Patient Registries 101

A self paced course for advocacy groups building and managing disease registries

Dr. Danielle Boyce · EpilepsyLive

Part 1 · Planning

Module 1: Fit for Purpose, Defining Your Registry's Goals

Goal: Define your registry's primary purpose before a single line of data is collected. This decision shapes every subsequent choice about data elements, platform, governance, and partnerships.

The most important question you'll answer

Before choosing a platform, hiring staff, or designing a questionnaire, you must answer: **What is this registry for?**

Patient registries are not one-size-fits-all. A registry designed for disease characterization collects different data, operates under different governance, and uses different technology than one designed for clinical trial recruitment. Building without a clear purpose produces a dataset that serves no purpose well.

The six primary registry types

1. Natural History / Disease Characterization Registry

Purpose: Document the course of a disease over time, who gets it, when symptoms begin, how it progresses, what complications arise, and how participants are currently managed.

Typical data collected: Demographics, symptom onset and severity, diagnostic journey, comorbidities, current treatments, functional outcomes, quality of life measures, biosamples.

Value: Natural history studies are foundational to the entire drug development pipeline. Regulators (FDA, EMA) increasingly use patient registry natural history data as a comparator arm for single arm trials in rare diseases.

Key resource: FDA Guidance: Natural History Studies for Drug Development

2. Clinical Trial Recruitment Registry

Purpose: Identify and pre-qualify participants who may be eligible for clinical trials.

Typical data collected: Diagnosis confirmation, key inclusion/exclusion criteria (genotype, age at onset, current medications, prior treatments), willingness to participate, geographic location, treating physician contact.

Value: Rare disease trials fail most often due to insufficient enrollment. A recruitment registry can cut enrollment timelines dramatically and reduce screen failure rates.

> **Important distinction.** A recruitment registry is **not** a clinical trial. It does not require IND or IDE. However, it does require IRB review if it collects identifiable information, and consent language must clearly distinguish registry participation from trial enrollment.

Key resource: NORD Rare Disease Registry Program

3. Contact / Participant Community Registry

Purpose: Maintain a list of participants and caregivers who want to stay connected to research updates, participate in surveys, or be contacted about future studies.

Typical data collected: Name, email, diagnosis, age, geographic region, expressed interests.

Value: Builds community, enables rapid surveys, and serves as a pipeline for more detailed data collection over time.

Privacy consideration: Even a simple contact registry requires privacy policy, consent, and a plan for data security. HIPAA may or may not apply depending on whether the registry operates within a covered entity.

4. EHR-Linked Registry

Purpose: Systematically extract structured clinical data from electronic health records, at scale, with minimal participant burden.

Data sources: Problem lists, medication records, lab results, imaging reports, clinical notes (via NLP), encounter data.

Value: Rich longitudinal clinical data without requiring participants to self report. Particularly powerful when linked to biobanks.

Complexity: High. Requires technical infrastructure (FHIR APIs, data harmonization), site agreements, and careful attention to the 21st Century Cures Act framework. See Module 9 and Module 10.

5. Biosample Registry

Purpose: Link participant clinical data to stored biological samples, DNA, plasma, tissue, CSF, available for future research.

Typical data collected: Sample type, storage location, collection date, associated clinical phenotype, consent scope for sample use.

Value: Enables biomarker discovery, genetic studies, and drug target identification. Samples with linked phenotype data are extraordinarily valuable to industry.

Key resource: ISBER Best Practices for Biobanking

6. Quality of Care / Outcomes Registry

Purpose: Track real world outcomes, treatment patterns, complications, health resource utilization, to improve standards of care.

Typical data collected: Treatments received, dosing, adverse events, hospitalizations, functional status over time.

Value: Informs clinical guidelines, supports comparative effectiveness research, and provides postmarket surveillance data valuable to industry partners.

Registries are not mutually exclusive

Most successful advocacy organization registries start with one purpose and expand over time. A common pathway:

```
Contact registry → Disease characterization → Recruitment registry → EHR-linked biosample registry
```

> **Design for expansion from the start.** Even if you're starting with a simple contact registry, use standardized identifiers and data elements from the beginning. Retrofitting a contact list to become a natural history database is much harder than building in that capability from day one.

The purpose statement

Before proceeding to the next module, write a one-paragraph purpose statement for your registry. It should answer:

1. Who are we collecting data about?
2. What data will we collect?
3. What scientific or clinical questions will this answer?
4. Who will use this data and how?
5. What does success look like in 5 years?

This statement will guide your IRB application, your platform selection, your data governance policy, and your partnership conversations.

Checklist

- [ ] Identified primary registry type(s)
- [ ] Written registry purpose statement
- [ ] Confirmed purpose with scientific advisory board (or identified need to form one, see Module 3)
- [ ] Identified 3 to 5 key scientific questions the registry will answer
- [ ] Mapped purpose to likely data elements (rough draft)
- [ ] Identified whether biosamples will be collected
- [ ] Determined geographic scope (single site, national, international)

Key resources

- AHRQ Registries for Evaluating Patient Outcomes: A User's Guide (4th Ed.), the definitive reference
- FDA: Participant-Focused Drug Development Guidance Series
- NORD Rare Disease Registry Program
- PCORI: Patient-Centered Outcomes Research

Continue to Module 2: Governance & IRB →

Module 2: Governance & IRB

Goal: Establish the governance structures and regulatory approvals your registry needs before collecting any data.

Why governance comes first

Governance failures are the most common reason patient registries stall or collapse. Data collected without proper consent cannot be shared with researchers. Registries without clear data access policies cannot attract industry partners. Registries without defined oversight become targets for legal challenges.

Getting governance right at the start protects participants, protects your organization, and makes your data usable.

IRB Review

Do you need IRB approval?

The short answer for most patient registries: **yes**.

IRB (Institutional Review Board) review is required when:

- Research involves human subjects (identifiable individuals)
- The intent is to generate generalizable knowledge
- Data will be published or shared for research purposes

Common misconception: Advocacy organizations sometimes believe that because they are not a university or hospital, IRB requirements don't apply to them. This is incorrect. The regulatory standard follows the *activity*, not the *organization type*.

Types of IRB review

Review type	When it applies	Timeline
Exempt	Minimal risk; surveys/interviews with adults; secondary use of deidentified data	Days to weeks
Expedited	Minimal risk; identifiable data; no vulnerable populations	2 to 6 weeks
Full board	More than minimal risk; children; cognitive impairment; sensitive topics	4 to 8 weeks

Getting IRB coverage without an institutional affiliation

Advocacy organizations without a home institution have several options:

1. **Commercial/Independent IRBs**, WCG (formerly Western IRB), Advarra, and Copernicus are commonly used. Expect $3,000, $8,000 for initial review plus annual continuation fees.
2. **Partner with an academic institution**, An academic collaborator can serve as the PI and provide IRB coverage through their institution.
3. **WIRB-Copernicus Group or similar**, Accepts non-institutional sponsors; experienced with participant led research.

Key resource: OHRP IRB Registration and Federalwide Assurance (FWA)

Informed Consent

Consent is not just a regulatory checkbox, it is the foundation of the participant-researcher relationship. A well-designed consent process builds trust and improves retention.

What consent must cover

- What data will be collected and why
- Who will have access to the data
- How the data will be stored and for how long
- Whether samples will be collected and what they may be used for
- Whether data may be shared with commercial entities (industry sponsors)
- The right to withdraw and what happens to data upon withdrawal
- Whether the registry is HIPAA-covered and how PHI is protected
- Whether participants will receive results or incidental findings

Consent models

Traditional one-time consent: Participant signs consent document once. Simple but inflexible, if the registry expands its scope, you may need to re-consent.

Broad consent: Participant consents to future unspecified research uses within a defined scope. Requires explicit IRB approval and specific regulatory language (45 CFR 46.116(d)).

Dynamic/tiered consent: Participant makes granular choices about data uses (e.g., "yes to academic research, no to commercial use, yes to contact for future studies"). More complex to implement but increasingly preferred by participants.

Electronic consent (eConsent): Enables multimedia consent, remote participation, and timestamped audit trails. Accepted by most IRBs. Tools include REDCap (free), Medidata Rave eConsent, and others.

> **Recommendation.** For rare disease registries, **dynamic consent** is worth the extra investment. Participants are sophisticated advocates who appreciate control over their data, and tiered consent dramatically expands partnership options later.

Data Governance Policy

Your data governance policy defines who can access registry data, under what conditions, and with what oversight. This document is essential for:

- Protecting participant privacy
- Enabling data sharing with researchers
- Negotiating with industry partners
- Satisfying IRB requirements

Core components of a data governance policy

1. **Data Access Committee (DAC)**, Who reviews and approves data access requests? Composition typically includes a participant representative, a scientific advisor, a legal/privacy officer, and a member of the advocacy organization's leadership.
2. **Access tiers**, Define levels of data access (e.g., summary statistics only → deidentified

individual records → identified data under DUA → linked biosamples)

3. **Data Use Agreements (DUAs)**, Legal contracts specifying permitted uses, prohibitions (e.g., no reidentification attempts, no selling), and data destruction requirements.

4. **Publication policy**, Who must be acknowledged or included as coauthor when registry data is published? Does the advocacy organization have review rights before submission?

5. **Data retention and destruction**, How long is data kept? What is the process for destroying data upon participant withdrawal or registry closure?

Key resource: Global Alliance for Genomics and Health (GA4GH) Framework for Responsible Sharing of Genomic and Health related Data

HIPAA and Participant Data Privacy

HIPAA applies if your registry is operated by or in partnership with a **covered entity** (healthcare provider, health plan, or healthcare clearinghouse) or their **business associates**.

