

Golden Table – Supporting Literature Review and Interview Insights
May 2025

Document Introduction

MassHealth developed a Golden Table that stores a single member record for each Race, Ethnicity, Language, Disability, Sexual Orientation, and Gender Identity (RELD SOGI) element (for each MassHealth member) based on a set of business rules that help determine the “best” available value(s) across 90 select data sources in the MassHealth Data Warehouse (MH DW).

Prior to this development effort, MassHealth underwent an effort to better understand the current landscape of RELD SOGI data hierarchies and identify leading practices for building a MassHealth-specific RELD SOGI data hierarchy.

This document is intended to compile key insights from that effort as part of MassHealth’s commitment to transparency on how the Golden Table was made, how it works, and how it will be used. For questions on this document or other Golden Table related question, please email Health.Equity@mass.gov.

Approach

MassHealth followed a three-step approach to gain insight into the current landscape of RELD SOGI data hierarchies:

1. **Conduct external interviews:** MassHealth facilitated conversations with key external parties and subject matter experts focused on gathering common practices and lessons learned from their experiences in data analytics and hierarchy development. Interviews focused on RELD SOGI data collection and data flows, approaches to data cleansing and standardization, and considerations for developing a data hierarchy. MassHealth also reviewed documentation shared by interviewees including information on RELD SOGI data imputation, potential approaches to source prioritization logic, and RELD SOGI data mapping.
2. **Complete literature review:** MassHealth conducted a literature review, analyzing 40 publicly available references, to build a foundational understanding of existing data management strategies and source prioritization recommendations. Sources included white papers, federal guidelines on RELD SOGI data collection, and articles that broadly explore the construction of data hierarchies, extending beyond frameworks focused on RELD SOGI data.

Note: MassHealth discovered that the concept of a RELD SOGI data hierarchy is an emerging one. Therefore, literature review insights focus on leading practices for designing and building data hierarchies at a general level.

3. **Facilitate internal exploration meetings:** MassHealth facilitated internal exploration meetings with the Massachusetts Executive Office of Health and Human Services (EOHHS) technical teams to identify a high-level approach to building the data hierarchy, to discuss Master Person Index (MPI) strategies and Master Data Management (MDM) tools and conduct deep dives into potential tools for the Golden Table solution.

The following sections include insights from each of these three steps.

Interview Insights

MassHealth conducted external interviews with five key parties outside of MassHealth (as shown in Table 1). These interviews have been deidentified to protect the privacy of participants.

Key Party	Interviewee Titles
State Agency #1	<i>Race, Ethnicity, Language and Disability (REALD) and SOGI Director</i>
State Agency #1	<i>REALD and SOGI Data Analytics Manager</i>
State Agency #2	<i>Health Analytics & Informatics Administrator</i>
State Agency #3	<i>Director of the Bureau of Social Care and Community Supports</i>
State Agency #3	<i>Deputy Director</i>
State Agency #3	<i>Director of Strategic Operations and Planning</i>
State Agency #3	<i>Medicaid Redesign Analyst</i>
Research Institute	<i>Deputy Director</i>
Research Institute	<i>Senior Research Fellow</i>
Sister Agency	<i>Director, Special Analytic Projects</i>
Sister Agency	<i>Epidemiologist II</i>

Table 1: List of key parties and interviewees from the data hierarchy external interviews

Six key insights emerged across interviews. These insights are described below.

Interview Insight #1: Most RELD SOGI data hierarchies are in their infancy and should be considered highly iterative.

1. Interviewees noted that MassHealth is among the pioneering Medicaid agencies to develop a RELD SOGI data hierarchy, with only one other state agency indicating that they have begun working on this concept at a Statewide level. The Agency noted that it is still a work in progress — particularly for SOGI data.
2. Other agencies indicated their interest in building a data hierarchy and noted that, through trial and error of certain concepts, they have identified more lessons learned than best practices. For example, one interviewee attempted to prioritize sources but did so exclusively on data concordance (i.e., how closely aligned sources were to others in the data set to determine which were most likely to be accurate) to establish its source ranking. Reflecting on that approach, the team identified other factors such as recency and trust in sources that will be considered in future iterations of its hierarchy.
3. Across all agencies that have been designing, developing, or preparing to create data hierarchies, there was a consistent sentiment that developing a RELD SOGI data hierarchy should be an iterative process — one that should be consistently updated over time. Data collection practices improve each year, and an individual's RELD SOGI data may change over time.

Interview Insight #2: Data hierarchies should prioritize sources that are closest to the individual directly reporting the data.

1. RELD SOGI data is collected from a wide variety of sources, including distinct Eligibility & Enrollment systems, Acute Hospitals, MCEs, and CBHCs. This data is collected in multiple methods and settings, including patient self-reporting on forms, collection by medical staff, or through enrollment applications completed by the head of household. Due to these differences,

certain sources may reflect a member's identity with a higher level of confidence based on how the data was collected.

