

2

The Nigerian Health Data Ecosystem

A Map of the Terrain

Before a single line of analysis code is written, a Nigerian health data scientist must know the landscape of data sources available, because each has its own unit of observation, reporting frequency, coverage, and known limitations. This chapter maps that terrain, distinguishing routine health information systems from periodic surveys, disease-specific surveillance platforms, and primary research data.

Data source	Unit of observation	Typical frequency	Primary custodian
DHIS2 / National Health Management Information System (NHMIS)	Health facility (aggregated)	Monthly	Federal/State Ministries of Health
Nigeria Demographic and Health Survey (NDHS)	Household / individual	~Every 5 years	National Population Commission, ICF
Nigeria Multiple Indicator Cluster Survey (MICS)	Household / individual	~Every 4-5 years	National Bureau of Statistics, UNICEF
NCDC surveillance (SORMAS, IDSR)	Case / event	Continuous (weekly)	Nigeria Centre for Disease Control
Global Fund malaria M&E data (RDT/microscopy/PCR registers)	Health facility / patient test	Monthly	National Malaria Elimination Programme, State programmes
Institutional/clinical cohort studies (e.g., UCH, teaching hospitals)	Patient / participant	Study-specific	University research units, hospitals
Vital registration (birth/death)	Individual event	Continuous, under-reported	National Population Commission

Routine Health Information Systems

DHIS2 and NHMIS

DHIS2 is the software platform underlying Nigeria's NHMIS, aggregating monthly indicator counts (for example, confirmed malaria cases, antenatal first visits, or DPT3 immunisations) from facility level up through LGA, state, and national levels. Its great strength is coverage and timeliness relative to surveys; its great weakness is that it inherits every error made at the point of paper or electronic entry in the facility, and it reports counts, not denominators drawn from an independent population frame, which means indicators such as coverage rates depend on population projections that can drift substantially from the true catchment population, especially in areas with displacement or rapid urban growth.

A recurring, well-documented pattern in Nigerian malaria data is that the test positivity rate calculated directly from monthly HMIS summary reports tends to run higher than the TPR recomputed from source diagnostic registers and laboratory archives at the same facilities. This gap is not random noise; it typically reflects incomplete transcription of negative test results, selective reporting emphasis on confirmed positive cases, and monthly aggregation choices. A data scientist working with HMIS/DHIS2 exports should therefore always ask: has this indicator been validated against a source register (a Lot Quality Assurance Sampling or Data Quality Assessment exercise), and over what date range?

Household and Population-Based Surveys

NDHS and MICS

Where HMIS captures what happens inside the formal health system, the NDHS and MICS are designed to describe the whole population — including those who never visit a facility — through

a stratified, multi-stage cluster sample of households. These surveys are the backbone of Nigeria's estimates for indicators such as under-five mortality, stunting prevalence, contraceptive use, and HIV prevalence, and they are essential for triangulating against HMIS-derived estimates. Their limitations are the mirror image of HMIS: they are infrequent (a five-year gap between rounds means a survey estimate can be badly out of date for a fast-moving indicator), and their cluster design means that a naive analysis that ignores sampling weights and clustering will produce standard errors that are too small and can lead to false statistical significance — a point developed formally in Chapter 7.

Disease Surveillance Platforms

The Nigeria Centre for Disease Control (NCDC) operates the Integrated Disease Surveillance and Response (IDSR) system and the Surveillance Outbreak Response Management and Analysis System (SORMAS) for case-based reporting of priority diseases such as Lassa fever, cholera, measles, and mpox. Unlike aggregate HMIS counts, these systems capture line-listed, case-level data (in principle allowing analysis of case characteristics, geographic clustering, and time-to-detection), but they are also the systems most sensitive to reporting fatigue during a prolonged outbreak and to inconsistent case definitions across states.

Neglected tropical disease (NTD) surveillance — for example, for cutaneous and visceral leishmaniasis — sits somewhere between the routine and the research categories: baseline mapping exercises produce facility- or LGA-level case counts that can be modelled geographically (Chapter 14) and temporally (Chapter 10), but coverage is typically far patchier than for malaria, because diagnostic capacity for NTDs is concentrated in a small number of referral centres.