If your advocacy organization collects data directly from participants outside a covered entity context, HIPAA may not technically apply, but you should still implement equivalent protections because:

- Participants expect and deserve strong privacy protections
- Industry and academic partners will require them
- State privacy laws (especially California CCPA/CPRA) may apply regardless

Safe Harbor deidentification: Removing 18 specific identifiers from a dataset creates a "Safe Harbor" deidentified dataset that is no longer subject to HIPAA. Key identifiers to remove: names, geographic data smaller than state, dates (except year), phone numbers, email addresses, SSN, medical record numbers, device identifiers, URLs, IP addresses, biometric identifiers, photos.

Checklist

- [ ] Determined whether IRB review is required
- [ ] Selected IRB (institutional or commercial)
- [ ] Drafted consent document covering all required elements
- [ ] Chosen consent model (traditional, broad, or dynamic)
- [ ] Drafted data governance policy
- [ ] Established Data Access Committee with participant representation
- [ ] Drafted Data Use Agreement template
- [ ] Determined HIPAA applicability and implemented appropriate protections
- [ ] Established publication policy

Key resources

- AHRQ Registry User's Guide Chapter 8: Legal and Liability Issues
- OHRP Human Subjects Regulations (45 CFR 46)
- GA4GH Data Access Framework

- WCG IRB (Independent IRB)
- Advarra IRB
- NIH Broad Consent Template
- IAPP Privacy Resource Center

← Module 1 | Module 3: Scientific Advisory Boards →

Module 3: Scientific Advisory Boards

Goal: Build a Scientific Advisory Board (SAB) that provides genuine scientific direction, not just credibility optics, and structure it to serve your registry's long term needs.

Why your registry needs a SAB

A well-constituted SAB is not a prestige accessory. It provides:

- **Scientific validity:** Ensures your data elements, study design, and analysis plans meet publishable standards
- **Regulatory credibility:** FDA and EMA look favorably on participant led research with independent scientific oversight
- **Industry credibility:** Pharma and biotech partners evaluate SAB composition when deciding whether registry data is worth licensing
- **Academic credibility:** Journal reviewers assess whether the registry was designed with appropriate scientific rigor
- **IRB support:** IRBs are more comfortable with participant led research when there is independent scientific oversight
- **Endpoint development:** Helps identify and validate clinically meaningful endpoints for future trials

SAB composition

Core roles to fill

Disease specific clinician scientists (2 to 4), Physicians or scientists with direct expertise in your disease. These are your most important members. They validate your data elements, interpret natural history findings, and connect you to the clinical research community.

Biostatistician (1), Essential for designing data collection that produces analyzable results. Without statistical input at the design stage, registries often collect data that cannot answer the scientific questions they were built to address.

Participant/caregiver representative (1 to 2), Ensures the registry captures outcomes that matter to participants, not just outcomes that are easy to measure. This is non-negotiable for a advocacy organization.

Clinical trial methodologist (1), Advises on how registry design can support future trial endpoints, comparator arms, and recruitment.

Bioinformatician or data scientist (1), Critical if you are collecting genomic, proteomic, or other high-dimensional data.

Regulatory scientist (1, optional but valuable), Advises on FDA/EMA guidance, natural history data acceptance, and the regulatory pathway from registry to trial support.

Ethicist (1, optional), Particularly valuable for registries collecting genomic data, pediatric data, or planning broad consent.

What to look for in members

- Active publishing record in your disease area
- No conflicts of interest that would compromise independence (industry employment is a common conflict, can be managed with disclosure and recusal policies, but must be explicit)
- Genuine interest in participant led research (not just a CV line)
- Willingness to commit meaningful time, quarterly meetings minimum, plus ad hoc consultations

Structure and operations

Charter

Every SAB needs a written charter covering:

- Mission and scope of the SAB's authority
- Composition requirements and term limits
- Meeting frequency and quorum requirements
- Conflict of interest disclosure and management policy
- Compensation policy (honoraria, travel)
- Relationship to the advocacy organization's board of directors
- Decision making process (advisory vs. binding)

Meeting cadence

- **Full SAB meeting:** Quarterly (90 minutes minimum), or semi-annually for very busy members
- **Working groups:** Monthly or as-needed for specific projects (e.g., data element committee, publication committee)
- **Annual meeting:** In-person or hybrid intensive; review registry progress, set annual priorities, discuss publications

Compensation

SAB members should be compensated. Uncompensated advisory roles lead to disengagement. Typical honoraria:

- Meeting attendance: $500, $1,500 per meeting
- Manuscript review/authorship work: $500, $2,000 per project
- Travel expenses reimbursed at cost

Compensation levels vary significantly by organization size and the seniority of advisors. Even small honoraria signal that you value their time.

SAB responsibilities specific to registries

Data element review

The SAB should formally review and approve your core data element set before data collection begins. This review should address:

- Are these elements sufficient to answer our scientific questions?
- Are these elements consistent with the literature and with other registries in this disease space?
- Are there validated instruments or scales we should use rather than custom questions?
- What are the minimum data elements required for a natural history analysis?

Publication oversight

Define clearly before the first publication:

- Who is eligible for authorship vs. acknowledgment when registry data is published?
- Does the advocacy organization have a right of review before submission?
- How are participant coauthors identified and supported?
- What is the embargo policy for data shared with industry partners?

Endpoint development

For rare diseases pursuing FDA approval, the SAB should actively participate in FDA meetings (e.g., Type B or C meetings, Natural History Workshop) and help develop clinically meaningful endpoints that regulatory agencies will accept.

How to recruit SAB members

1. **Start with your clinical network**, Physicians treating your community are natural candidates. Ask your most engaged participants who their best doctor is.
2. **Conference outreach**, Attend disease specific conferences and approach researchers whose work aligns with your registry's goals.
3. **NORD and Global Genes networks**, Both organizations connect participant groups with scientific advisors.
4. **Reach out cold with a strong pitch**, Many researchers are genuinely motivated by participant led science. A clear, compelling ask with defined time commitments gets more yeses than you expect.

Checklist

- [ ] Drafted SAB charter
- [ ] Identified required expertise areas
- [ ] Recruited disease specific clinician scientists
- [ ] Recruited biostatistician
- [ ] Included participant/caregiver representative(s)
- [ ] Established conflict of interest policy
- [ ] Defined compensation structure
- [ ] Scheduled inaugural meeting before data collection begins
- [ ] SAB has reviewed and approved data element set
- [ ] Defined publication policy with SAB input

Key resources

- NORD Scientific Advisory Board Resources
- Global Genes RARE Advocacy Resources
- FDA Natural History Workshop Proceedings
- PCORI Stakeholder Engagement Resources

← Module 2 | Module 4: Choosing a Platform →

Module 4: Choosing a Platform or Vendor

Goal: Evaluate and select the technology platform that best fits your registry's purpose, budget, and long term data needs.

Platform decisions are architecture decisions

The platform you choose determines what data you can collect, how it can be structured, whether it can interoperate with EHRs, and how easily data can be exported for analysis. A platform that seems easy to launch can become a barrier to research if it lacks FHIR support, cannot export in standard formats, or locks your data in a proprietary schema.

Evaluate platforms against your scientific requirements, not just their ease of setup.

The major platform categories

Purpose built registry platforms

Designed specifically for patient registries. Typically include patient portals, clinical site interfaces, data management, and export functionality.

Platform	Best for	Notes
PatientCrossroads / IAMRARE	Rare disease advocacy orgs; NORD partnership	Strong community features; NORD-affiliated orgs get preferential pricing
Castor EDC	Clinical-grade data with patient portals	European origin; GDPR-native; good FHIR support
OpenClinica	Research-grade; open source option available	Requires more technical setup; very flexible
Medidata Rave	Larger orgs with industry partnerships	Enterprise cost; very common in industry-sponsored trials
Veeva Vault	Orgs with existing Veeva relationships	Industry-grade; high cost

REDCap (Research Electronic Data Capture)

Widely used in academic research. Free to institutions with a license (over 6,000 institutions worldwide hold licenses). Highly flexible, validated, and well-understood by academic collaborators.

Pros: Free (with institutional access), flexible, validated, large user community, good export options
Cons: Not designed as a patient portal; requires technical setup; hosting requires an institutional partner or paid hosting

Key resource: REDCap Consortium

FHIR-native platforms

If EHR integration is a priority, consider platforms built on or natively supporting HL7 FHIR:

- **Firely**, FHIR server and tooling
- **Smile CDR**, Enterprise FHIR platform
- **Microsoft Azure Health Data Services**, Cloud FHIR infrastructure
- **Google Cloud Healthcare API**, FHIR R4 compatible cloud storage

These are infrastructure components, not turnkey registry solutions, you'll typically build on top of them.

Key questions to ask every vendor

Data ownership and portability

- Who owns the data? (Answer must be: the advocacy organization owns the data, always.)
- In what format can data be exported? (Require: CSV, JSON, and ideally FHIR R4)
- What is the process for full data export if we change vendors?
- Are there any data lock-in provisions?

Security and compliance

- What is your SOC 2 Type II certification status?
- Are you HIPAA-compliant? Do you sign a Business Associate Agreement (BAA)?
- Where are servers located? (Matters for GDPR if collecting international data)
- What is your data breach notification policy and history?

Interoperability

- Do you support HL7 FHIR R4?
- Can participants import their health records via SMART on FHIR / 21st Century Cures Act APIs?
- Do you support standard terminologies (SNOMED CT, LOINC, ICD-10, HPO)?

Participant experience

- Is there a patient portal with longitudinal data entry?
- Can participants view their own data?
- Is the interface accessible (WCAG 2.1 AA)?
- Does it support mobile?

Research features

- Can you define branching logic in questionnaires?
- Is there audit trail / version control on data changes?
- Does the platform support validated patient-reported outcome instruments (PROMIS, EQ-5D, etc.)?
- What analysis tools are built in vs. what requires export?

Pricing model

- Is pricing per participant, per site, per year, or by data volume?
- What does scaling cost as your registry grows?
- Are there costs for data export?
- Is there a nonprofit or advocacy organization pricing tier?