- a. For example, one interviewee noted that a teen's SOGI data reported by their doctor may have a higher level of confidence than enrollment system data where a head of household can report data on behalf of all household members.
 - b. In another instance, an interviewee indicated that when it first began exploring an integrated data set, race was sometimes assumed in historical data rather than self-reported, raising concerns about data quality.
2. Prioritizing data sources that do not transform self-reported data may be key to maintaining data integrity and mitigating potential risks such as member misrepresentation, especially for SOGI values that are more likely to change over time and may be more sensitive for members to share.

Interview Insight #3: Data snapshots should be captured routinely for resulting values of a RELD SOGI data hierarchy as RELD SOGI data can change over time and data collection practices evolve.

1. A member's responses may change over time. It is important to incorporate business rules to account for historical data and updated responses to allow for data and population changes to be identified over time.
2. One interviewee emphasized the importance of retaining historical data since collection methods and cultural competency trainings are still evolving and could influence how a member responds to a question.
3. Others indicated the need for data snapshots to provide reports that capture data over time, noting that they have received data snapshot requests for a particular point in time or for multiple years to help identify population health insights.
4. An agency identified that its system's ability to capture data has evolved significantly in the last decade (e.g., the ability to select multiple responses for each category), highlighting the importance of revisiting a data hierarchy as the data improves.

Interview Insight #4: The structure and design of a RELD SOGI data hierarchy is highly dependent on use cases, emphasizing the need to document design decisions and processes.

1. Data needs may vary depending on use case (e.g., reporting requirements may ask for one or multiple values which may require a different interpretation of the collected data). The ability to retain and track a robust set of historical and current values as well as the logic for selecting them can support functionality across evolving use cases and business needs.
2. Additionally, due to the iterative nature of RELD SOGI data, keeping documentation updated is critical for documenting and sharing best practices and lessons learned. For example, one agency maintains robust documentation about its RELD SOGI Golden Table design and hierarchy approach, so that any updates or design decisions can be clearly tracked.

Interview Insight #5: Data sharing can be implemented to support broader organizational and community goals, while upholding data privacy and security standards.

1. Given the sensitivities of RELD SOGI data, sharing and transparency options must be considered against access, permission, and security criteria.
2. One interviewee noted that community groups have expressed interest in accessing demographic data to assess population health, but this has been difficult to do in a meaningful way. Data has been shared in files that are too broad (i.e., categories are displayed at too high a level) or too granular (i.e., there is so much information that only a data analyst could evaluate it) to effectively use. One takeaway shared was that states can consider alternative methods of sharing the outputs (e.g., dashboard or other visual formats) to increase accessibility and transparency while also maintaining member privacy.

Interview Insight #6: Transparency in how hierarchy logic is developed and applied can foster external engagement which can lead to a more robust and thorough hierarchy.

1. To support community involvement and transparency, certain interviewees have conducted community outreach sessions to increase understanding of how RELD SOGI data can be used to analyze population health. While they do not provide data access to these groups, they offer a platform for understanding the impacts that a data hierarchy may have on member outcomes.

Literature Review Insights

In addition to external interviews, MassHealth conducted a literature review of 40 references to identify leading practices and approaches to developing a data hierarchy. References are captured in [Table 3](#) below and used to cite throughout the literature review. Eight key insights from the literature review are presented below — many of which align with the interview insights outlined in the previous section. Importantly, given the cross-sectional nature of this approach, the need to both evaluate hierarchical approaches and the intersection with RELD SOGI implementation, the literature review aimed to evaluate both methods and qualitative practices. Additionally, best practices for reference and master data management were reviewed and incorporated into the following insights to provide a generalized approach to the master data management (MDM) concepts that will be applicable to these guiding principles and provide a foundational MDM approach.

Methodology

Review of established data management best practices was coupled with an analysis of current literature related to data hierarchy management and RELD SOGI specific data management to produce the literature insights. The Data Management Body of Knowledge (DAMA DMBOK) serves as foundational guide for data management practices¹⁰. It is a comprehensive framework developed by the Data Management Association (DAMA) International to provide a standardized approach to data management practices. Established to address the growing need for structured and effective data management in organizations, DAMA DMBOK serves as a guide for data professionals, offering best practices, principles, and terminology. It was created through the collaborative efforts of industry experts and practitioners who recognized the importance of a unified approach to managing data as a critical asset. The framework is widely used by organizations to enhance their data management capabilities, ensuring data quality, governance, and security. Key concepts covered in DAMA DMBOK include data governance, data architecture, data modeling and design, data storage and operations, data security, data integration and interoperability, document and content management, reference and master data management, data warehousing and business intelligence, metadata management, and

data quality management. For development of literature insights and guiding principles the reference and master data management sections of DMBOK were used as foundational principles, along with recent advancements in master data management accounting for additional MDM components such as factors that influence improvement of master data quality^{4,33}, master data management lifecycle and support^{18,31}, assessing and advancing master data management maturity³², and the specific implications that the inclusion of RELD SOGI data may have on the MDM process^{16,17,22}.