Primary Research and Cohort Data

Much of the most methodologically rich Nigerian health data science happens on purpose-built research datasets: longitudinal adolescent health cohorts tracking mental health and risk behaviours across survey waves, respondent-driven samples (RDS) of key populations such as female sex workers, and clinical studies conducted in partnership with teaching hospitals — for instance, dermatoglyphic (fingerprint pattern) studies comparing children affected by a congenital condition against unaffected controls. These datasets typically have small-to-moderate sample sizes, rich individual-level covariates, and sampling designs (respondent-driven sampling in particular) that require specialised estimators rather than standard survey-weighting approaches; RDS analysis is treated in Chapter 7.

Triangulation

Using Multiple Sources Together

The mature practice of Nigerian health data science rarely relies on a single source. A robust analysis of, say, adolescent depression risk will draw on a cohort survey for individual-level modelling (Chapters 8-9) while checking prevalence against MICS; a malaria elimination programme will triangulate HMIS trends against Global Fund diagnostic registers and, where available, community prevalence surveys. This chapter's central lesson, carried forward through the rest of the book, is that the choice of data source is itself a methodological decision with consequences as large as the choice of statistical model.

Worked Example

Triangulating Three Sources for a Single Estimate

A State Malaria Elimination Programme wants a single, defensible estimate of malaria test positivity for the 2024 rainy season. Three

candidate figures are available: the HMIS/DHIS2 summary report (68%), the Global Fund diagnostic register audit for a sample of 40 facilities (59%), and a community-based prevalence survey using microscopy in the same LGAs (41%, but measuring prevalence in the general population rather than positivity among those presenting with fever, a materially different quantity — see Section 3.4 on the relationship between prevalence and predictive value). The correct response is not to average these three numbers, which measure different things at different stages of the care-seeking pathway, but to report each with its definition made explicit: HMIS-derived TPR among tested febrile attendees (68%, likely inflated per Section 2.2), register-audited TPR among tested febrile attendees at the sampled facilities (59%, treated as the better estimate of true clinic-level positivity), and community prevalence among the general population regardless of care-seeking (41%, a lower and conceptually distinct figure reflecting the fact that not everyone with parasitaemia seeks care or presents with fever). Presenting all three, correctly labelled, is more useful — and more honest — to a technical working group than collapsing them into one number.

Practical Data Access

Working with DHIS2 Exports and Survey Microdata

In practice, DHIS2 data typically arrives to an analyst as a wide-format CSV or Excel export, with one row per facility/period combination and one column per indicator, often requiring reshaping into a long (tidy) format before analysis; NDHS/MICS microdata typically arrives as separate individual-, household-, and birth-recode files in a statistical software format (.dta, .sav) that must be merged on shared household and cluster identifiers. The code below demonstrates the essential reshaping and merging operations a Nigerian health data scientist performs routinely.

R example

```
# R: reshaping a wide DHIS2 export into tidy long format, and merging NDHS files
library(tidyr); library(dplyr); library(haven)

# --- DHIS2: wide to long ---
dhis2_wide <- read_csv("dhis2_export_wide.csv")
dhis2_long <- dhis2_wide %>%
  pivot_longer(cols = starts_with("indicator_"),
               names_to = "indicator", values_to = "value",
               names_prefix = "indicator_")

# --- NDHS: merging individual and household recode files ---
ir_data <- read_dta("NGIR_individual_recode.dta") # individual recode
hr_data <- read_dta("NGHR_household_recode.dta") # household recode

ndhs_merged <- ir_data %>%
  left_join(hr_data, by = c("v001" = "hv001", "v002" = "hv002")) # cluster + household ID
```

A Note on Data Access Permissions in Practice

Gaining legitimate access to any of the sources catalogued in this chapter typically requires navigating a specific institutional process, and a new Nigerian health data scientist should expect this to take real time: DHIS2/NHMIS access is normally granted through a State or Federal Ministry of Health data request, often requiring an institutional affiliation letter and a stated analytic purpose; NDHS/MICS microdata is available through a formal registration process with the DHS Program or UNICEF MICS website, free for legitimate research use but requiring an account and a brief description of the intended analysis; NCDC surveillance data access is generally restricted to authorised public health staff and formal research collaborations, reflecting its outbreak-sensitive nature; and institutional cohort or clinical data always requires the originating institution's ethical and data-governance approval, per Chapter 16. Building in several weeks of lead time for these approvals, rather than assuming data will be available immediately, is a basic but frequently underestimated project-planning step.

