding**   * Disease or Disorder Finding | **Condition Occurance** | Diseases and disorders (e.g., diagnoses) map to **Condition Occurrence** because they represent conditions experienced by the patient. |
| **Clinical Finding**   * Functional Finding * Compliance Finding | **Observation** | These “findings” are generally not formal diagnoses. They represent patient states or observations (e.g., functional status, compliance with a treatment) and are best stored in the **Observation** domain |
| **Death** | **Death** | OMOP provides a dedicated **Death** table for capturing the date and (optionally) cause of death |
| **Demographics** | **Person** | Demographic attributes (e.g., gender, year of birth) belong in the **Person** table, which captures core patient-level attributes. |
| **Risk Factors** | **Observation or Measurement** | If recorded as qualitative or categorical risk factors (e.g., “smoking status,” “family history of diabetes”), **Observation** is appropriate. If they are quantitative risk scores (e.g., Framingham risk score), use **Measurement** |
| **Sign or symptom**   * General sign * Cardiac Sign * Respiratory Sign * Neurological Sign * Vital Sign * Symptom | **Condition Occurrence or Measurement** | In OMOP, signs and symptoms are typically treated as conditions (e.g., “chest pain,” “fever”), thus placed in **Condition Occurrence**. Note that measured vitals like blood pressure or heart rate go to **Measurement**, but the presence of a symptom is a Condition |
| **Treatment**   * Exposure   + Device Exposure | **Device Exposure** | When the treatment involves medical devices (e.g., implants, instruments), use **Device Exposure** to capture these data |
| **Treatment**   * Exposure   + Medication Exposure | **Drug Exposure** | Medication usage or prescriptions belong in the **Drug Exposure** table |
| **Treatment**   * + Procedure | **Procedure** | All performed procedures, surgeries, or interventions (e.g., “appendectomy,” “MRI scan”) belong in the **Procedure** table. |
| **Follow up Attrition** | **Observation or Visit** | Follow-up attrition” is not a standard clinical event; it often reflects a study or administrative detail (e.g., loss to follow-up). You can store it as an **Observation** if you need to record it as a fact about the person, or potentially use **Visit** if tied to an encounter end date |
| **Medical History**   * Family History * Disease or Disorder History * Hospitalization History * Medication History * Procedure History | **Observation** | Historical data about conditions, procedures, or medications is typically captured in the **Observation** table if it is not actively ongoing at the time of data capture. (Some groups use Condition Occurrence for past conditions but they must carefully set start/end dates.) |
| **Dietary Intake** | **Measurement** | Dietary intake often represents measured or surveyed consumption (e.g., calorie count, nutrient intake), so **Measurement** is appropriate. If purely qualitative (e.g., “patient follows vegetarian diet”), it could go into **Observation** instead |
| **Number of Occurrences** | **Observation or Visit** | If representing an actual inpatient encounter in the data system, it should be recorded in **Visit** |

A diagram of a person's work flow

AI-generated content may be incorrect.

Figure 3 Patient-Level Data Modelling in CHEO

* 1. **CHEO Harmonization Specification Model**

To integrate **harmonization specifications** into CHEO, we define a **Plan Specification** that:

* **Evaluates each data element** from x number of studies.
* Assess the study design difference.
* **Assigns a harmonization score**.
* **Indicates required transformations**.
* **Identifies potential harmonization matches**.
* **Classifies harmonization difficulty**.
* **Provides a report explaining challenges** for difficult data elements.