Master Data Management deals with the creation, maintenance, and use of a single, authoritative source of master data for core business entities such as customers, products, employees, and suppliers¹⁰. This process involves data consolidation, data quality management, and data stewardship to create and manage master data that is accurate, consistent, and available across various systems and business processes. The key concepts in MDM as outlined in DAMA DMBOK include data governance, data stewardship, data quality, data integration, and data lifecycle management and are detailed below.

1. **Data Governance:**

- **Establish Clear Ownership and Accountability:** Define roles and responsibilities for data governance, including data stewards and data owners, to ensure accountability for data quality and consistency.
- **Develop Data Policies and Standards:** Create and enforce policies, standards, and procedures for data management to ensure uniformity and compliance across the organization.

2. **Data Stewardship:**

- **Define Data Stewardship Roles:** Clearly define the roles and responsibilities of data stewards, ensuring they have the authority and resources needed to manage master data effectively.
- **Assign Data Stewards to Key Domains:** Allocate data stewards to specific data domains or subject areas to ensure focused and specialized oversight as needed.
- **Provide Training and Support:** Offer ongoing training and support to data stewards to keep them informed about best practices, tools, and methodologies.
- **Establish Stewardship Processes:** Develop and document processes for data stewardship activities, including data quality monitoring, issue resolution, and data lifecycle management.
- **Foster Collaboration:** Encourage collaboration between data stewards, data owners, and other key parties to ensure a cohesive approach to master data management.
- **Monitor Stewardship Performance:** Regularly review and assess the performance of data stewards, providing feedback and making adjustments as necessary to improve effectiveness.

3. **Data Quality Management:**

- **Implement Data Quality Metrics:** Define and monitor key performance indicators (KPIs) for data quality, such as accuracy, completeness, consistency, and timeliness.
- **Conduct Regular Data Audits:** Perform periodic data quality assessments and audits to identify and rectify data issues.

4. Data Integration and Consolidation:

- **Centralize Master Data:** Consolidate master data from disparate sources into a single, authoritative repository to ensure consistency and reduce redundancy.
- **Use Data Integration Tools:** Leverage data integration technologies and tools to facilitate the seamless flow of master data across different systems and applications.

5. Data Lifecycle Management:

- **Implement Data Lifecycle Policies:** Establish policies for the creation, maintenance, archiving, and deletion of master data to ensure its accuracy and relevance over time.
- **Automate Data Processes:** Use automation to manage the lifecycle of master data, including data validation, enrichment, and synchronization.

In addition to the review of industry best practices from leaders such as DAMA and Gartner, a current review of continuous MDM influencing factors, and because the concept of developing data hierarchies for RELD SOGI data has been elevated in recent years, MassHealth used online search engines (Google) and two databases (Google Scholar, PubMed) to identify relevant research and leading practices. Search terms used are outlined in Table 2 below. A variety of search terms and combinations were used given the limited amount of literature available that is specific to RELD SOGI data hierarchies and associated leading practices. These literature reviews in conjunction with the review and incorporation of industry leading MDM practices were synthesized to create the insights below.

RELD SOGI Data Hierarchy Literature Review Search Terms
• Master Data Management review • Master Data Management quality • Master Data Management data lifecycle • Race AND OR Ethnicity in Master Data Management • CDC and data hierarchy and health equity • CMS and data hierarchy and health equity • "Data hierarchy" • "Data hierarchy" AND race OR ethnicity OR disability OR language OR "sexual orientation" OR "gender identity" • Data hierarchy AND race OR ethnicity • Data hierarchy and health equity • "Data source hierarchy" • "Data source hierarchy" AND race OR ethnicity OR disability OR language OR "sexual orientation" OR "gender identity" • Data taxonomy AND race OR language OR disability • Electronic Health Record source prioritization • Health data prioritization

- Health data prioritization AND race OR ethnicity
- "Hierarchical data logic"
- "Hierarchical data logic" AND race OR ethnicity
- "Hierarchical data modeling" AND race data
- "Hierarchical data modeling" AND race OR language OR disability
- "Mastering logic" AND race OR language OR disability

Table 2: Summary of search terms used in the literature review. Search terms were entered in Google, PubMed, and Google Scholar.

Literature Review Insight #1: Most publicly available literature on RELD SOGI data concentrates on data collection and standardization.