Build vs. buy

Some organizations consider building a custom registry. This is almost always a mistake for organizations without a dedicated software engineering team.

Custom builds are justified when:

- Your data architecture is uniquely complex (e.g., multiomics, imaging, biosample tracking)
- You have a large technical team and sustained engineering budget
- No commercial platform meets your FHIR/interoperability requirements

For most advocacy organizations: Choose an established platform. The ongoing maintenance burden of a custom solution, security patches, hosting, feature development, is far greater than it appears at launch.

Hosted vs. self-hosted

Most advocacy organizations should use **hosted/SaaS platforms**. Self-hosting requires server management, security patching, backup infrastructure, and technical staff, capabilities most advocacy organizations lack.

Exception: If you have an academic partner willing to host REDCap, this is a cost-effective and academically credible option.

Checklist

- [ ] Listed platform requirements based on registry purpose and data elements
- [ ] Confirmed HIPAA BAA is available from shortlisted vendors
- [ ] Confirmed data ownership is explicitly assigned to your organization in vendor contracts
- [ ] Confirmed full data export capability (format and process)
- [ ] Evaluated FHIR support if EHR integration is a priority
- [ ] Requested nonprofit/advocacy pricing
- [ ] Verified SOC 2 Type II certification
- [ ] Piloted platform with a small test dataset before committing
- [ ] Confirmed SAB approval of platform selection

Key resources

- AHRQ Registry User's Guide Chapter 3: Registry Design
- NORD IAMRARE Registry Program
- REDCap Consortium
- Castor EDC
- HL7 FHIR Implementation Guide Registry

← Module 3 | Module 5: Common Data Elements →

Part 2 · Data Architecture

Module 5: Common Data Elements

Goal: Understand what Common Data Elements (CDEs) are, why they matter, and how to select and implement them to make your registry data maximally reusable.

What are Common Data Elements?

A Common Data Element (CDE) is a data element, a question, measurement, or variable, that has been precisely defined with:

- A standard definition of what is being measured
- A standard set of permissible values or units
- A standard name and code
- Documentation of how it is collected

CDEs allow researchers to combine and compare data across registries, studies, and institutions. Without them, a "disease severity" question in your registry may measure something completely different from the same question in a collaborator's study, making the datasets impossible to pool.

Why CDEs matter for your registry

Data pooling: If your registry uses the same CDEs as other registries in your disease space, your combined dataset is dramatically more powerful than either dataset alone.

Regulatory acceptance: FDA increasingly requires or strongly prefers CDEs from registries used to support drug development. The FDA's own CDE repository is the gold standard.

Industry partnerships: Pharmaceutical companies evaluating your registry for trial support will check whether your data elements align with established standards. Proprietary data schemas are a barrier to partnership.

Publication: Journal reviewers evaluate data element choices. Using established, validated CDEs strengthens the methodology section of any paper.

The CDE ecosystem

NIH Common Data Element Repository

The FDA and NIH jointly maintain the **CDE Repository**, the most important source for CDEs in US-based registries.

cde.nlm.nih.gov, Search by disease, research domain, or data element name. Includes CDEs used in FDA-cleared instruments and NIH funded studies.

NINDS CDE Project

The National Institute of Neurological Disorders and Stroke developed a comprehensive

CDE framework now widely used across neurological disease registries. Even if your disease isn't neurological, the NINDS CDE methodology is instructive.

Common Data Elements: Standards and Tools

PhenX Toolkit

PhenX provides standardized measures for phenotypes and exposures. Particularly strong for epidemiological and population health measures.

phenxtoolkit.org

PROMIS (Patient-Reported Outcomes Measurement Information System)

PROMIS is an NIH-developed library of validated patient-reported outcome measures covering physical, mental, and social health. Use PROMIS instruments rather than writing your own quality-of-life questions.

Key PROMIS domains for rare disease registries:

- Physical Function
- Fatigue
- Pain Interference and Pain Intensity
- Sleep Disturbance
- Anxiety and Depression
- Social Participation

healthmeasures.net

Structuring your data element set

Core (required) vs. supplemental (optional)

Divide your data elements into:

- **Core elements:** Required for all participants at enrollment and each follow up. Keep this list short, every additional required element reduces completion rates.
- **Supplemental elements:** Collected when available, or for specific subpopulations (e.g., genetic data only for participants who consent to genotyping).
- **Time-stamped longitudinal elements:** Collected at defined intervals to track disease progression.

The minimum dataset problem

A common mistake: collecting too many data elements. A 200-question enrollment form drives away participants and produces a dataset with massive missingness.

Design principle: What is the minimum set of data elements that answers your core scientific questions? Start there.

A well-designed rare disease registry often has:

- 15 to 30 core enrollment elements
- 5 to 10 core follow up elements (collected every 6 to 12 months)

- 20 to 50 supplemental elements (collected once or as available)

Disease specific CDEs

For many rare diseases, disease specific CDE sets already exist. Before designing your own, search:

- NORD Registry Database
- Orphanet
- NCI Thesaurus, Standardized cancer and biomedical terminology
- ClinicalTrials.gov, Review outcome measures used in trials in your disease

Data element documentation

For each data element in your registry, document:

Field	Description
Element name	Short, unique identifier
Definition	Precise definition of what is being measured
Data type	String, integer, date, coded value, etc.
Permissible values	For coded fields: the complete value set
Unit of measure	For numeric fields
Collection method	Self reported, clinician-assessed, EHR-extracted, etc.
Source/reference	CDE ID from NIH repository, validated instrument, etc.
Collection timepoint	Enrollment, 6-month follow up, etc.

Checklist

- [ ] Searched NIH CDE Repository for disease specific CDEs
- [ ] Reviewed PROMIS for patient-reported outcome measures
- [ ] Checked PhenX Toolkit for epidemiological measures
- [ ] Identified existing disease specific registries and their data elements
- [ ] Divided elements into core vs. supplemental
- [ ] SAB has reviewed and approved data element set
- [ ] Each element is documented with definition, type, permissible values, and source
- [ ] Pilot-tested questionnaire for completion time and participant comprehension

Key resources

- NIH CDE Repository
- NINDS Common Data Elements
- PROMIS Health Measures
- PhenX Toolkit
- AHRQ Registry User's Guide Chapter 4: Data Elements

← Module 4 | Module 6: Standardized Vocabularies →

Module 6: Standardized Vocabularies

Goal: Learn the key medical terminologies and coding systems your registry should use to

ensure data is interoperable, searchable, and ready for research.

Why standardized vocabularies matter

A diagnosis of "Duchenne muscular dystrophy" in your registry might be entered as "DMD", "Duchenne MD", "Duchenne muscular dystrophy", or "Duchenne's". Without standardization, these are four different values, impossible to aggregate or query reliably.

Standardized vocabularies solve this by assigning unique, stable codes to every concept. When your registry uses the same codes as EHRs, biobanks, and other registries, data becomes interoperable by default.

The essential vocabularies

ICD-10-CM (International Classification of Diseases, 10th Edition, Clinical Modification)

Use for: Diagnoses, symptoms, procedures. The billing standard in US healthcare.

Why it is important: EHRs store diagnoses as ICD-10 codes. Using ICD-10 in your registry allows direct linkage to clinical records.

Limitation: ICD-10 is designed for billing, not research. Many rare diseases are collapsed into a single non-specific code. Supplement with disease specific codes (ORDO, HPO).

Resource: ICD-10-CM Browser

SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms)

Use for: Clinical findings, procedures, body structures, organisms, substances. The most comprehensive clinical terminology.

Why it is important: Used natively in EHR systems (Epic, Cerner) and required for FHIR interoperability. Provides far more granular clinical coding than ICD-10.

Resource: SNOMED CT Browser

LOINC (Logical Observation Identifiers Names and Codes)

Use for: Laboratory tests, clinical measurements, patient-reported outcomes, survey questions. Every lab test has a LOINC code.

Why it is important: Required for FHIR Observation resources. Enables lab result comparison across institutions.

Example: Serum creatinine = LOINC 2160-0. Using this code means any system that receives your data knows exactly what was measured, in what units, by what method.

Resource: LOINC Search

RxNorm

Use for: Medications. Provides normalized names and codes for drugs.

Why it is important: Participants report medications in countless ways ("methotrexate", "MTX", "Rheumatrex", "25mg methotrexate weekly"). RxNorm maps all of these to a single concept.

Resource: RxNorm Browser

Orphanet Rare Disease Ontology (ORDO)

Use for: Rare disease classification. Provides codes for over 10,000 rare diseases and their subtypes.

Why it is important: ICD-10 doesn't cover most rare diseases specifically. ORDO codes are the standard for rare disease registries, used by EMA and European rare disease networks.

Resource: Orphanet

MedDRA (Medical Dictionary for Regulatory Activities)

Use for: Adverse events, symptoms, medical history, primarily in a regulatory/clinical trial context.

Why it is important: Required for adverse event reporting to FDA and EMA. If your registry is used to support regulatory submissions, MedDRA is essential.

Resource: MedDRA MSSO

Practical implementation

You don't need to implement every vocabulary at once. Prioritize based on your registry's purpose:

Registry purpose	Must-have vocabularies
Natural history	ICD-10, SNOMED CT, LOINC, ORDO, HPO (Module 7)
EHR-linked	SNOMED CT, LOINC, RxNorm, ICD-10
Trial recruitment	ICD-10, SNOMED CT, MedDRA
Genomic	HPO, OMIM, HGNC (see Module 7)

Key resources

- NLM Unified Medical Language System (UMLS), Maps between vocabularies
- OBO Foundry, Open biological ontologies
- NCBO BioPortal, Browse and search biomedical ontologies
- HL7 Terminology, FHIR-aligned value sets

← Module 5 | Module 7: HPO, GA4GH & Phenopackets →

Module 7: HPO, GA4GH & Phenopackets

Goal: Understand the Human Phenotype Ontology (HPO), the GA4GH data sharing framework, and Phenopackets, the emerging standard for exchanging participant phenotype and genomic data.

