Supplementary Appendix: IRSF-NHS Custom Vocabulary Mapping to OMOP Standard Terminologies

A1. Vocabulary Mapping Methods

The IRSF Natural History Study (IRSF-NHS) registry data were transformed to OMOP CDM v5.4 using a custom Python ETL pipeline. Because the registry contains Rett syndrome–specific clinical instruments (the Clinical Severity Scale and Motor Behavioral Assessment) and mutation-level genotyping data that have no representation in standard OMOP vocabularies, a custom vocabulary (`IRSF-NHS`) of 117 concepts was created. Concept identifiers were assigned in the 2-billion range (concept_id ≥ 2,000,000,000) per OMOP convention for locally defined vocabularies.

Prior to the mapping effort described here, all 117 IRSF-NHS concepts were completely isolated within the OMOP vocabulary infrastructure: no `Maps to` relationships to standard concepts, no hierarchical `Is a` relationships, and no ancestor/descendant entries. This meant that IRSF-NHS patients were invisible to any standard SNOMED- or LOINC-based cohort definition, and the dataset could not participate in federated OHDSI network studies.

We performed a systematic manual curation of all 117 concepts against three target vocabularies: SNOMED CT (for clinical findings), OMOP Genomic (for genetic variants), and LOINC (for clinical assessments). Each source concept was reviewed and assigned one of four mapping outcomes: (1) direct match to an external standard concept, (2) approximate match to a broader external concept, (3) self-mapping with hierarchical placement (for concepts with no external equivalent), or (4) self-mapping only (for composite scores). Mapping quality was documented for each concept to enable downstream users to assess the precision of cross-vocabulary queries.

A2. Mapping Coverage

Table A1 summarizes the mapping coverage across the four IRSF-NHS concept domains. Overall, 107 of 117 concepts (91.5%) received a mapping to an external standard concept, with the remaining 10 receiving self-mappings (OHDSI convention for source concepts that serve as their own standard representation). A total of 264 relationship rows were inserted into the `concept_relationship` table.

Table A1. IRSF-NHS Vocabulary Mapping Coverage by Domain

Domain	Concept Class	Count	Mapped to External	Self-Mapped	Coverage
Condition	Clinical Finding	14	14	0	100%
Observation	Observable Entity	48	48	0	100%
Measurement	Clinical Observation	14	13	1	100%

Observation	Clinical Observation	41	32	9	90%
Total	**—**	**117**	**107**	**10**	**97%**

Note: 5 MBA items received both a self-mapping and an approximate broader mapping; 5 composite scores (CSS Total, MBA Grand Total, 3 MBA subtotals) received self-mappings only.

Table A2. Relationship Types Inserted

Relationship	Count	Purpose
`Maps to (external)`	107	Primary standard concept mapping
`Maps to (self)`	10	OHDSI convention for source-standard concepts
`Mapped from (reverse)`	107	Reverse navigability from standard to source
`Is a (hierarchy)`	10	Places concepts in SNOMED hierarchy
`Subsumes (reverse)`	10	Reverse of Is a
Total	**264**	

A3. Complete Mapping Tables

A3.1. Condition Diagnoses (14 concepts → SNOMED CT)

All 14 condition concepts mapped to SNOMED CT at 100% coverage. Six concepts have direct matches; five atypical Rett subtypes map to the single SNOMED parent "Atypical Rett syndrome" (37397680) because SNOMED does not differentiate the Zapella, Rolando, Hanefeld, and delayed-onset clinical variants. The IRSF-NHS source concepts preserve this subtype granularity for variant-specific analyses.