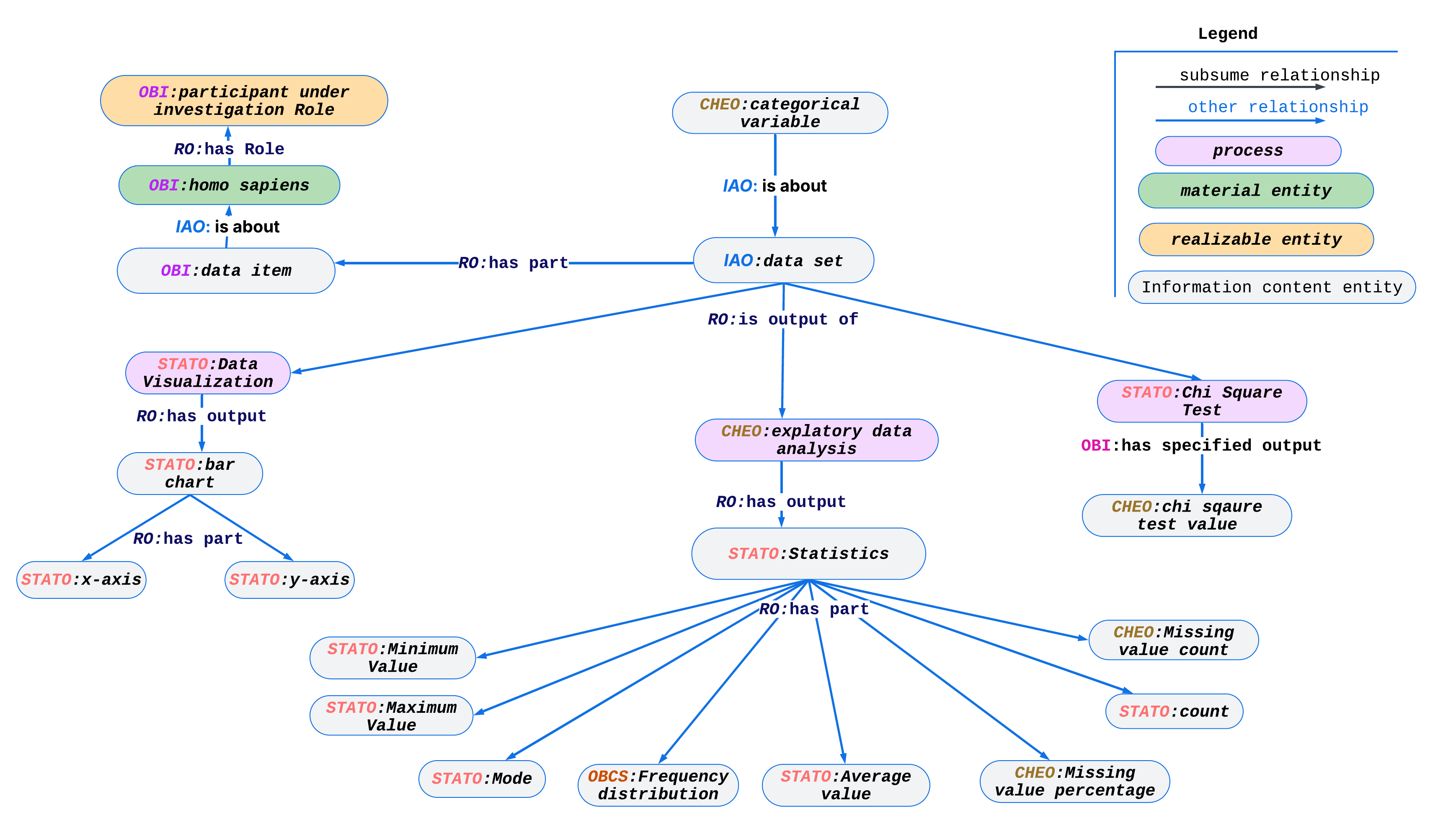


Figure 4 Categorical Data Element Statistical Modelling in CHEO

A harmonization plan is a directive information entity (obcs:PlanSpecification) that formalizes how a data element should be harmonized across studies. Each **Data Element** will have a corresponding **Harmonization Plan**.

**13. Future Steps and Next Iterations**

* **Refine the Ontology**: Add detail or constraints (SHACL) for data validation (e.g., permissible units, coded value sets).
* **Align with External Standards**: E.g., map or import classes from upper level ontologies for alignment.
* **Pilot Implementation**: Use the ontology in an actual data analysis pipeline, gather feedback from end users, and iterate.

**Conclusion**

The development of **CHEO** represents a significant step toward achieving **semantic harmonization** in clinical research. By providing a **unified ontological framework**, CHEO ensures that study metadata, patient-level data, and statistical methodologies are represented in a structured, interoperable, and reusable manner. Key contributions of CHEO include:

1. **Standardized metadata representation** – enabling clear distinctions between data element definitions, observed data, and analytical models.
2. **Enhanced interoperability** – facilitating mapping between local dataset structures and standardized terminologies such as **LOINC, SNOMED, and OMOP**.
3. **Support for automated reasoning and advanced queries** – allowing researchers to perform complex data retrieval tasks (e.g., "Find patients with NYHA Class ≥ III and EF < 40%").
4. **Improved data harmonization for multi-study integration** – introducing harmonization scoring mechanisms, transformation requirements, and integration challenges for cohort comparability assessments.
5. **Scalability for multi-center studies** – enabling structured representation of clinical trials, patient registries, and epidemiological datasets with standardized reporting and statistical modeling.