1. There is limited publicly available information regarding established industry practices, potential approaches, or lessons learned from developing a RELD SOGI data hierarchy. However, the importance of improving capture and standardization of this data has been noted^{17,22} as have the impacts the lack of collection and standardization of this data when used for analysis¹⁶.
2. Federal agencies such as CMS and the Office of Minority Health (OMH) have not released specific guidance on building a RELD SOGI data hierarchy. Existing literature and guidelines focus on data collection guidelines and methodology^{11,23-25,27,34,35,38,40}. However, these references can still provide guidance for the creation for data management and hierarchy approaches to support development of a data source hierarchy. A report jointly published by Grantmakers in Health and NCQA outlines the different levers that the federal government can take to improve race and ethnicity data, one of which is standardizing the collection and reporting of racial, ethnic, and other demographic data across the federal government (e.g., Department of Agriculture, CMS, Health Resources and Services Administration reporting requirements) while “providing states, local governments, and grantees the flexibility to collect data on other population groups residing in their area”¹¹. These recommendations for data collection can be used to guide evaluation and prioritization of data sources providing RELD SOGI data.
3. MassHealth’s approach to collecting RELD SOGI data (data collection practices, RELD SOGI data standards) is based on federal guidelines and industry-leading practices. Federal guidelines include: [U.S. Department of Health & Human Services](#) and [CMS](#); additionally, the [Massachusetts Quality Measurement Alignment Taskforce Health Equity Data Standards](#) are informed by OMB, CDC, and the U.S. Census Bureau.
4. A report released by SHADAC in October 2020, [Exploring Strategies to Fill Gaps in Medicaid Race, Ethnicity, and Language Data](#), outlines the most commonly cited data collection challenges for race, ethnicity, and language. These data collection challenges include: lack of mandatory reporting standards; rapidly changing demographics; evolution in how people self-identify; voluntary reporting; lack of understanding on why the data are important; and mistrust and enrollee concerns about how data will be used. These data collection challenges impact states’ ability to collect complete, high quality REL data consistently³⁶.
5. The National Committee for Quality Assurance’s (NCQA) publication, [Current Health Plan Approaches to Race and Ethnicity Data](#), outlines key challenges highlighted by health plans to data completeness for race and ethnicity data. These challenges include: categories for

documenting and reporting race and ethnicity; rationale for data collection; access to employer data; opportunities to improve government enrollment data; consent/transparency of data use; source of truth where there are multiple data sources; frequency of data collection and verification; and imputation uses and methods²⁴.

6. A report released by SHADAC in March 2024, [Sexual Orientation and Gender Identity Data: New and Updated Information on Federal Guidance and Medicaid Data Collection Practices](#), noted that “very few state Medicaid agencies are currently collecting [SOGI data],” underscoring MassHealth’s position as a leader in this area and the nascency of building a RELD SOGI data hierarchy³⁷. Additionally, the variability rates in RELD compared to SOGI data (and cumulative) are still being evaluated. The rates at which individuals vary their response for RELD questions compared to SOGI questions is not well established but may lead to differing rates of discrepancy for these values across sources.

Literature Review Insight #2: While RELD SOGI data can be used to identify and reduce health inequities, there are challenges and important considerations when collecting complete and accurate RELD SOGI data.

1. Organizations can use health equity data to identify health inequities and use insights from the data to inform policy or programmatic decisions^{11,12,39}. However, there are challenges presented in the collection of complete and accurate RELD SOGI data. These challenges can include:
 - a. Legal and privacy concerns¹²: A major barrier to collecting race and ethnicity data in health insurance plans is the belief that it is restricted or prohibited. The 2004 National Research Council report identified four states with such restrictions, but another legal analysis indicated that data collection is generally allowed under federal and most state laws. Despite the age of these reports, similar concerns were raised in a 2021 open letter to the National Association of Insurance Commissioners, citing potential liabilities and privacy issues. Additionally, there are worries about the consequences of identifying health disparities and the costs of ensuring data privacy.
 - b. Lack of standardized RELD SOGI data standards and data collection practices: See *Literature Review Insights 1.4, 1.5, 1.6*. Additionally, it has been shown that race and ethnicity data standards vary dramatically across different types of health insurance, which results in wide variation in race and ethnicity data completeness across health plan coverage types (e.g., public vs. private) and lines of business (e.g., plan type)^{12,39}.
 - c. Community trust providing RELD SOGI data^{12,39}: Individuals that identify with historically marginalized groups may be concerned that providing RELD SOGI data could have adverse impacts such as discrimination, misuse of RELD SOGI data, or that providing the data could impact the types of services provided. Due to historical and ongoing discriminatory treatment of vulnerable patients by healthcare institutions, race and ethnicity data collection procedures must emphasize the reasons for data collection, ensure patients' privacy protections, and incorporate other trust-building measures.
 - d. Technological readiness and interoperability concerns¹²: Improving health equity data faces challenges with data compatibility, interoperability, and sharing. Providers and employers often hesitate to share detailed race and ethnicity data due to privacy

concerns. Technical interoperability issues exist between EHRs and administrative data across different systems. Standards like HL7® are not universally used, and there is a need to clarify data provenance and reconcile conflicting records. Standardization and "semantic interoperability" are essential for effective health data management.