Table A3. Condition Diagnosis Mappings

IRSF-NHS Concept	ID	SNOMED Target	Target ID	Quality
Classic Rett Syndrome	2000004000	Rett syndrome	4288480	Direct
Variant Rett Syndrome	2000004001	Atypical Rett syndrome	37397680	Direct
CDKL5 Deficiency Disorder	2000004003	Cyclin-dependent kinase-like 5 deficiency	36676367	Direct
MECP2 Duplication Syndrome	2000004002	MECP2 duplication syndrome	45765797	Direct
FOXG1 Syndrome	2000004006	FOXG1 syndrome	45765499	Direct
MECP2 Mutation-Positive Non-Rett	2000004005	MECP2 related disorder	1245147	Broader
Atypical Rett – Unspecified	2000004007	Atypical Rett syndrome	37397680	Direct
Atypical Rett – Preserved Speech (Zapella)	2000004008	Atypical Rett syndrome	37397680	Parent
Atypical Rett – Congenital (Rolando)	2000004009	Atypical Rett syndrome	37397680	Parent
Atypical Rett – Early Seizure (Hanefeld)	2000004012	Atypical Rett syndrome	37397680	Parent
Atypical Rett – Delayed Onset	2000004011	Atypical Rett syndrome	37397680	Parent
FOXG1 Duplication Syndrome	2000004010	FOXG1 syndrome	45765499	Parent
Other Mutation Diagnosis	2000004013	Genetic disease	37204336	Broadest
Other Non-Rett Diagnosis	2000004004	Neurodevelopmental disorder	45771096	Broadest

A3.2. Clinical Severity Scale Measurements (14 concepts → SNOMED CT)

Each CSS item was mapped to the SNOMED clinical finding that the scale item quantifies. CSS items are ordinal severity ratings (scored 0–4 or 0–5), so the mapping represents "this measurement assesses the severity of [SNOMED concept]." The CSS Total Score has no LOINC panel code or SNOMED equivalent and received a self-mapping with an `Is a` hierarchy relationship to SNOMED "Assessment score" (36684305).

Table A4. CSS Measurement Mappings

CSS Item	ID	SNOMED Target	Target ID	Quality
CSS Total Score	2000001000	Self-mapped	—	Self + Is a Assessment score
Age of Onset of Regression	2000001001	Developmental regression	43531507	Direct
Onset of Stereotypies	2000001002	Repetitive hand wringing	4171878	Direct
Head Growth	2000001003	Microcephaly	606878	Direct
Somatic Growth	2000001004	Failure to thrive	437986	Direct
Independent Sitting	2000001005	Unable to sit	4106332	Direct
Ambulation	2000001006	Abnormal gait	437643	Direct
Hand Use	2000001007	Clumsiness	4320789	Direct
Scoliosis	2000001008	Scoliosis deformity of spine	72418	Exact
Language	2000001009	Mutism	4339188	Direct
Nonverbal Communication	2000001010	Poor eye contact	4260763	Direct
Respiratory Dysfunction	2000001011	Abnormal breathing	4305080	Direct
Autonomic Symptoms	2000001012	Peripheral cyanosis	4318406	Direct
Epilepsy/Seizures	2000001013	Epilepsy	380378	Exact

A3.3. Genotype Mutation Observations (48 concepts → OMOP Genomic)

All 48 mutation-level concepts map to three gene-level OMOP Genomic standard concepts: *MECP2* gene variant measurement (35960472, n = 37), *CDKL5* gene variant measurement (35959553, n = 6), and *FOXG1* gene variant measurement (35955187, n = 5). This gene-level mapping strategy was chosen because IRSF mutation nomenclature uses protein-level notation (e.g., R168X) that does not directly correspond to ClinVar's transcript-level identifiers. Variant-level distinction is preserved only in the IRSF-NHS source concepts.

Table A5a. *MECP2* Mutation Mappings (37 concepts → OMOP Genomic 35960472)