Extended Case Study

Reconciling Three Conflicting Population Denominators

A recurring, specifically Nigerian data ecosystem problem deserves its own worked illustration: for almost any LGA, at least three different population figures may be in active use simultaneously — the National Population Commission’s projection from the last full census (now more than a decade old and increasingly uncertain in fast-growing or conflict-affected areas), a State Ministry of Health facility catchment-population estimate (often based on an even older assumption and rarely reconciled against migration or displacement), and a survey-based population estimate derived from a recent NDHS or MICS cluster sample covering only a fraction of LGAs directly and requiring modelled small-area estimation for the rest. When a data scientist is asked to compute a rate — malaria cases per 1,000 population, DPT3 coverage as a percentage of the eligible birth cohort — the choice of denominator among these three candidates is not a minor technicality; it can change the resulting rate, and therefore the LGA’s apparent ranking relative to its neighbours, substantially.

Consider a specific LGA that has experienced significant internal displacement due to insecurity, with a Ministry of Health catchment estimate of 210,000 (based on pre-displacement figures), an NPC projection of 245,000 (extrapolated from the last census with an assumed uniform growth rate that likely does not reflect the actual displacement-driven population loss), and a recent rapid assessment suggesting an actual current resident population closer to 175,000. Using the Ministry catchment figure as the denominator for a reported 350 confirmed malaria cases yields a rate of 1.67 per 1,000; using the NPC projection yields 1.43 per 1,000; using the rapid assessment figure yields 2.00 per 1,000 — a range wide enough to change whether this LGA appears as a below-average, average, or above-average burden area relative to its neighbours purely as a function of which denominator was

chosen, with no change whatsoever in the actual underlying disease burden. The responsible practice, established in this chapter, is to state explicitly which denominator source was used in any reported rate, to flag LGAs (such as this one) where displacement or rapid population change makes the choice of denominator especially consequential, and, where feasible, to report a sensitivity range across the plausible denominators rather than a single unqualified figure.

A Comparative Summary Table for Quick Reference

Given how frequently this book returns to the distinct strengths and limitations of Nigeria's major health data sources, it is useful to consolidate the comparative judgments scattered across this chapter into a single quick-reference summary, distilling each source's primary strength and primary caution into a form that can be consulted quickly during project planning, without needing to re-read the chapter in full each time.

Source	Primary strength	Primary caution
DHIS2/NHMIS	Near-complete facility coverage, monthly timeliness	Denominators depend on ageing population projections; TPR-type indicators can run high
NDHS/MICS	Population-representative, independent of care-seeking	Infrequent (4-5 year gaps); requires design-based (weighted) analysis
NCDC/SORMAS-IDSR	Case-level detail, near-real-time for priority diseases	Sensitive to reporting fatigue during prolonged outbreaks
Global Fund malaria M&E	Cross-validated against source registers	Coverage limited to supported states/facilities
Institutional cohort/clinical studies	Rich individual-level covariates	Small-to-moderate sample sizes; site-specific generalisability limits

Interoperability Standards

Why Systems Struggle to Talk to Each Other

A persistent, structural reason Nigerian health data sources are often difficult to combine (beyond the denominator and definitional issues discussed elsewhere in this chapter) is a lack of shared interoperability standards across different data systems — DHIS2, a hospital’s own electronic medical record system, and a research team’s custom data collection tool may each use different facility identifiers, different disease coding systems, and different variable naming conventions for conceptually identical information. International standards such as HL7 FHIR (for exchanging clinical health data) and standardised facility master lists are gradually being adopted to address this, but a working Nigerian health data scientist should expect, in the near term, to spend real time manually reconciling facility names, disease codes, and variable definitions across systems that were never designed to communicate with each other directly — precisely the kind of fuzzy-matching and careful cross-referencing work introduced in Chapter 4’s discussion of near-duplicate detection, now understood as a symptom of a broader, system-level interoperability gap rather than a series of isolated, one-off data entry errors.

