Photos used with permission from Avinash Sharma

ARGO GRANT WRITING COURSE · WEEK 4 (JULY 15)

Data Sources & Feasibility

Mengmeng (Margaret) Du, ScD

Associate Attending Epidemiologist

Associate Professor

Memorial Sloan Kettering Cancer Center

Weill Cornell Medicine

Slide deck created with the assistance of [Gamma]

Disclosures: No conflicts of interest beyond funding from the following organizations

National Cancer Institute

National Human Genome Research Institute

National Institute of Environmental Health Sciences

National Center for Advancing Translational Sciences

Per and Astrid Heidenreich Family Foundation

David M. Rubenstein Center for Pancreatic Cancer Research

The Society of MSK

What We'll Cover Today

01

Identifying Appropriate Data Sources

Where your data comes from and how to justify it

02

Defining Key Variables & Outcomes

Primary outcomes, covariates, and measurement clarity

03

Assessing Feasibility

Accrual, resources, personnel, and timelines

04

Avoiding Common Feasibility Pitfalls

Mistakes that sink pilot proposals and how to prevent them

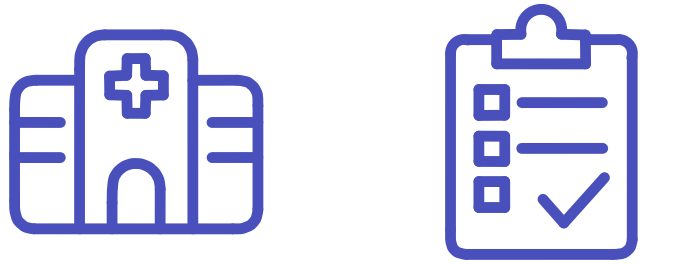
SECTION 1 OF 4

Identifying Appropriate Data Sources

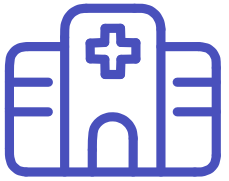
Our scientific question and available resources determine the most appropriate data source

Our data source shapes everything, including methods, sample size, feasibility, and our credibility with reviewers

Important: Our data capture who shows up or is approached in a study. In LMICs, this is often NOT THE SAME as the general population



Facility Records



In many LMICs, **paper-based records** or **community networks** are the primary data source. These are rich in local context but require careful data extraction and quality checks.

- Medical records
- Patient registries
- Lab records

✓ Strengths	✗ Limitations
Existing infrastructure and established data collection routines	Incomplete or inconsistently maintained records
Local institutional buy-in and staff familiarity	Staff turnover disrupts data quality and continuity
Captures real-world care patterns in resource-limited settings	Transition from paper to digital systems can create gaps and duplication

Primary Data Collection

Requires careful instrument design/validation, ethical approval, and, most critically, **investing in training local field staff** to ensure data consistency and cultural/regional appropriateness.

- Surveys
- Interviews
- Direct measurements (e.g., anthropometrics, biospecimens)

✓ Strengths	✗ Limitations
Purpose-built for your research question; fill gaps where routine data is absent	High cost, time-intensive, and logistically complex in low-resource environments
Patient or community engagement can strengthen trust and response rates	Field staff attrition and inconsistent training can introduce bias
Allows culturally adapted/validated instruments and local language versions	Ethical approval processes may be slower or less standardized across multiple sites

Public & National Datasets

Credible and often free but may lack granularity or have significant lag time between collection and release.

- Demographic and Health Surveys (DHS)
- National census data
- WHO repositories
- LMIC-specific surveys like UNICEF Multiple Indicator Cluster Surveys (MICS) and WHO Service Availability and Readiness Assessment (SARA)
- Genomics: database of Genotypes and Phenotypes (dbGaP)

✓ Strengths	✗ Limitations
Often free, publicly accessible, and widely recognized by reviewers and funders	Significant lag time (often 3–5 years) between data collection and public release
Nationally representative samples increase generalizability	Limited sub-national or facility-level granularity
Standardized methodology enables cross-country comparisons	Coverage gaps in fragile states, conflict-affected regions, or under resourced/underrepresented regions

Partner & Consortium Data

Can help expand population diversity and generalizability. Consortium data requires harmonized protocols and data-sharing agreements across institutions.