- e. Sufficient staff and resources^{12,39}: Communicating the value of improved data collection and demonstrating how it can be used to help reduce disparities could support data collection efforts. However, adequate resources for data collection (technology, staff) are critical. Even with increased funding for data collection, smaller healthcare organizations, especially local public health agencies, may lack the staff and technology needed for effective data collection and analysis.
 - f. Cost of collection and lack of incentives¹²: Financial factors can hinder the collection of race and ethnicity data, including costs for updating processes, health IT investments, and staff retraining. State and local governments face significant expenses to upgrade systems, and clinical time may be diverted from billable activities. The lack of formal incentives or requirements may lead to de-prioritization of data collection. While there is support for greater incentives, designing them is challenging, and patients often lack incentives to share their data.
2. Varied data collection standards across agencies and systems: According to an [article by the Urban Institute](#), there are no standardized federal requirements for collecting RELD SOGI data, and as a result, an individual may encounter differing RELD SOGI questions and response options depending on the organization³⁹. A data hierarchy can help an organization reconcile multiple demographic data values for the same person. Thus, supporting the identification of a "best" value for an individual's RELD SOGI information amidst multiple sources of truth with varying collection methods and standards. As an organization is standardizing its data collection practices, the data hierarchy could prioritize sources that more closely align to the most current data collection standards.
3. Importance of self-reported data for health equity analytics: Self-reported demographic data is largely considered the "gold" standard across the industry^{3,5,9,14,24-26,36,37}. The implementation of a data hierarchy that uses self-reported data as its foundation can more closely reflect the ways that an individual identifies.
4. Given the challenges associated with standardized collection of RELD SOGI data, it is rare to have a single source with fully complete data. The literature supports different methods and use cases for integrating multiple data sources to create a more complete dataset^{1,22}.
 - a. Several research studies used data linkage to improve the quality of race and ethnicity data²². Data linkage is the process of connecting data from different sources to create a unified dataset²². For example, to improve missing race information for death records of American Indian and Alaskan Native individuals, researchers linked registration records from the Indian Health Services to data from the CDC's National Center for Health Statistics' (CDC-NCHS's) National Death Index (NDI) to identify deaths in this population²². On the other hand, Researchers at Kaiser Permanente Southern California, a large managed health-care system, validated race and ethnicity data from health

administrative records using race and ethnicity data from birth certificates available through their hospitals²².

- b. Data pooling is a method that can be used to support equitable approaches to collect and analyze race and ethnicity survey data; data pooling combines data from multiple surveys or data sources to create a more comprehensive dataset¹. While this method is specific to survey research, it is applicable to MassHealth because the Golden Table will include RELD SOGI data on the same individual from multiple sources to help create a more complete record.

Literature Review Insight #3: Leading data hierarchy practices suggest a variety of potential factors for MassHealth to consider when developing its RELD SOGI data hierarchy prioritization.

1. Source of truth when there are multiple data sources: NCQA conducted a series of key informant interviews with a sample of health plans representing different geographic regions and insurance markets (e.g., Medicare Advantage, Medicaid, commercial). Identifying the “source of truth” when there are multiple data sources was raised as a challenge that has more recently developed as a result of increased opportunities for data collection and sharing²⁴. This issue is particularly challenging since an individual’s response to a question may be dependent on who is asking the question, what the available response options are, and the setting that they are asked. NCQA provided the following guidance to help health plans prioritize data sources based on anticipated accuracy:
 - a. Use specific categories over non-specific categories
 - b. Use the most frequent or consistently reported category
 - c. Select data with clear provenance (source and method of collection) over data without clear provenance

Additionally, NCQA gathered feedback from an advisory panel, which emphasized the importance of establishing a process to identify a source of truth among multiple data sources²⁴. However, the panel also noted that current data collection and sharing methods present significant challenges to achieving this. Insights from two panel members include:

- d. One panel member prioritizes data they collected because they can confirm and validate all aspects of the data collection process
 - e. A different panel member emphasized the importance of engaging a diverse group of key parties when developing a process
2. Size of the datasets: The amount of data collected may inform how sources are prioritized and utilized²⁴. For example, the amount of data in an electronic health record (EHR) system compared to the amount of data collected by a small ACO may be drastically different. Larger datasets allow for a more statistically based review of the data and may include a larger population distribution with potentially less bias. Importantly, these considerations are most relevant at the population level. Statistically larger sample sizes are likely to contain a normal distribution and error rate, whereas smaller data samples have the propensity to be skewed at the population level. This is a relevant consideration when establishing source prioritization based on the probability of any one source providing a more unbiased information set relative to the population¹.