The Human Phenotype Ontology (HPO)

What is HPO?

The Human Phenotype Ontology is a standardized vocabulary of over **18,000 terms** describing human phenotypic abnormalities, symptoms, clinical findings, and disease features. Each term has a unique identifier (e.g., HP:0001250 = Seizure) and is organized in a hierarchical structure.

HPO was developed by the Monarch Initiative and is now the international standard for phenotype description in rare disease research.

hpo.jax.org

Why your registry needs HPO

Without HPO, "seizure" in your registry and "convulsion" in a collaborator's registry may or may not mean the same thing. HPO maps both terms to HP:0001250, making them computationally equivalent.

This is very important for:

- **Genotype phenotype analysis**, Linking specific genetic variants to specific clinical features
- **Cross-registry comparison**, Comparing phenotype frequencies across institutions
- **Diagnosis support**, Tools like Phenomizer and LIRICAL use HPO terms to suggest diagnoses
- **GA4GH / Phenopackets**, HPO is the phenotype backbone of the Phenopackets standard (see below)

Implementing HPO in your registry

1. Map each clinical feature you collect to the most specific applicable HPO term
2. For each feature, record: HPO term ID, observation status (present/absent/unknown), age at onset, severity modifier
3. Use HPO's own browser to identify appropriate terms: hpo.jax.org
4. Avoid creating custom phenotype terms, if the concept exists in HPO, use it

Related ontologies

- **OMIM (Online Mendelian Inheritance in Man):** Gene-disease relationships, omim.org
- **HGNC (HUGO Gene Nomenclature Committee):** Standardized gene symbols, genenames.org
- **ClinVar:** Variant-disease assertions, ncbi.nlm.nih.gov/clinvar

The Global Alliance for Genomics and Health (GA4GH)

What is GA4GH?

GA4GH is an international standards-setting body that develops frameworks and technical standards for responsible genomic and health data sharing. It is the closest thing the field has to a global governance authority for genomic data.

ga4gh.org

The GA4GH Framework for Responsible Sharing

The GA4GH framework, adopted by hundreds of institutions globally, establishes principles for:

- **Data access:** Who can access genomic data and under what conditions
- **Data security:** Technical standards for protecting genomic data in transit and at rest
- **Ethics:** Consent standards, benefit sharing, vulnerable populations
- **Identity management:** How individuals are identified across systems

Even if your registry doesn't immediately involve genomics, aligning with the GA4GH framework builds credibility with academic and industry partners who operate within it.

Key GA4GH Standards relevant to registries

Standard	Purpose
Phenopackets	Structured exchange of phenotype + genomic data
Beacon API	Query whether a variant or phenotype exists in a dataset
Data Connect	Federated search across multiple datasets
Passport/Visas	Identity and access management for data sharing
VRS (Variant Representation Spec)	Standardized genomic variant description

Phenopackets

What is a Phenopacket?

A Phenopacket is a structured, computable representation of a participant's phenotype, medical history, and genomic data. Think of it as a standardized medical summary designed for machines to read and exchange, a participant's clinical story in a format that any compliant system can interpret.

The Phenopacket schema (v2) includes:

- **Subject:** Individual ID, date of birth, sex, taxonomy
- **Phenotypic features:** HPO coded observations with onset, severity, and status (present/absent)
- **Diseases:** OMIM/ORDO-coded diagnoses with onset
- **Measurements:** LOINC-coded lab values and clinical measurements
- **Genomic interpretations:** Variant data with ACMG classifications
- **Medical actions:** Treatments, procedures, doses
- **Family history:** Pedigree information

github.com/phenopackets/phenopacket-schema

Why Phenopackets matter for your registry

Phenopackets are becoming the submission standard for:

- **NCBI ClinVar**, Variant submissions increasingly require Phenopacket format

- **Matchmaker Exchange**, Connecting undiagnosed participants globally
- **ERDERA**, European rare disease data sharing networks
- **Monarch Initiative**, Cross-species phenotype analysis
- **GA4GH Beacon**, Federated data discovery

Designing your registry to export Phenopackets from the start means your data is immediately accessible to this entire ecosystem.

Getting started with Phenopackets

1. Review the Phenopackets v2 schema documentation: phenopackets.readthedocs.io
2. Use the Python or Java SDK for programmatic creation
3. Validate Phenopackets with the official validator before submission
4. Map your existing data elements to Phenopacket fields

Checklist

- [ ] HPO terms mapped to all phenotypic features in registry
- [ ] HPO term IDs stored alongside display labels (not just text)
- [ ] OMIM and/or ORDO codes assigned to diagnoses
- [ ] HGNC gene symbols used for all genetic data
- [ ] Reviewed GA4GH Framework for data sharing alignment
- [ ] Evaluated whether Phenopacket export is feasible with chosen platform
- [ ] Consent language covers genomic data sharing via GA4GH-compliant networks

Key resources

- Human Phenotype Ontology
- GA4GH
- Phenopackets Schema v2
- Monarch Initiative
- Matchmaker Exchange
- ClinVar

← Module 6 | Module 8: OMOP & Data Standards →

Module 8: OMOP & the OHDSI Network

Goal: Understand the OMOP Common Data Model and the OHDSI network, and evaluate whether converting your registry to OMOP is right for your goals.

What is OMOP?

The **Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)** is a standardized data structure for observational health data. Originally developed by the FDA and now maintained by OHDSI (Observational Health Data Sciences and Informatics), OMOP defines:

- A standard relational database schema (tables, fields, relationships)
- Standard vocabularies for every concept (SNOMED CT, LOINC, RxNorm, ICD-10, CPT, etc., all mapped to OMOP concept IDs)
- Standard analytic tools that run identically across any OMOP formatted dataset

ohdsi.org
ohdsi.org/omop

Why OMOP is important for patient registries

Federated research at scale

The OHDSI network includes **over 300 databases in 70+ countries**, collectively representing over 800 million participant records in OMOP format. When your registry is in OMOP format, it can participate in federated studies where the same analysis runs simultaneously across all OHDSI network databases, without any individual site sharing raw data.

This means: a natural history analysis of your disease can include participants from academic medical centers worldwide, massively increasing statistical power.

Regulatory use

FDA uses OMOP formatted real world data for regulatory submissions. If your registry aspires to support drug development, OMOP compatibility is increasingly expected.

Standard analytics tools

OHDSI has developed a library of open-source analytics tools that run on any OMOP database:

- **ATLAS**, Web-based cohort definition, incidence analysis, treatment pathways
- **ACHILLES**, Data quality and characterization
- **HADES**, R package library for population level estimation and participant level prediction
- **Strategus**, Orchestrates large-scale network studies

OMOP core tables relevant to registries

Table	Contains
`PERSON`	Demographic information
`CONDITION_OCCURRENCE`	Diagnoses (SNOMED CT coded)
`DRUG_EXPOSURE`	Medication records (RxNorm coded)
`MEASUREMENT`	Lab results, vital signs (LOINC coded)
`OBSERVATION`	Clinical findings, survey responses
`PROCEDURE_OCCURRENCE`	Procedures (CPT/SNOMED coded)
`VISIT_OCCURRENCE`	Encounter records
`DEATH`	Death records
`NOTE`	Clinical notes

Is OMOP right for your registry?

When OMOP conversion is worthwhile

- You want to participate in federated OHDSI network studies

- You are collecting EHR-sourced data that naturally aligns with clinical data structures
- You have a technical team or academic partner capable of implementing and maintaining OMOP
- Your registry will be large enough to contribute meaningfully to federated analyses

When OMOP may not be necessary

- You are a small registry focused on rare disease natural history with highly disease specific data elements
- Your data is primarily patient-reported and doesn't align naturally with clinical data tables
- You lack technical resources for ETL (Extract, Transform, Load) development
- Your primary goal is within-registry analysis, not federated research

The middle path

Many registries maintain their native schema and create an **OMOP export**, a periodic conversion of their data to OMOP format for participation in specific studies, without rebuilding their entire infrastructure in OMOP. This is a practical compromise.

Getting started with OMOP

1. **Review the OMOP CDM documentation:** ohdsi.github.io/CommonDataModel
2. **Assess your vocabulary coverage:** Use the ATHENA vocabulary browser to check whether your concepts have OMOP standard equivalents, athena.ohdsi.org
3. **Use Usagi for vocabulary mapping:** OHDSI's Usagi tool helps map source concepts to OMOP standard concepts, github.com/OHDSI/Usagi
4. **Run ACHILLES for data quality:** After conversion, run ACHILLES to characterize your data and identify quality issues
5. **Connect to the OHDSI community:** forums.ohdsi.org is an active community with extensive help resources

Key resources

- OHDSI
- OMOP Common Data Model
- ATHENA Vocabulary Browser
- ATLAS, Try the demo
- Book of OHDSI, Free, comprehensive textbook
- OHDSI Community Forums

← Module 7 | Module 9: EHR Integration →

Part 3 · Data Collection

Module 9: EHR Integration, CCDA & FHIR

Goal: Understand how to extract structured clinical data from EHR systems using industry standards, reducing participant burden and improving data quality.

Why EHR integration is important

Patient-reported data is invaluable, but it has limitations: participants may not know their exact lab values, may misremember medication doses, or may not be aware of diagnoses documented in their records. EHR-sourced data provides clinically precise, longitudinal records that complement what participants report.

The good news: federal law (the 21st Century Cures Act) now requires EHR vendors to provide patient access to their data, and your registry can leverage this.

C-CDA: The Legacy Standard

What is C-CDA?

The **Consolidated Clinical Document Architecture (C-CDA)** is an HL7 XML standard for clinical documents, the format EHRs have used for years to exchange participant records. When a participant "downloads" their health records from a patient portal, they typically receive a C-CDA document (or a PDF).