- **Multi-country networks:** ARGO, Men of African Descent and Carcinoma of the Prostate (MADCaP) Consortium, International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH), Human Heredity and Health in Africa (H3Africa), South-South Cooperation

✓ Strengths	✗ Limitations
Larger, more diverse samples strengthen statistical power and generalizability	Data-sharing agreements and IRB harmonization across countries are time-consuming
Established networks signal credibility and research infrastructure to funders	Protocol differences between sites can introduce heterogeneity
Promotes equitable authorship and capacity building across partner institutions	Power imbalances between high-income and low-income partners may affect data access and ownership

SECTION 2 OF 4

Defining Key Variables & Outcomes

Be specific because reviewers WILL notice!

Primary Outcome

The one variable your study is powered to detect

Secondary Outcomes

Supportive measures: limit to 2 to 3

Covariates

Define how you'll measure and adjust for confounders

Operationalize Everything with Specifics

"Hypertension" = SBP ≥140 mmHg on two readings

Vague outcomes are a top reason proposals get flagged for scientific weakness.

Assessing Feasibility

Feasibility isn't one question, it's four! Each dimension carries unique structural challenges that reviewers expect you to anticipate and acknowledge explicitly

📋 Accrual

Can we realistically recruit enough participants?

💻 Resources

What already exists vs. what must be procured (from where/how?)
Do we have the right tools in the right environment?

👥 Personnel

Do we have the right team with the right expertise from day one?

📅 Timelines

What are the major milestones/metrics of success, and can we achieve them within the project timeframe?

⚠️ **Feasibility gaps are the leading cause of delayed or underpowered studies.** Reviewers from funding agencies, e.g., NIH Fogarty, Wellcome Trust, and European and Developing Countries Clinical Trials Partnership (EDCTP), look for explicit mitigation strategies and not just acknowledgment of challenges.

📋 Accrual

Can we realistically recruit enough participants?

Show historical enrollment, participation, and acceptability data from your site or population

- Account for anticipated **seasonal patient volume shifts** (e.g., wet-dry season, harvest seasons) and **migration patterns** that reduce follow-up retention
- Rural-to-urban migration can cause significant attrition mid-study
- Consider participant incentives to help increase participation and retention

💻 Resources

Do we have the right tools in the right environment?

Highlight existing capacity-building (infrastructure), partnerships, and tools that make us uniquely well-suited to complete the project

Address any concerns with existing infrastructure and list what exists vs. what must be procured and from where and how

- Frequent internet/power outages: off-line data collection/back-up, backup generators, cold-chain contingencies
- Limited or shared lab capacity
- Supply chain delays for reagents and consumables that can stretch weeks or months.

Personnel

Do we have the right team with the right expertise from day one?

Personnel scrutiny goes beyond retention; reviewers actively assess whether our team has the expertise to execute and not just fill roles. They look for:

- **Demonstrated experience** with your research context, tools, methods, and constraints
- **Local co-investigators** with deep community ties and institutional knowledge, not just token partnerships
- **Clear roles and qualifications** for every named team member, tied directly to study tasks
- **Evidence of prior successful collaboration** between PI, co-investigators, and site teams

Beyond qualifications, can also be helpful to address career development plans (if training grant) and build in structured training plans for research coordinators and data collectors

Timelines

What are the major milestones/metrics of success, and can we achieve them within the project time frame?

- Multi-country regulatory approvals are rarely fast. **IRB and ethics clearances in Nigeria, Tanzania, and similar settings can take 6+ months and** sometimes longer if amendments are required

- Map every milestone with realistic buffers: regulatory review, site-level approvals, permits, authority sign-offs

- Clearly describe each milestone

- Reviewers look for logical sequencing and realistic buffers. Ignoring IRB review, important milestones, potential issues (e.g., seasonal patient volume), or is overly ambitious will raise red flags

Building a Realistic Timeline: Is your study doable within the grant time frame?