3. Data collection setting and modality: How a source collects data (e.g., self-reported, standardized input, free form) may influence source prioritization^{1,16,22}. For example, when standardized data collection aligns most closely with the business need, a data source that collects standardized data may be prioritized over sources that collect free-form data.
4. Data quality and reliability (veracity of the data source): The quality of data submitted by sources may also inform source prioritization^{4,10,33}. For example, a source reporting RELD SOGI data where the majority of fields are “unknown” or “unable to collect” may be prioritized lower than a source that reports substantive RELD SOGI values for most of its records. It is important to note that the criteria for determining data quality are closely tailored to an organization’s specific business needs.
5. Latency of the data source¹⁰: How frequently data is updated by a source is an important consideration, especially for RELD SOGI data that can change over time^{4,10,32}. Latency refers to how long it takes for data to be transmitted from a data collection source to the receiving entity (e.g., if an ACO only reports information to MassHealth once a year, there is a potentially a year-long latency in the data reported by that ACO). Sources that update data frequently may be prioritized higher than a source that updates its data less frequently.
6. Data level of detail: The level of detail submitted by a source (e.g., more descriptive, less descriptive) may inform how sources are prioritized based on the business need^{15,18}. For example, if a detailed understanding of ethnicity is desired, a data source that collects more detailed ethnicity information (such as providing 20 ethnicity fields for selection), as opposed to a data source that collects only Hispanic/Non-Hispanic may be prioritized.

Literature Review Insight #4: There are two popular methods for building a prioritized categorical data hierarchy: deductive (bottom-up) and inductive (top-down). Utilizing both approaches (inductive and deductive) in a hybrid approach can be an effective method to build a prioritization hierarchy.

1. A categorical data hierarchy is an organized structure that arranges data into different levels or categories, each with a specific relationship to the others²⁰. This hierarchy allows for the classification, categorization, prioritization, and sorting of information in a systematic way. At the top level, broad categories are defined, which are then subdivided into more specific subcategories at lower levels. This hierarchical structure facilitates the organization of data, making it easier to analyze and interpret by grouping similar items together and establishing clear relationships between different data points. By using a hierarchical approach, data can be more efficiently managed and prioritized²⁰. A hierarchical model is an effective approach for prioritizing information from multiple sources due to its organized structure, efficient data management, enhanced decision-making, improved accessibility and retrieval, and consistency and standardization. By providing a clear and simple structure, a hierarchical model allows for layered prioritization, making complex data more understandable and navigable²⁰. It supports scalability and reduces redundancy by grouping similar data points together. Decision-makers benefit from focused analysis and contextual understanding. In general, a hierarchical model is a powerful tool for managing and prioritizing information to support informed decision-making.
 - a. Deductive approach²⁰: When building a categorical data hierarchy using the deductive method, all the values in a system are reviewed by end users or subject matter experts

who identify which values can represent the broadest category. The remaining values are allocated to the categories based on the relationships according to the use case. It is important to maintain a consistent level of specificity within a category.

- i. An example of this approach would be to first group sources for prioritization based upon a set of common features. For example, grouping provider sources into categories based upon reporting frequency, size of the population served, compliance with reporting standards, etc. This type of grouping then allows further prioritization and categorization within groups as needed. This may be helpful when defining source prioritization for RELD/SOGI reporting as it categorizes sources to enable standardized guidelines or rules for each category of data providers.
- b. Inductive approach²⁰: The inductive method of building a hierarchy is used when there is not access to historical data or content (e.g., building a taxonomy for a new website). This method is recommended in instances where the kind of information that will be indexed is unknown. It is a continuous and iterative approach, where new terms are added to the hierarchy as they are encountered. This concept also informs data source prioritization, especially when considering available terms (or selections) for individuals to select.
 - i. When considering RELD SOGI data and information, the inductive approach is likely to be encountered as new fields or definitions are adopted.
 - ii. One example for Massachusetts is the inclusion of “Cape Verdean” as an ethnicity category. While this ethnicity may not be defined initially, its inclusion as a category in the hierarchy would be considered through the inductive approach.

Literature Review Insight #5: The benefits of using a data hierarchy for RELD SOGI data outweigh potential challenges given MassHealth’s business needs.

1. Benefits to using a hierarchal data model include²:
 - a. Hierarchical models are predictable since the relationship between values is documented using tree-like architecture.
 - b. Hierarchal models can support data integrity via the parent-child relationship across values, as data remains consistent and there is a lower likelihood of redundant data.
 - c. Hierarchical models work well for large datasets since they are capable of quickly retrieving data.
 - d. Hierarchal models generally are more secure, as defined access paths make unauthorized data access more challenging. Data access for authorized users is limited to specific data or data paths.
 - e. Hierarchal models are scalable — a hierarchal model can be updated (e.g., remove data values and/or logic) more easily than other data management strategies.
2. Challenges to using a hierarchal data model include²:

- a. There is limited flexibility with a hierarchal data model, and not all types of data are suited to linear relationships.
- b. Data relationships can be highly complex, especially in systems with many-to-many data relationships. Developing a hierarchal, linear logic can be especially challenging since there can be replication of data across multiple sources.
- c. It may be challenging to modify a hierarchal database structure when data requirements or business needs shift.

Literature Review Insight #6: Data hierarchies are iterative and should be updated over time as new federal guidelines are released, data specifications change, and business needs evolve.