C-CDA documents include sections for:

- Problems (diagnoses, coded in SNOMED CT or ICD-10)
- Medications (RxNorm)
- Allergies
- Lab results (LOINC)
- Procedures
- Vital signs
- Immunizations
- Social history

C-CDA for registries

Participants can download their C-CDA from their patient portal and upload it to your registry. Your platform can then parse the XML to extract structured data, diagnoses, labs, medications, without the participant needing to manually enter it.

Limitations of C-CDA:

- Snapshot in time (not continuously updated)
- Inconsistent implementation across EHR vendors
- XML parsing requires technical infrastructure
- Requires participant action to download and upload

FHIR: The Modern Standard

What is FHIR?

HL7 FHIR (Fast Healthcare Interoperability Resources) is the modern standard for healthcare data exchange. Unlike C-CDA's document-based approach, FHIR uses a RESTful API model, meaning data can be queried, retrieved, and continuously synchronized, just like any web API.

FHIR represents data as **Resources**, discrete objects like Participant, Condition, Observation, Medication, and Procedure. Each resource has a defined structure, standard fields, and uses standard terminologies (SNOMED CT, LOINC, RxNorm).

FHIR R4 is the current stable version. FHIR R5 is available but not yet widely implemented.

hl7.org/fhir

FHIR for registries

With FHIR APIs, your registry can:

- Pull a participant's diagnoses, labs, and medications directly from their EHR in real time
- Receive automatic updates when new lab results are available
- Query across multiple participants at clinical sites that grant API access
- Submit data back to EHRs (e.g., registry-generated phenotype scores visible to the treating physician)

SMART on FHIR

SMART on FHIR is a framework that enables third party apps (like your registry) to securely access EHR data with patient authorization. It uses OAuth 2.0 for authorization, the same standard that powers "Login with Google."

A participant can:

1. Visit your registry's patient portal
2. Click "Connect my health records"
3. Authenticate with their EHR's patient portal (e.g., MyChart)
4. Authorize specific data types to be shared with your registry
5. Your registry receives a FHIR access token and can pull their records

smarthealthit.org

Key FHIR Resources for registries

FHIR Resource	What it contains	Terminology
`Patient`	Demographics, identifiers	,
`Condition`	Diagnoses, problems	SNOMED CT, ICD-10
`Observation`	Lab results, vital signs, survey answers	LOINC
`MedicationRequest`	Prescribed medications	RxNorm
`Procedure`	Procedures performed	SNOMED CT, CPT
`DiagnosticReport`	Lab panels, imaging reports	LOINC
`DocumentReference`	Clinical notes, C-CDA documents	LOINC

`Questionnaire / QuestionnaireResponse`	Structured forms and participant responses	LOINC

Implementation considerations

Site agreements

Pulling FHIR data from clinical sites (rather than directly from participants) requires:

- Data Use Agreement with the health system
- IRB approval covering EHR data access
- Possible Business Associate Agreement (BAA) under HIPAA

Bulk FHIR

For population level data extraction (pulling data for many participants at once), use the **FHIR Bulk Data Access** specification, designed for exactly this use case.

hl7.org/fhir/uv/bulkdata

Checklist

- [ ] Determined whether participant-initiated C-CDA upload or SMART on FHIR is the right approach
- [ ] Selected platform with FHIR R4 support
- [ ] Mapped registry data elements to FHIR resources
- [ ] Drafted site agreements for clinical site FHIR access
- [ ] Confirmed IRB protocol covers EHR data collection
- [ ] Tested SMART on FHIR patient authorization flow

Key resources

- HL7 FHIR R4 Specification
- SMART on FHIR
- FHIR Bulk Data Access
- CMS Interoperability Rule FHIR APIs

← Module 8 | Module 10: Patient-Reported Data & 21st Century Cures →

> **Aligned with EpilepsyLive.** The Registry Toolkit on this site frames record acquisition as five pathways a family or registry can use. This module covers the mechanics of the first two; the rest are summarized in Module 10 and demonstrated in the toolkit.
>
> - **Patient portal login**, SMART on FHIR (covered above).
> - **Download & upload**, a C-CDA file the person exports and re-uploads (covered above).
> - **HIPAA authorization**, a signed release directing a provider to send records.
> - **Network exchange**, TEFCA / Individual Access Services (see Module 10).
> - **EEG & imaging studies**, large binary studies requested separately (see Module 10).
>
> See the interactive Registry Toolkit health records demo for each pathway.

Module 10: Patient-Reported Data & the 21st Century

Cures Act

Goal: Understand how the 21st Century Cures Act creates participant rights to access their EHR data via APIs, and how your registry can leverage this.

The 21st Century Cures Act

The **21st Century Cures Act** (passed 2016, final rules 2020 to 2021) transformed participant data access. Its key provisions for registries:

Information Blocking Prohibition

EHR vendors and health systems are **prohibited from blocking patient access** to their electronic health information. This means:

- Participants have the right to access their complete EHR data in electronic form
- Health systems cannot charge unreasonable fees for this access
- EHR vendors must provide standardized API access

Mandatory FHIR APIs

All EHR vendors certified under ONC's 2015 Edition must now implement **FHIR R4 APIs** allowing patient access via SMART on FHIR. This means that as of 2022, essentially every major EHR system, Epic, Cerner/Oracle, Meditech, Athenahealth, must support patient-directed FHIR data access.

This is transformative for patient registries. Participants can now legally direct their EHR data to any app or registry they choose, with their treating institution required to comply.

OAuth 2.0 and Patient Authorization

How it works

OAuth 2.0 is the authorization protocol that makes patient-directed data access secure. When a participant authorizes your registry to access their EHR:

1. **Authorization Request:** Your registry redirects the participant to their EHR's authorization server
2. **Participant Authentication:** The participant logs in to their patient portal (e.g., MyChart)
3. **Consent:** The participant reviews and approves the specific data types your registry is requesting
4. **Authorization Code:** The EHR returns a short-lived authorization code to your registry
5. **Token Exchange:** Your registry exchanges the code for an access token
6. **Data Access:** Your registry uses the access token to call FHIR APIs and retrieve the participant's data

The participant can revoke this authorization at any time.

Scopes, What you can request

SMART on FHIR uses OAuth 2.0 scopes to define what data an app can access:

```
patient/Condition.read         - Patient's diagnoses
patient/Observation.read       - Lab results, vital signs
patient/MedicationRequest.read - Prescriptions
patient/Procedure.read         - Procedures
patient/DocumentReference.read - Clinical notes, C-CDA documents
```

Request only what you need. Requesting broad scopes reduces participant trust and authorization rates.

Direct from Participant: The Registry Opportunity

The Cures Act creates a direct path from participant to registry that bypasses institutional barriers:

1. Participant enrolls in your registry
2. Participant clicks "Connect my health records"
3. Participant logs in to their EHR portal and authorizes specific data sharing
4. Registry receives structured FHIR data directly, diagnoses, labs, medications
5. Data updates automatically as new information appears in the EHR

No health system agreement required. No IRB coverage of the health system. The participant is exercising their own data rights.

> **Important caveat.** While no health system agreement is needed for patient-directed access, your registry still needs IRB approval covering the collection and use of EHR data obtained through patient authorization.

Health Apps and the Patient Portal Ecosystem

Major patient portal apps now support patient-directed FHIR access:

- **Apple Health Records**, iOS users can aggregate records from thousands of institutions; data can be exported in FHIR format
- **CommonHealth (Android)**, a nonprofit, Apple-Health-style aggregator from The Commons Project. Its consumer app has effectively wound down, so treat it as historical rather than a current integration target; verify any specific aggregator is still operating before relying on it.
- **Particle Health**, Aggregates participant records from 270M+ participant records nationally via CareQuality and CommonWell networks

Some registries integrate with these aggregators rather than building direct EHR connections, significantly reducing development complexity.

Practical implementation

Using Bulk FHIR for site level data

For clinical sites that agree to participate in your registry as data contributors, **FHIR Bulk Data** allows the site to export FHIR data for all consented participants at once, rather than participant-by-participant.

This requires:

- Site participation agreement
- BAA with the health system
- IRB approval covering the health system

Testing your FHIR implementation

Before going live, test against public FHIR sandboxes:

- SMART on FHIR App Launcher
- Inferno, ONC's official FHIR testing tool
- Epic, Cerner, and Athena all provide developer sandboxes

Key resources

- ONC 21st Century Cures Act Final Rule
- SMART App Launch Framework
- Apple Health Records
- CommonHealth
- ONC Cures Act Developer Resources

← Module 9 | Module 11: Designing Questionnaires →

Network exchange: TEFCA and Individual Access Services (IAS)

Aligned with EpilepsyLive. Beyond a single participant-portal login, records can be matched across many providers at once through national networks.

- **ASTP/ONC.** The federal health-IT office formerly called ONC was renamed **ASTP/ONC** (Assistant Secretary for Technology Policy / Office of the National Coordinator) in July 2024. It administers the frameworks below, so recent materials use the new name.
- **TEFCA**, the Trusted Exchange Framework and Common Agreement sets a nationwide floor for health-data exchange. It runs through **Qualified Health Information Networks (QHINs)** that connect participating providers.
- **Individual Access Services (IAS)**, the TEFCA pathway that lets a person use an app or service of their choice to request their own records across participating providers, after identity verification. Reach is broad in theory and grows as the networks mature.
- In practice, portal login (SMART on FHIR), C-CDA download, HIPAA authorization, and TEFCA/IAS usually run *through a vendor or platform* rather than being wired directly into a registry.

EEG & imaging studies

Especially relevant for epilepsy. The actual EEG recordings (e.g., EDF) and imaging studies (DICOM) are large binary files that generally do **not** travel through C-CDA documents or the standard FHIR document flows. Report text may appear, but the studies themselves usually have to be requested separately, from the neurology or radiology department, or via an image-exchange service, often under a HIPAA authorization. Plan for these as their own acquisition pathway.