Mutation	Concept ID	OMOP Genomic Target
Missense mutations (11)		
R106W (c.C316T)	2000003000	MECP2 gene variant (35960472)
R133C (c.C397T)	2000003001	MECP2 gene variant (35960472)
T158M (c.C473T)	2000003002	MECP2 gene variant (35960472)
L100V (c.C298G)	2000003003	MECP2 gene variant (35960472)
R106Q (c.G317A)	2000003004	MECP2 gene variant (35960472)
P152R (c.C455G)	2000003005	MECP2 gene variant (35960472)
P101R (c.C302G)	2000003006	MECP2 gene variant (35960472)
S134C (c.C401G)	2000003007	MECP2 gene variant (35960472)
D156E (c.C468G)	2000003008	MECP2 gene variant (35960472)
P225R (c.C674G)	2000003009	MECP2 gene variant (35960472)
P322L (c.C965T)	2000003010	MECP2 gene variant (35960472)
Nonsense mutations (6)		
R168X (c.C502T)	2000003011	MECP2 gene variant (35960472)
R255X (c.C763T)	2000003012	MECP2 gene variant (35960472)
R270X (c.C808T)	2000003013	MECP2 gene variant (35960472)
R294X (c.C880T)	2000003014	MECP2 gene variant (35960472)
Y141X (c.C421G)	2000003015	MECP2 gene variant (35960472)
Y141X (c.C423G)	2000003016	MECP2 gene variant (35960472)
Small deletions (9)		
710del1	2000003017	MECP2 gene variant (35960472)
1157del41	2000003018	MECP2 gene variant (35960472)
1163del35	2000003019	MECP2 gene variant (35960472)
806del1	2000003020	MECP2 gene variant (35960472)
1157del44	2000003021	MECP2 gene variant (35960472)
1164del44	2000003022	MECP2 gene variant (35960472)
807del1	2000003023	MECP2 gene variant (35960472)
1163del26	2000003024	MECP2 gene variant (35960472)
1168del6	2000003025	MECP2 gene variant (35960472)
Large deletions (5)		
Large Deletion Exon 1-2	2000003026	MECP2 gene variant (35960472)
Large Deletion Exon 3	2000003027	MECP2 gene variant (35960472)

Large Deletion Exon 3-4	2000003028	MECP2 gene variant (35960472)
Large Deletion Exon 4	2000003029	MECP2 gene variant (35960472)
Large Deletion Exon 1-4	2000003030	MECP2 gene variant (35960472)
Other / Duplication-associated (6)		
Other MCP2 Mutations (legacy)	2000003031	MECP2 gene variant (35960472)
Other MECP2 Mutations	2000003032	MECP2 gene variant (35960472)
MCP2 Duplications	2000003033	MECP2 gene variant (35960472)
Duplication-Associated FLNA	2000003034	MECP2 gene variant (35960472)
Duplication-Associated L1CAM	2000003035	MECP2 gene variant (35960472)
Duplication-Associated IRAK1	2000003036	MECP2 gene variant (35960472)

Table A5b. *CDKL5* and *FOXG1* Mutation Mappings

Mutation	Concept ID	OMOP Genomic Target
CDKL5 → OMOP Genomic 35959553 (6 concepts)		
A40V	2000003037	()
R59X	2000003038	()
Polymorphism Q791P	2000003039	()
Polymorphism R952X	2000003040	()
Other Mutation	2000003041	()
Other Mutation (variant)	2000003046	()
FOXG1 → OMOP Genomic 35955187 (5 concepts)		
N253D	2000003042	()
560dupG	2000003043	()
Other (underscore)	2000003044	()
Other (no underscore)	2000003045	()
Other (text)	2000003047	()

A3.4. Motor Behavioral Assessment Observations (41 concepts → SNOMED CT)

MBA items were mapped in three tiers: 25 items with exact or near-exact SNOMED matches, 7 items with approximate broader matches, 5 items with both self-mapping and approximate mapping (bidirectional or Rett-specific behaviors lacking SNOMED representation), and 4 composite scores with self-mappings only.