Months 1–3

Approvals & Study Setup

- IRB & ethics approvals
- REDCap/database setup
- Staff training, certification, and standard operating protocols

Months 4–6

Recruitment & Early Enrollment

- Patient/Community outreach
- Participant screening & enrollment

Months 7–9

Continue Enrollment & Interim QC

- Enrollment & data collection
- Monitor participation/retention
- Interim data QC

Months 10–12

Complete Enrollment

- Complete target enrollment
- Start follow-up if appropriate

Months 13–15

Data Quality Assurance

- Data quality audit
- Protocol compliance review
- Resolve data issues

Months 16–18

Complete Follow-up & Database Lock

- Complete participant follow-up
- Resolve data queries
- Lock study database
- Finalize statistical analysis plan

Months 19–21

Analysis Planning & Initial Analysis

- Conduct primary analyses
- Prepare results tables

Months 22–24

Final Analysis & Dissemination

- Complete analyses
- Submit manuscript
- Disseminate findings

How to Format a Timeline When Space is Limited?

STUDY TIMELINE

		Year 1				Year 2				Year 3				Year 4			
		Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4
Ongoing CRC case accrual to the ARGO biobank and clinical database in Nigeria*																	
Aim 1	Selection of population-based controls in HDSS database																
	1,200 population-based controls recruited in Nigeria																
	QC of questionnaire data																
	Analysis of risk factor associations																
Aim 2	Tumor and matched blood samples sent to MSK for sequencing (~30 patients/quarter)																
	MSK-IMPACT sequencing with QC and mutation calling (including MSIsensor microsatellite instability detection)																
	Analysis of somatic and germline mutation profiles																
Final data analysis and manuscript preparation																	

Common Feasibility Pitfalls in Pilot Proposals

⚠ Top scientific reasons pilot proposals fail review:

- **Overly ambitious aims**
- **Undefined success criteria**
- **Unclear how pilot will lead to or inform a larger study**

A pilot study without achievable, measurable benchmarks signals poor planning

For every major milestone, including our final outcome, define explicit, quantitative success criteria. **What determines if our pilot can progress to a larger project**? For example:

- **Recruitment target by Month 6:** e.g., ≥50% of total enrollment goal achieved
- **Participation rate:** e.g., ≥60% of approached participants consented and enrolled
- **Data quality benchmark:** e.g., <5% missing data fields per case report form at each quarterly QC review
- **Protocol adherence:** e.g., ≥90% of visits conducted within the allowed visit window

But don’t reserve success metrics only for the final outcome. Link each milestone with a success metric to complete our feasibility argument

What Else Can Go Wrong and How to Fix It

Overestimating Recruitment Rates or Quality Of Existing Data

Assume 60–70% of your ideal. Account for seasonal recruitment patterns, competing health priorities (e.g., vaccination campaigns), and community-level gatekeeping. Cite your site's historical enrollment data and build in buffer months. Build in ample time for data QC/cleaning of existing data.

Ignoring Data Quality & Infrastructure Concerns

Many LMIC sites rely on **paper-based systems** with limited IT support. Describe your paper or electronic questionnaire design, double-entry protocols, QC checklists, and how discrepancies will be resolved. If using REDCap offline mode, explain sync procedures and who manages data quality locally. Describe contingency plan for internet/power outages.

Underestimating Regulatory Timelines

Multi-country approvals are the norm, not the exception. IRB and national ethics approvals can each take **3–6 months** and must often run sequentially, not in parallel. Build full regulatory timelines into your timeline from day one and identify a local regulatory liaison at each site.

Lack of appropriate expertise

Highlight your expertise for the proposed project and institutional capacity. Clearly articulate how the pilot can help build local infrastructure and provide data to justify a larger project. Name co-investigators with substantive expertise (e.g., clinician, biostatistician), a role-to-qualification mapping that justifies each member's function, and evidence of successful collaboration (e.g., shared publications, completed studies, training grants).

Key Takeaways

1 Know your data source before you write

Justify why it's appropriate, accessible, and sufficient for your research question.

2 Define outcomes with precision

Operational definitions signal scientific rigor and make your analysis plan credible.

3 Feasibility is a story we have to sell

Weave accrual, resources, personnel, and timeline into a coherent narrative.

4 Anticipate reviewer concerns

Address the pitfalls proactively, which shows experience and builds trust.

Next session: Week 5 (July 22): Statistics & Analysis [Mithat Gonen]