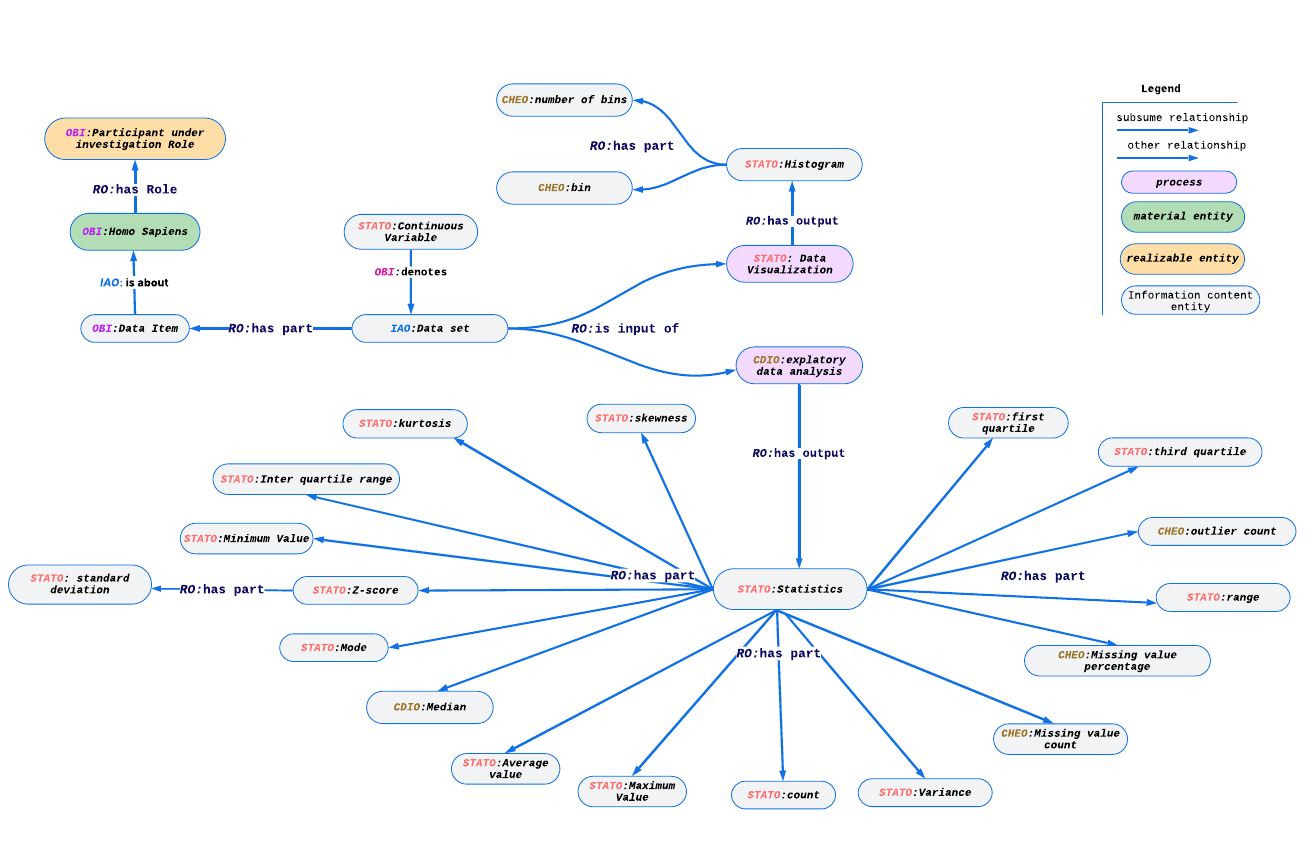


Figure 5 Continuous Data Element Statistical Modelling in CHEO

By aligning with **established ontologies** and adopting best practices in **metadata governance (ISO 11179, CDISC)**, CHEO fosters a **reliable and scalable ecosystem** for clinical research. The **next phase** of this project involves:

* **Refining data validation constraints** (e.g., SHACL-based quality checks),
* **Aligning with regulatory standards** (e.g., CDISC SDTM, Define.xml),
* **Testing in real-world data environments** to assess feasibility and impact.

Ultimately, CHEO aspires to become a **widely adopted resource** for **ontology-based clinical data standardization**, enhancing the **reusability, interoperability, and analytical potential** of clinical datasets across research networks. Through continuous iteration and community engagement, CHEO will contribute to a **more cohesive, transparent, and scalable approach** to clinical data integration and analysis.

Additional (Backup) about data element and phenomena it is about.