Table A6. MBA Observation Mappings (41 concepts)

Direct and Near Matches (25 items)

MBA Item	ID	SNOMED Target	Target ID	Quality / Note
Bruxism	2000002021	Bruxism	4099943	Exact
Seizures	2000002018	Seizure	377091	Exact
Dystonia	2000002033	Dystonia	375800	Exact
Bradykinesia	2000002032	Bradykinesia	4161417	Exact
Myoclonus	2000002036	Myoclonus	441553	Exact
Hyperreflexia	2000002039	Hyperreflexia	4313132	Exact
Aggressive Behavior	2000002017	Aggressive behavior	4266361	Exact
Hyperventilation	2000002023	Hyperventilation	316814	Exact
Scoliosis	2000002035	Scoliosis deformity of spine	72418	Exact
Stereotypic Hand Activities	2000002028	Repetitive hand wringing	4171878	Exact
Feeding Difficulties	2000002012	Feeding difficulties	434750	Exact
Chewing Difficulties	2000002013	Difficulty chewing	4012491	Exact
Breath Holding	2000002022	Breath holding spell	37016868	Exact
Self-Mutilating/Scratching	2000002016	Self-injurious behavior	4092411	Exact
Poor Eye/Social Contact	2000002006	Poor eye contact	4260763	Exact
Masturbation	2000002015	Masturbation	4069957	Exact
Lack of Toilet Training	2000002014	Delayed toilet training	4089215	Exact
Insensitivity to Pain	2000002019	Indifference to pain	4057307	Exact

Truncal Rocking	2000002030	Involuntary truncal rocking	4311298	Exact
Motor Skills Regression	2000002004	Developmental regression	43531507	Near
Verbal Skills Regression	2000002005	Expressive language delay	4039748	Near
Hand Clumsiness	2000002027	Clumsiness	4320789	Near
Irritability/Crying	2000002008	Feeling irritable	4184149	Near
Lack of Sustained Interest	2000002007	Inattention	4318665	Near
Speech Disturbance	2000002020	Motor speech disorder	4047111	Near

Approximate Broader Matches (7 items)

MBA Item	ID	SNOMED Target	Target ID	Quality / Note
Ataxia/Apraxia	2000002029	Ataxia	437584	Ataxia only; apraxia lost
Chorea/Athetosis	2000002037	Chorea	440990	Chorea only; athetosis lost
Hypertonia/Rigidity	2000002038	Muscular hypertonicity	4218799	Hypertonia only
Air/Saliva Expulsion	2000002024	Excessive salivation	4207204	Saliva only; air lost
Oculogyric Movements	2000002031	Oculogyric crisis	4205592	"Crisis" implies more acute
Biting Self/Others	2000002026	Self-injurious behavior	4092411	Self-biting only
Vasomotor Disturbance	2000002040	Cold feet	4154763	Main Rett manifestation

Self-Mapped with Approximate External Mapping (5 items)

MBA Item	ID	SNOMED Target	Target ID	Quality / Note
Over-active/Over-passive	2000002009	Psychomotor agitation	4187507	Bidirectional item
Does Not Reach	2000002010	Gross motor impairment	4027312	No SNOMED "failure to reach"
Does Not Follow Verbal Acts	2000002011	Receptive language disorder	443443	Approximate
Mouthing Hands/Objects	2000002025	Stereotypy habit disorder	4207660	Oral behavior not in SNOMED
Hypomimia	2000002034	Motor function behavior finding	433453	Hypomimia not in SNOMED

These 5 items received both a self-mapping (retaining as IRSF-NHS standard concepts) and an approximate Maps to relationship to the listed SNOMED target.

Composite Scores — Self-Mapped Only (4 items)

MBA Item	ID	Hierarchy	Parent ID
Grand Total	2000002000	Assessment score (Is a)	36684305
Behavioral/Social Subtotal	2000002001	Assessment score (Is a)	36684305
Orofacial/Respiratory Subtotal	2000002002	Assessment score (Is a)	36684305
Motor/Physical Subtotal	2000002003	Assessment score (Is a)	36684305

A4. Mapping Limitations

Variant-Level Information Loss in Genotype Mappings

The most significant information loss occurs in the genotype domain. All 37 *MECP2* mutation-level concepts collapse to a single OMOP Genomic gene-level concept (35960472). A federated OHDSI query for "any MECP2 variant" will correctly discover all IRSF-NHS patients, but a query for "R168X specifically" cannot use standard concept navigation and must instead reference the IRSF-NHS source concept directly (2000003011). This is an inherent limitation of the current OMOP Genomic vocabulary, which provides gene-level but not variant-level standardization for most loci. Future adoption of ClinVar variant-level concepts in OMOP could enable more granular mapping.