> See the Registry Toolkit health records demo, which walks through all five pathways, and the Guide glossary for the underlying terms.

Module 11: Designing Questionnaires

Goal: Design questionnaires that produce high-quality, analyzable data, and that participants actually complete.

The questionnaire design imperative

A poorly designed questionnaire produces poor data. The most common problems:

- **Too long**, Completion rates drop dramatically beyond 15 to 20 minutes
- **Ambiguous questions**, "How severe are your symptoms?" means different things to different people
- **Double-barreled questions**, "Do you have pain and fatigue?" cannot be answered with a single yes/no
- **Custom scales instead of validated instruments**, Produces uninterpretable data
- **No skip logic**, Showing irrelevant questions wastes participant time and increases abandonment

Use validated instruments wherever possible

Always search for a validated instrument before writing a custom question.

A validated instrument has been:

- Developed with participant and clinician input
- Tested for reliability (consistent results across administrations)
- Tested for validity (actually measures what it claims to measure)
- Normed against a reference population
- Accepted by journals and regulatory agencies

Key validated instrument libraries

PROMIS (NIH): Physical function, fatigue, pain, sleep, anxiety, depression, social participation
→ healthmeasures.net

EQ-5D: Generic health related quality of life; widely used in health economics
→ euroqol.org

SF-36 / RAND-36: General health status
→ rand.org/health-care/surveys_tools/mos/36-item-short-form.html

NIH Toolbox: Cognitive, emotional, motor, and sensory function
→ nihtoolbox.org

Participant-specific instruments: Many diseases have gold-standard disease specific instruments. Your SAB should identify these.

Question types and best practices

Response scales

For symptom severity, use numeric rating scales (NRS) or Likert scales, not free text.

- **NRS 0 to 10:** "On a scale of 0 to 10, how severe is your pain today?", Simple, widely understood
- **Likert (5-point):** Never / Rarely / Sometimes / Often / Always, Good for frequency
- **Visual analog scale (VAS):** Continuous line from "None" to "Worst imaginable", More sensitive but harder to implement on paper

Date questions

For onset dates, provide a structured date picker with a "year only" or "approximate" option, many participants know the year of symptom onset but not the exact date. A "don't know / unsure" option is essential.

Branching logic

Use skip logic so participants only see relevant questions:

- "Have you ever had seizures?" → If NO, skip to next section
- "What medications are you currently taking?" → If NONE, skip medication detail section

Cognitive burden

- Use plain language (aim for 6th grade reading level)
- Avoid medical jargon; define terms when needed
- Group related questions in logical sections
- Provide progress indicators on long questionnaires
- Offer "save and return later" for multi-section forms

Questionnaire development process

1. **Draft** based on scientific questions and SAB input
2. **Participant review**, share draft with 3 to 5 participants for comprehension testing
3. **Cognitive interviewing**, ask participants to "think aloud" as they answer questions
4. **Pilot test** with 20 to 50 participants; measure completion rate and time
5. **Analyze pilot data**, Are there items with very high "don't know" or skip rates? Do items perform as expected statistically?
6. **Revise** based on pilot findings
7. **Final SAB approval**

Frequency and burden management

- **Enrollment questionnaire:** Budget 20 to 30 minutes maximum for core elements
- **Annual/biannual follow up:** 10 to 15 minutes
- **Brief check-ins:** 5 minutes or less for high-frequency (quarterly or monthly) touchpoints

Shorter and more frequent is often better than long and infrequent, and yields better longitudinal data.

Key resources

- PROMIS
- EQ-5D
- NIH Toolbox
- COSMIN, Measurement instrument quality standards
- FDA PRO Guidance

← Module 10 | Module 12: Verifying Clinical Data →

Module 12: Verifying Clinical Data

Goal: Understand when and how to verify patient-reported data against clinical records, and build a data quality framework for your registry.

Patient-reported vs. clinician verified data

Most patient registries start with self reported data because it is scalable and low-cost. But for many research and regulatory purposes, verified clinical data is required.

Data type	Scalability	Cost	Research value
Participant self reported	High	Low	Good for PROs, demographics, functional status
Clinician verified	Medium	Moderate	Required for diagnosis confirmation, genetic data
Source document verified (SDV)	Low	High	Required for regulatory submissions

What typically needs verification

Diagnosis confirmation

For rare diseases, self reported diagnosis is unreliable. Many participants have the wrong diagnosis; others have multiple diagnoses and may report a different primary. For research purposes, and especially for clinical trial recruitment, **diagnosis must be verified against medical records or genetic testing**.

Methods:

- Medical record review by a study coordinator or clinician
- Genetic test report review (for genetic diseases)
- Clinician attestation (treating physician confirms diagnosis)
- Death certificate review (for mortality data)

Genetic/molecular data

Mutation or variant data should be extracted directly from clinical genetic test reports, not transcribed by participants. Participants frequently report variant details incorrectly.

Build a document upload feature so participants can upload their genetic test report, and have a study coordinator extract the variant data into standardized fields (HGVS notation, gene symbol, variant type).

Key clinical measurements

For outcomes analysis, key measurements (functional scores, lab values, imaging findings) should be extracted from the medical record or entered by the treating clinician, not relying on participant memory.

Data quality framework

Edit checks and validation rules

Build validation into your data collection forms:

- Range checks (age cannot be negative; weight cannot be 500kg)
- Logic checks (symptom onset cannot be before birth date)
- Completeness checks (flag records with missing required fields)
- Consistency checks (diagnosis date should precede treatment start date)

Missing data management

Design your analysis plan before data collection, including a missing data strategy:

- What percentage of missingness in a field triggers exclusion from analysis?
- Will you use imputation? Which methods?
- How will you handle "not applicable" vs. "unknown" vs. genuinely missing?

Duplicate detection

Rare disease communities are small. The same participant may enroll multiple times, or enroll in multiple registries. Build duplicate detection:

- Match on date of birth, sex, and diagnosis date
- Consider a unique participant identifier (with appropriate privacy protections)
- Coordinate with other registries in your disease space

Audit trail

Maintain a complete audit trail of:

- Who entered each data element
- When data was entered and modified
- What changes were made
- Source documentation for verified data

This is required for regulatory submissions and is good research practice.

Study coordinator verification workflow

For registries with clinical site participation, a typical verification workflow:

1. Participant enrolls and completes self report
2. Participant signs medical records release authorization
3. Study coordinator at clinical site receives notification
4. Coordinator reviews medical record and completes clinical data form
5. Discrepancies between patient-reported and record-sourced data are flagged
6. Data manager reviews and resolves discrepancies

7. Record is marked as "clinician verified"

Data quality metrics to track

- **Completeness rate:** % of required fields with non-missing values, by element and by time period
- **Timeliness:** Median days from enrollment to complete baseline data
- **Verification rate:** % of enrolled participants with clinician verified diagnosis
- **Inconsistency rate:** % of records with at least one logic or range check failure
- **Attrition rate:** % of participants who complete each follow up assessment

Report these metrics to your SAB quarterly.

Key resources

- AHRQ Registry User's Guide Chapter 5: Data Collection
- FDA Data Standards for Clinical Trials
- CDISC CDASH (Clinical Data Acquisition Standards Harmonization)
- TransCelerate Data Standards

← Module 11 | Module 13: Recruitment Strategies →

Part 4 · Participants

Module 13: Recruitment Strategies

Goal: Build a recruitment strategy that reaches your community effectively, with particular attention to the unique challenges of rare disease recruitment.

The rare disease recruitment challenge

For common diseases, standard recruitment channels (physician referral, hospital outreach, advertising) work reasonably well. For rare diseases, where the participant population may number in the hundreds or thousands globally, recruitment requires a fundamentally different approach.

Key principles for rare disease recruitment:

1. Participants are your best recruiters, the community recruits itself
2. Meet participants where they already are (online communities, support groups, specialty clinics)
3. Trust is everything, participants evaluate the organization behind the registry before enrolling
4. Reduce friction relentlessly, every extra step loses participants

Recruitment channels

Community and support organizations

Your own organization's community is your primary recruitment source. Leverage:

- Email newsletter to members
- Social media (Facebook groups are particularly active for rare disease communities)
- Annual conference / family day
- Participant ambassadors who share their registry participation story

Clinical site partnerships

Specialty clinics (neuromuscular centers, metabolic disease clinics, etc.) see concentrated participant populations. Work with these sites to:

- Train physicians and nurses to mention the registry at clinic visits
- Provide participant-facing materials for waiting rooms and exam rooms
- Enable direct referral from the clinic to the registry

Registries and disease networks

Many disease communities have existing participant directories, foundations, or networks. Coordinate recruitment with:

- Other disease foundations in your space
- International advocacy organizations
- Academic centers with existing participant cohorts

Social media and online communities

- Facebook participant groups for your disease
- Rare disease social networks (Inspire, RareConnect)
- Disease specific forums and Reddit communities
- Targeted social media advertising (Facebook/Instagram allow targeting by condition interest)

ClinicalTrials.gov listing

List your registry on ClinicalTrials.gov (it's free). Participants and family members actively search this database. A registry listing increases findability and establishes scientific credibility.

clinicaltrials.gov

NORD RareConnect and Global Genes

Both organizations have participant-facing platforms and can help publicize your registry to their communities.

Reducing enrollment friction

Every step between a participant hearing about your registry and completing enrollment is an opportunity for dropout. Analyze your funnel:

```
Awareness → Interest → Eligibility check → Consent → Registration → Baseline questionnaire → Complete
```

Measure dropout at each stage. Common friction points:

- Lengthy eligibility screening before consent
- Consent document that requires printing and signing
- No mobile-optimized enrollment form
- Required fields that participants don't know (exact diagnosis date, genetic mutation details)
- No "save and return later" option

Best practices:

- Enable enrollment from a mobile device
- Use eConsent with digital signature
- Make baseline questionnaire completable in 20 minutes or less
- Send follow up reminders at 24 hours, 1 week, and 1 month for incomplete enrollments
- Offer enrollment assistance (phone support, email help desk)

Retention strategies

Enrollment is only the beginning. Long term registry value depends on participant retention for follow up data.