1. Data hierarchies are not static and should be periodically updated to align with revised policies or federal guidelines, as well as to accommodate new use cases and evolving business needs^{19,20,21}. The [Office of Management and Budget \(OMB\) recently updated its standards](#) for collecting and presenting federal data on race and ethnicity. These revisions enhance the ability to compare information and data across federal agencies and to assess how effectively federal programs serve diverse populations. The OMB is committed to ongoing research and regular reviews to keep its standards up to date with shifting demographic and evolutions in how individuals identify. As a result, the software used to implement a data hierarchy should be flexible to support hierarchy updates (e.g., ingest new data sources, incorporate new data values, update logic).
2. Guidance released by CMS in May 2024, [Resource of Health Equity-related Data Definitions, Standards, & Stratification Practices](#), specifically notes that “standards continuously evolve and may be subject to change,” and that CMS guidance will be updated as applicable⁹. As an industry leader in RELD SOGI data collection, many of MassHealth’s current data collection practices (e.g., prioritizing self-reported data, treating sexual orientation and gender identity as distinct and independent values) already align with guidance included in the document.
3. The way in which individuals identify may shift and evolve over time. A [report](#) released by NORC emphasizes that as racial and ethnic identities evolve, the methodologies used by researchers to capture and collect race and ethnicity data must also adapt and progress¹. This highlights the importance of continuously evolving and refining the data hierarchy to ensure accuracy and relevance.

Literature Review Insight #7: There are six key operational considerations for data governance to support prioritization hierarchy maintenance and operations.

1. Addition of new values¹⁹: New values or entities within the hierarchy, in this case potential RELD SOGI data providers, should not be added to the hierarchy without a formal approval or review process, since overlapping values or logic can impact the prioritization and logic flow of the hierarchy. When a new value or data is added to the hierarchy, standard business processes should be followed to archive and re-index data as needed.
2. Removal of values¹⁹: Existing values or data elements should not be removed from the hierarchy without a formal approval or review process, and the rationale for removal and potential impacts should be documented.

3. Re-indexing¹⁹: Whenever a value or data element is added or removed from the hierarchy, the logic and values should be re-indexed as appropriate to maintain the prioritized hierarchy.
4. Review process (governance)¹⁹: It is important to establish a formal governance and review process for any proposed changes or updates to the data hierarchy. Best practices include:
 - a. Maintaining historical documentation of all changes to the data hierarchy, including the date of change and any cascading impacts from the update.
 - b. Conducting an annual review of the entire hierarchy to review the integrity of the hierarchy and confirm that business needs are still being met.
 - c. Developing a communication strategy to communicate changes to key parties (e.g., Enterprise Data Management & Engineering (EDME) team, Data Integrity team, MCEs).
 - d. Establishing a formal data hierarchy governance process to support implementation and operation of the data hierarchy.

Literature Review Insight #8: Developing the logic for a data hierarchy (e.g., determining which sources are the most “trustworthy” or determining which RELD SOGI value should be selected) is highly individualized based on organization, its objectives, and data collection processes.

1. In 2023, researchers conducted a [study](#) to develop a common data model to support sickle cell disease (SCD) surveillance. Population-level data on SCD is limited and there was not a standard process across states¹⁵. Researchers identified a series of variables such as birth year, sex, race, and ethnicity, and developed a hierarchy based on the most valid data source for each variable. In this study, newborn screening records are the primary source in the hierarchy, followed by birth certificates¹⁵. This hierarchy was informed by the literature on SCD and “expert input,” illustrating the highly individualized nature of hierarchy development.
2. The Oregon Health Authority (OHA) released a [summary of its findings from a data quality assessment](#) centered on data collection standards for race, ethnicity, spoken and written language, and disability status (REALD)³⁰. While the study’s findings did not directly lead to the creation of a data hierarchy, they demonstrate the substantial customization required when formulating the logic for such a hierarchy, as the evaluation focused on areas such as member response rate to REALD questions and concordance (e.g., to what extent are REALD responses consistent across sources).

Literature review references

Source ID	Citation
1	Alyssa Ghirardelli, et al., Advancing Equity in Race & Ethnicity Data in Population Surveys: Findings from Expert Interviews (2024)
2	Anina Ot, What Is A Hierarchical Data Model? Definition And Examples (2023)
3	ASPE, HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language, and Disability Status (2011)
4	Azira Ibrahim, et al., Factors Influencing Master Data Quality: A Systematic Review (2022)
5	BlueCross BlueShield Association, The Ethical and Transparent use of Data to Reduce Health Disparities (2022)