Semantic Gap in CSS/MBA Severity-to-Finding Mappings

CSS and MBA items are ordinal severity ratings, not binary diagnoses. Mapping CSS Scoliosis (2000001008) to SNOMED "Scoliosis deformity of spine" (72418) creates a semantic gap: a CSS scoliosis score of 0 (no scoliosis) will appear linked to the SNOMED scoliosis concept. Researchers using standard concept–based queries should be aware that the mapping indicates "this measurement assesses scoliosis severity" rather than "this patient has scoliosis." The value_as_number field in the measurement table must be consulted to determine whether the finding is present.

Compound Behavioral Items

Seven MBA items describe compound clinical phenomena (e.g., "Ataxia/Apraxia," "Chorea/Athetosis," "Hypertonia/Rigidity"). In each case, only one component could be mapped to SNOMED, losing the second. This reflects a limitation in the source instrument's design (combining related but distinct findings into single items) that cannot be fully resolved through vocabulary mapping alone.

Rett-Specific Behavioral Items Without SNOMED Representation

Five MBA items describe Rett syndrome–specific behavioral patterns that have no precise SNOMED equivalent: "Over-active or Over-passive" (a bidirectional item), "Does Not Reach for Objects/People," "Does Not Follow Verbal Acts/Appears Deaf," "Mouthing Hands/Objects," and "Hypomimia." These received both self-mappings and approximate broader mappings to ensure partial discoverability through standard queries while retaining their full semantic content as IRSF-NHS standard concepts.

A5. Implications for OHDSI Network Interoperability

The vocabulary mapping described here transforms the IRSF-NHS dataset from an isolated registry into a participant in the global OHDSI network. Four specific capabilities are now enabled:

Standard cohort definitions now discover IRSF patients. A cohort definition using SNOMED "Rett syndrome" (4288480) with `includeDescendants = true` will discover patients coded with IRSF Classic Rett Syndrome (2000004000) through the `Maps to` relationship. Prior to this mapping, these patients were invisible to standard cohort queries.

Cross-network federated studies can include IRSF data. Federated studies using OHDSI standard concept IDs—for example, an international study of epilepsy prevalence in neurodevelopmental disorders—can now include IRSF-NHS data alongside other OMOP-encoded databases without manual reconciliation.

Concept set expansion works correctly. Building a concept set with SNOMED "Epilepsy" (380378) and expanding via `Mapped from` will include both IRSF CSS Epilepsy/Seizures (2000001013) and MBA Seizures (2000002018), capturing all seizure-related assessments in the dataset.

Vocabulary search discovers IRSF-NHS data. Searching for "dystonia" in the OHDSI vocabulary browser will show the `Mapped from` relationship to IRSF MBA Dystonia (2000002033), enabling researchers to discover Rett-specific data through standard clinical terminology.

A6. Data Quality Assurance: Cross-Source Isolation

During the vocabulary mapping effort, a cross-source data contamination issue was identified and resolved in the Parthenon platform. Multiple CDM data sources share the `omop` schema with overlapping `person_id` ranges. For example, SynPUF person 100021 (a synthetic patient: "Barbara Harris," age 32, with diabetes) shared the same person_id as an IRSF patient (female, born 2003, with Rett syndrome). Without source isolation, patient profile queries could return clinical data from the wrong source.

The fix implemented filters all patient profile domain queries by `visit_occurrence_id` linkage: records must have either a NULL visit_id (the IRSF ETL convention for assessment data without explicit visit linkage) or a visit_id that matches the patient's actual visits within the same source. This eliminates 100% of cross-source contamination at the query layer without requiring physical schema isolation. The IRSF-NHS data integrity has been validated at the individual patient level to confirm that no cross-source records are returned.