- **Regular communication:** Quarterly or annual newsletters reporting registry progress and findings
- **Results return:** Share aggregate findings back to participants, what is the registry learning?

- **Feedback loops:** Surveys about registry experience; act on feedback
- **Acknowledgment:** Recognize participants in publications (with consent)
- **Reminders:** Automated follow up reminders at scheduled data collection timepoints
- **Easy withdrawal:** Make it simple to update preferences or withdraw, trust is maintained by respecting autonomy

Tracking recruitment metrics

Metric	Description
Monthly new enrollments	Track trend and seasonality
Enrollment by source	Which channels drive the most registrations
Completion rate	% who complete full baseline questionnaire
Retention at 1 year	% of enrolled participants with at least one follow up
Geographic distribution	Are you reaching participants nationally / internationally
Demographic diversity	Are underrepresented communities being reached

Key resources

- NORD Rare Disease Participant Recruitment Resources
- Global Genes Rare Advocacy Tools
- RareConnect
- Inspire Participant Community
- ClinicalTrials.gov Registration Guide
- PCORI Engagement in Research Resources

← Module 12 | Module 14: Study Coordinators →

Module 14: Working with Study Coordinators

Goal: Understand the role of study coordinators in multisite registries and how to build effective relationships with clinical sites.

What study coordinators do

Study coordinators (also called research coordinators or clinical research coordinators) are the backbone of clinical site participation in registries. They:

- Identify and approach eligible participants at their site
- Obtain informed consent
- Complete or verify clinical data entry
- Collect and process biosamples (if applicable)
- Manage site regulatory documentation (IRB approvals, delegation logs)
- Communicate with your registry team about site level issues

For a rare disease registry, your relationship with study coordinators at key clinical centers is as important as your relationship with the physicians.

Setting up clinical sites

Site agreements

Each participating clinical site needs:

- **Site agreement / participating site agreement (PSA):** Defines the site's responsibilities, your organization's responsibilities, compensation, and data handling
- **IRB authorization:** Either reliance on your central IRB or local IRB approval
- **HIPAA BAA:** If the site is sending identifiable data

Coordinator training

Build a training program for new coordinators:

- Registry overview and scientific purpose
- Eligibility criteria and how to identify participants
- Consent process and documentation
- Data entry procedures (platform walkthrough)
- Sample collection and shipping (if applicable)
- Regulatory requirements

Provide: training slides, written SOPs, a quick reference guide, and a direct contact for questions.

Coordinator compensation

Study coordinators' time should be compensated. Typical models:

- **Per-participant payment:** A set amount for each enrolled and verified participant ($100, $500 depending on data complexity)
- **Site grant:** Annual site support payment covering coordinator time

Keeping coordinators engaged

Sites with an enthusiastic coordinator enroll dramatically more participants than sites with a disengaged one. Invest in coordinator relationships:

- Personal outreach (email, phone calls), not just automated system messages
- Coordinator-specific newsletter with recruitment tips and registry updates
- Recognition of high-enrolling sites
- Regular coordinator calls (quarterly) to share learnings and address problems
- Site visit support (virtual or in-person) for sites that are struggling

Key resources

- ACRP: Association of Clinical Research Professionals
- SOCRA: Society of Clinical Research Associates
- ACRP Coordinator Training Resources

← Module 13 | Module 15: Working with Industry →

Part 5 · Partnerships

Module 15: Working with Industry Sponsors

Goal: Build productive partnerships with pharmaceutical and biotech companies while protecting participant interests, data integrity, and your organization's independence.

The opportunity and the risk

Patient registry data is extremely valuable to the pharmaceutical industry. A disease registry with natural history data, verified diagnoses, stored samples, and an engaged community can:

- Serve as a comparator arm for single arm clinical trials
- Provide pre-screened participants for trial recruitment
- Support regulatory submissions as real world evidence
- Inform trial endpoint selection and outcome measure development
- Enable postmarket surveillance studies

This creates leverage for your organization. You hold something valuable. The question is how to structure partnerships that advance science and benefit participants, without compromising your independence or your participants' trust.

Types of industry engagement

Data licensing

A company pays for access to your registry data, either aggregate analyses, deidentified individual records, or (with additional consent) identified data.

Key protections:

- You retain data ownership; the license is time-limited and purpose-limited
- Use restrictions must be explicit: no reidentification, no secondary commercial use, no data sale
- Define the expiration and data destruction obligations
- Reserve the right to publish independent analyses

Natural history partnership

A company funds registry operations in exchange for being able to use your natural history data to support their regulatory submission.

Key protections:

- This is a research partnership, not a sponsorship, maintain scientific independence
- Your SAB (not the company) controls data element selection and analysis plans
- You retain publication rights
- Any FDA meetings where registry data is presented should include advocacy organization representation

Recruitment partnership

A company compensates your registry for identifying and referring eligible participants to their clinical trial.

Key protections:

- Referral must be genuine and patient-centered, never pressure participants to enroll in trials
- Compensation should be transparent and fair market value
- Participants must understand the distinction between registry participation and trial enrollment
- Your organization should never receive per-enrollment payment in ways that create incentive to over-refer

Sponsored registry

A company funds the creation of a registry from scratch, with the explicit goal of generating data for their regulatory program.

This arrangement requires the most careful structuring. Ensure:

- Governance remains with the advocacy organization
- An independent SAB controls scientific decisions
- All participants are informed of industry funding in the consent document
- A firewall exists between your organization's advocacy activities and the sponsored registry

Negotiating the data use agreement

The DUA is the most important document in an industry partnership. Key terms to negotiate:

Term	Your position
Data ownership	Your organization owns the data, always
Publication rights	You retain the right to publish; company may have limited review period (30 to 60 days) for confidentiality concerns only, not to suppress findings
Reidentification prohibition	Absolute prohibition
Secondary use	Prohibited without separate agreement
Audit rights	Right to audit data use compliance
Term and termination	Define what happens to data upon agreement expiration
Intellectual property	Data generated by your registry is your IP; discoveries made using your data by the company belong to the company (define clearly)
Transparency	Require disclosure of the partnership to FDA and in publications

Protecting participant trust

Your community's trust is your most valuable asset, worth more long term than any individual industry partnership. Maintain it by:

- Being transparent with participants about industry partnerships in your consent

document
- Publishing an annual transparency report disclosing funding sources
- Never allowing industry partners to contact participants directly
- Maintaining independent scientific governance
- Being willing to decline partnerships that conflict with participant interests

Key resources

- PhRMA Principles on Conduct of Clinical Trials
- NORD Participant Advocacy Summit Resources
- EFPIA Participant Engagement Framework
- FDA Guidance: Natural History Data for Regulatory Submissions

← Module 14 | Module 16: Working with Academia →

Module 16: Working with Academia

Goal: Build productive research collaborations with academic institutions that advance science, produce publications, and respect advocacy organization interests.

Why academic partnerships matter

Academic collaborators bring:

- **Scientific credibility**, peer reviewed publications require academic coauthors
- **IRB access**, academic institutions can provide IRB coverage for registry research
- **Technical resources**, biostatisticians, bioinformaticians, data scientists
- **Funding access**, NIH grants, foundation grants, and clinical trial networks
- **Clinical expertise**, disease specialists, rare disease centers of excellence

Structuring the relationship

Data access agreements

Academic researchers accessing your registry need a signed Data Use Agreement covering:

- Permitted analyses and publications
- Prohibition on reidentification
- Data security requirements
- Publication notification and co-authorship expectations
- Data return and destruction upon project completion

IRB coordination

Define who holds IRB coverage for registry-linked research:

- Your central IRB covers the registry; academic collaborators add a linked study under their own IRB
- Academic institution provides a master IRB agreement covering all affiliated analyses
- Consider using a single central IRB (CIRB or NCI CIRB) for multisite academic networks

Authorship policy

Establish your authorship policy before the first paper:

- Registry team members are coauthors on all publications using registry data
- A participant coauthor is included on all publications (identify and support interested participants)
- Acknowledge registry participants in every paper
- First and last authorship for key papers goes to whom?

Funding considerations

NIH grants

Many registries are supported by NIH funding:

- **R01:** Standard research grant; advocacy organizations can be co-investigators with an academic lead PI
- **U01/U34:** Cooperative agreements; common for multisite registry networks
- **PAR funding opportunities:** NIH periodically issues targeted PARs for rare disease registries

Key resource: NIH Research Portfolio Online Reporting Tools (RePORTER)

PCORI

PCORI funds patient-centered outcomes research with a strong emphasis on participant engagement. Advocacy organizations as principal investigators are explicitly encouraged.

pcori.org/funding

Foundation and advocacy funding

Many disease foundations provide seed funding for registry development. Even modest funding ($50K, $250K) can launch a registry that then attracts larger academic or industry support.

Building a productive collaboration

- Define deliverables and timelines in writing at the start
- Hold regular (monthly or quarterly) working group calls
- Share data access promptly once agreements are in place
- Be explicit about publication timelines, academic timelines can be very slow
- Provide administrative support for grant applications that use your registry as a resource

Key resources

- NIH Office of Rare Diseases Research
- PCORI
- CTSA Hub Network, Academic clinical research network
- Rare Disease Clinical Research Network (RDCRN)

← Module 15 | Module 17: Analyzing Registry Data →

Part 6 · Using Your Data

Module 17: Analyzing Registry Data

Goal: Understand the analytic approaches used in registry research and how to design your registry for analysis from the start.

Design for analysis

The most common registry analysis failure: data was collected without a prespecified analysis plan. The result is either a fishing expedition for significant findings (p-hacking) or a dataset that cannot answer the questions it was supposed to address.