6*	CDC Foundation Health Equity Strategy Office, Principles for Using Public Health Data to Drive Equity (n.d.)
7*	Centers for Medicare & Medicaid Services (CMS), CMS Framework for Health Equity 2022-2032
8*	CMS, New SOGI Questions on the Marketplace Application (n.d.)
9	CMS, Resource of health equity-related data definitions, standards, and stratification practices (2024)
10	David Rethwisch, DAMA DMBOK: Data Management Body of Knowledge, Second Edition (2017)
11	Grantmakers In Health (GIH) and the National Committee for Quality Assurance (NCQA), Federal Action Is Needed to Improve Race and Ethnicity Data in Health Programs (2021)
12	Grantmakers In Health (GIH) and the National Committee for Quality Assurance (NCQA), Improving Data on Race and Ethnicity: A Roadmap to Measure and Advance Health Equity (2021)
13*	Hawaii Health Data Warehouse, 2023 Race/Ethnicity Documentation (2023)
14	Institute of Medicine, Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement (2009)
15	JAMIA, Common data model for sickle cell disease surveillance: considerations and implications (2023)
16	Kevin Nead, et al., Cautions When Using Race and Ethnicity in Administrative Claims Data Sets (2022)
17	Mandi Bishop, Use Social Determinants of Health Analytics to Inform Health Equity Strategy (2022)
18	Martin Ofner, et al., Management of the Master Data Lifecycle: A Framework for Analysis (2013)
19	Mohit Bhakuni, Guidelines for Taxonomy Maintenance (2020)
20	Mohit Bhakuni, How to build your organization's taxonomy — best practices (2020)
21	Mohit Bhakuni, Why is it Difficult to Build a Taxonomy? (2020)
22	Monica Ter-Minassian, Improving data capture of race and ethnicity for the Food and Drug Administration Sentinel database: a narrative review (2023)
23	National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Committee on Measuring Sex, Gender Identity, and Sexual Orientation, Measuring Sex, Gender Identity, and Sexual Orientation (2022)
24	National Committee for Quality Assurance, Current Health Plan Approaches to Race and Ethnicity Data Collection and Recommendations for Future Improvements (2023)
25	National Science and Technology Council, Federal Evidence Agenda on LGBTQI+ Equity (2023)
26	NORC, Data Collection in Medicaid to Advance Health Equity: Findings from Interviews with State Medicaid Agencies and Managed Care Organizations (2023)
27	Office of Management and Budget (OMB), Revisions to OMB's Statistical Policy Directive No.15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity (2024)
28*	Oregon Health Authority, REALD & SOGI Simple Template (2024)
29*	Oregon Health Authority, REALD Implementation Guide (2020)

30	Oregon Health Authority, Summary Assessment of REALD Data Quality in the Oregon Health Plan ONE System (2018)
31	Sally Parker and Helen Grimster, How D&A Leaders Can Assess and Advance Master Data Management Maturity (2023)
32	Sally Parker and Simon Walker, 3 Essentials for Starting and Supporting Master Data Management (2022)
33	Sanny Hikmawati, Improving Data Quality and Data Governance Using Master Data Management: A Review (2021)
34	State Health Access Data Assistance Center (SHADAC), Collection of Self-Reported Disability Data in Medicaid Applications: A Fifty-State Review of the Current Landscape (2024)
35	SHADAC, Collection of Sexual Orientation and Gender Identity (SOGI) Data: Considerations for Medicaid and Spotlight on Oregon (2021)
36	SHADAC, Exploring Strategies to Fill Gaps in Medicaid Race, Ethnicity, and Language Data (2020)
37	SHADAC, Sexual Orientation and Gender Identity Data: New and Updated Information on Federal Guidance and Medicaid Data Collection Practices (2024)
38	The National Academies of Sciences, Engineering and Medicine, Measuring Sex, Gender Identity, and Sexual Orientation (2022)
39	Urban Institute, Collection of Race and Ethnicity Data for Use by Health Plans to Advance Health Equity (2022)
40	WhiteHouse.gov, Recommendations on the Best Practices for Collection of SOGI Data (n.d.)

** Sources that were not a direct input into the Literature Review Insights but were reviewed for additional context for the development of a RELD SOGI data hierarchy.*

Table 3: Literature review sources

Internal Exploration Meeting Insights

Design sessions to inform the data hierarchy business rules are ongoing. In the meantime, MassHealth has initiated discussions to identify the most appropriate tool to develop a RELD SOGI data hierarchy and subsequently the Golden Table.

Exploration Meeting Insight #1: An MDM solution could support the implementation of the MassHealth RELD SOGI data hierarchy.

1. MassHealth conducted several information-gathering sessions with the EOHHS MDM team to explore the potential use case of building a Golden Table through the implementation of a RELD SOGI data hierarchy.
2. The EOHHS MDM team indicated that MDM could be used as a service to run the data hierarchy business rules on the RELD SOGI member data without a need for storing or accessing the data outside of the MassHealth Data Warehouse (MH DW).
 - a. MDM can act as a processing and integration layer that applies the business rules and pushes the result of the data hierarchy to the Golden Table in the MH DW.
 - b. The Golden Table will be formed as the rules select a single “best” value while not persisting any source data that is maintained in the MH DW.

- c. MDM as a service can help enforce security and privacy measures at the data integration stage so that only properly managed data is stored in the MH DW.
- 3. Additional sessions will be conducted as the EOHHS MDM team and the MassHealth EDME team collaborate on building the functional and technical components required for the MDM solution.