Pre-specify your primary analyses before collecting data. This means defining:

- Primary outcome(s)
- Primary exposure(s) or comparison groups
- Covariates to adjust for
- Statistical methods
- Missing data handling approach

Your SAB should review and approve the analysis plan.

Common registry analyses

Disease characterization

- **Prevalence / incidence estimation**, How common is the disease? How many new cases per year?
- **Symptom frequency and distribution**, What proportion of participants have each clinical feature?
- **Demographic analysis**, Age at onset, sex distribution, time to diagnosis
- **Genotype phenotype analysis**, Which genetic variants are associated with which clinical features?

Natural history

- **Disease progression modeling**, How do symptoms change over time?
- **Survival analysis (time-to-event)**, Time to disease milestones, time to treatment initiation
- **Longitudinal mixed models**, Track individual trajectories and population level trends

Treatment patterns and outcomes

- **Treatment use**, Which treatments are participants receiving, and for how long?
- **Comparative effectiveness**, Do participants on treatment A have better outcomes than participants on treatment B? (Requires careful confounding adjustment)

Statistical considerations for rare diseases

Small sample sizes

Most rare disease registries operate with small N. Implications:

- Pre-specify analyses to avoid overfitting
- Use Bayesian methods where appropriate (allow incorporation of prior knowledge)
- Be cautious about subgroup analyses, report effect sizes and confidence intervals, not just p-values
- Power analyses should be conducted before data collection, not after

Missing data

Missing data is ubiquitous in observational registries. Approaches:

- **Complete case analysis**, Easiest; valid only when data is missing completely at random (MCAR), which is rarely true
- **Multiple imputation**, Appropriate for data missing at random (MAR); widely used
- **Sensitivity analysis**, Test how your conclusions change under different missing data assumptions

Confounding

Registry data is observational, participants are not randomized to treatments or exposures. Confounding is a major threat to causal inference. Methods:

- **Propensity score matching or weighting**, Balance treatment groups on observed covariates
- **Instrumental variable analysis**, For when confounding by indication is severe
- **Target trial emulation**, Explicitly design the analysis as if it were a randomized trial

Tools and platforms

Tool	Use
R (CRAN)	General statistical analysis; extensive rare disease packages
Python (pandas, statsmodels)	Data manipulation and analysis
SAS	Industry/regulatory standard; expensive
ATLAS (OHDSI)	Cohort analysis on OMOP data
REDCap built-in reports	Basic frequency tables and exports
Tableau / Power BI	Data visualization and dashboards

Key resources

- AHRQ Registry User's Guide Chapter 6: Analysis
- Book of OHDSI, Characterization Chapter
- FDA Guidance on Natural History Data
- STROBE Statement, Reporting observational studies

← Module 16 | Module 18: Data Sharing →

> **From the author.** The analyst mindset behind this module, planning an analysis, exploring and verifying your data, and communicating results clearly, is covered in depth

> in Dr. Boyce's book, **Ten Habits of Great Data Analysts**, available on Leanpub as PDF, iPad, and Kindle. See also the companion site in the Learn section.

Module 18: Data Sharing

Goal: Develop a data sharing framework that maximizes the scientific value of your registry while protecting participant privacy and your organization's interests.

Why data sharing is important

Rare disease registries hold irreplaceable data. Hoarding it is not just bad for science, it is at odds with the fundamental mission of advocacy. Data that could benefit participants should be shared.

At the same time, sharing without governance, without consent coverage, security protections, and use restrictions, can harm participants and expose your organization to legal and reputational risk.

The goal is a principled sharing framework: maximally open within appropriate protections.

Data sharing frameworks

Federated analysis (no data transfer)

Researchers submit an analysis code that runs on your data behind your firewall; only aggregate results are returned. No individual-level data leaves your system.

Tools: OHDSI distributed network studies, PCORnet, TriNetX
Best for: Large queries where individual data transfer would be impractical; reduces privacy risk

Deidentified data sharing

Individual-level records, stripped of identifying information per HIPAA Safe Harbor or Expert Determination standards, are shared under a data use agreement.

Best for: Most registry research uses; wide applicability
Limitations: Some reidentification risk with rare diseases and rare variants; "deidentified" rare disease data is not as anonymous as common disease data

Controlled access (identified data)

Identified or potentially re-identifiable data shared only with approved researchers under strict DUA, subject to Data Access Committee review.

Best for: Genomic data; linkage studies; longitudinal matching
Implementation: Use access control systems like dbGaP or GA4GH Passport/Visa

Data repositories and sharing platforms

NIH repositories

- **dbGaP (Database of Genotypes and Phenotypes):** NIH's primary controlled access repository for genomic and phenotypic data from human studies. NIH funded studies are increasingly required to deposit data here.
  ncbi.nlm.nih.gov/gap
- **NCBI BioProject / BioSample:** For genomic sequence data
- **ClinVar:** For variant-disease assertions

Global repositories

- **EGA (European Genome-phenome Archive):** European equivalent of dbGaP; GDPR-native
  ega-archive.org

Open data platforms

- **Synapse (Sage Bionetworks):** Supports both open and controlled access sharing; widely used by participant led research
  synapse.org
- **Zenodo:** General open data repository; appropriate for fully deidentified summary data
  zenodo.org

Data sharing agreements

Every data sharing arrangement needs a Data Use Agreement (DUA) covering:

- Permitted uses of the data
- Prohibition on reidentification
- Data security requirements (encryption at rest and in transit, access controls)
- Prohibition on data redistribution without separate approval
- Publication notification requirements
- Data retention period and destruction upon expiration
- Reporting requirements (annual reports to your DAC)

FAIR data principles

FAIR stands for Findable, Accessible, Interoperable, and Reusable, a framework for maximizing the value of shared scientific data.

Principle	What it means for your registry
Findable	Data is registered in a searchable repository; metadata is published
Accessible	Clear process for requesting access; metadata accessible even when data requires controlled access
Interoperable	Uses standard vocabularies (SNOMED, LOINC, HPO) and formats (FHIR, OMOP)
Reusable	Data use license is clear; provenance is documented

go-fair.org

Key resources

- NIH Data Sharing Policy
- dbGaP Submission Guide
- GA4GH Data Access Framework
- Synapse Data Sharing Platform
- FAIR Principles
- AHRQ Registry User's Guide Chapter 7: Dissemination

← Module 17 | Module 19: Publications →

Module 19: Publications

Goal: Turn your registry data into peer reviewed publications that establish your organization as a scientific leader and advance the field.

Why publish?

Publications are how science moves. A registry that collects data but never publishes it does not advance the field, does not attract researchers or industry partners, and cannot demonstrate value to funders.

Strategic reasons to publish:

- Establish scientific credibility for your organization
- Demonstrate registry data quality and utility
- Attract academic collaborators and industry partners
- Support regulatory submissions with published natural history data
- Fulfill obligations to participants ("we will use your data to advance science")
- Build the evidence base for clinical trial endpoints

Types of registry publications

Registry description / methods paper

Describes the registry's design, governance, data elements, and enrolled population. This is typically the first publication, it establishes the scientific record for the registry and enables future citation.

Target journals: Rare diseases journals, clinical informatics journals, disease specific specialty journals
When to publish: After first 50 to 100 participants are enrolled and baseline data quality is established

Natural history paper

Describes the disease course, symptom prevalence, progression, and outcomes in your participant population. The core scientific product of a natural history registry.

Value: This paper becomes a reference for every clinical trial in your disease space.

Genotype phenotype paper

Links specific genetic variants to clinical features. Enormously valuable for diagnosis, prognosis, and drug target identification.

Requires: Genetic data (verified molecular diagnoses) + HPO coded phenotype data

Patient-reported outcomes paper

Describes participant burden, quality of life, and unmet needs from the participant perspective. Increasingly important for FDA participant-focused drug development.

Treatment patterns / real world evidence paper

Describes which treatments participants are receiving and their outcomes in real world practice.

Reporting standards

Use established reporting guidelines, reviewers and editors expect them:

- **STROBE**, Strengthening the Reporting of Observational Studies in Epidemiology
  strobe-statement.org
- **RECORD**, Reporting of Studies Conducted Using Observational Routinely-Collected Data (extension of STROBE for registry/EHR-based studies)
  record-statement.org
- **CONSORT**, For any controlled component
  consort-statement.org
- **GRIPS**, Genetic Risk Prediction Studies
  grips-statement.org

Participant authorship and acknowledgment

Every publication from your registry should:

1. **Acknowledge participants**, "The authors gratefully acknowledge the participants and families who contributed data to the [registry name]."
2. **Include participant coauthors** where possible, participants who contributed meaningfully to study design, questionnaire development, or interpretation of findings should be considered for co-authorship per ICMJE criteria
3. **Disclose funding sources** including industry partnerships
4. **Disclose conflicts of interest** for all authors

ICMJE authorship criteria: icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

Open access

Make your publications open access whenever possible. Your participants, and the broader community, deserve to read what was learned from their data. Most funding agencies (NIH, PCORI) now require open access publication.

NIH PubMed Central deposit: Required within 12 months for NIH funded research
publicaccess.nih.gov

Plan S / cOAlition S: International open access mandate increasingly adopted by funders

coalition-s.org

Target journals for rare disease registry publications

Journal	Focus
Orphanet Journal of Rare Diseases	Rare diseases; open access
Genetics in Medicine	Genetic disease; high impact
American Journal of Human Genetics	Genetics; very high impact
JAMIA (Journal of the American Medical Informatics Association)	Informatics, registry methods
PLOS ONE / PLOS Genetics	Open access; broad scope
Disease specific specialty journals	Highest relevance for clinical audience

Key resources

- STROBE Statement
- RECORD Statement
- ICMJE Authorship Guidelines
- NIH Public Access Policy
- Orphanet Journal of Rare Diseases
- EQUATOR Network, Reporting guidelines

← Module 18 | Reference Card →