Worked Example

Estimating a Population Denominator from Two Conflicting Sources

A recurring practical problem in Nigerian health data science is that no single, universally trusted population denominator exists for computing a coverage or incidence rate: the National Population Commission’s projections (based on the last full census, now dated) and DHIS2’s own facility catchment population estimates (often derived from historical service statistics rather than an independent count) can disagree by 15% or more for a

fast-growing urban LGA. Rather than arbitrarily choosing one, a defensible approach reports the resulting rate under both denominators explicitly, flags the sensitivity of the finding to the choice of denominator, and, where the distinction matters materially for a decision (as it typically will for a coverage target close to a policy threshold), recommends the more recent, methodologically transparent source while documenting exactly why the choice was made — turning what could be an invisible, undocumented assumption into an explicit, auditable methodological decision.

Metadata Standards

DHIS2 Indicator Definitions and Why They Matter

Every indicator within a DHIS2 instance has a formal metadata definition specifying exactly which data elements and category combinations feed into it, and this definition can, in practice, be silently modified over time as a system administrator adjusts an indicator's formula to correct an error or accommodate a new reporting requirement. An analyst comparing an indicator's trend across several years without checking whether its underlying metadata definition remained stable throughout that period risks attributing an artefactual jump (caused by a definition change) to a genuine epidemiological shift. Nigerian DHIS2 instances typically maintain a metadata version history accessible through the platform's API, and a careful analyst working with any multi-year DHIS2 trend should, at minimum, confirm that the indicator's defining formula did not change materially across the period under study.

A More Robust DHIS2 Import Function

The reshaping code in Section 2.8 assumes a well-formed export with a predictable column-naming convention, an assumption that does not always survive contact with a real DHIS2 instance, where

an indicator name can contain irregular characters or an export can include unexpected extra metadata columns. A production-quality import function should isolate exactly the columns it expects to reshape, warn (rather than silently mis-parse) when an unexpected column pattern is encountered, and preserve the facility identifier as an explicit, typed column throughout, so that a downstream join (Section 2.8's NDHS merge, or any later analysis) cannot silently fail due to a type mismatch between, for example, a numeric and a string-typed facility ID.

PYTHON example

```
# Python: a more defensive DHIS2 wide-to-long import function
import pandas as pd
import warnings

def import_dhis2_export(path, indicator_prefix="indicator_"):
    df = pd.read_csv(path, dtype={"facility_id": str})
    indicator_cols = [c for c in df.columns if c.startswith(indicator_prefix)]
    if not indicator_cols:
        warnings.warn(f"No columns found with prefix '{indicator_prefix}' - check export format")
    long_df = df.melt(id_vars=["facility_id"], value_vars=indicator_cols,
                     var_name="indicator", value_name="value")
    long_df["indicator"] = long_df["indicator"].str.replace(indicator_prefix, "", regex=False)
    return long_df

dhis2_long = import_dhis2_export("dhis2_export_wide.csv")
```

Chapter Summary

- Nigeria's health data ecosystem includes routine systems (DHIS2/NHMIS), periodic surveys (NDHS, MICS), case-based surveillance (NCDC/SORMAS/IDSR), disease-specific programme data (malaria, NTDs), and primary research/cohort data.
- Each source has a distinct unit of observation, update frequency, and known bias pattern; HMIS-derived indicators such as test positivity rate can differ systematically from source-register-verified values.

- Survey data (NDHS/MICS) requires attention to sampling design (weights, clustering, stratification) before formal inference.
- Triangulating across sources is a core analytic skill, not an optional extra — but triangulation means presenting distinct, clearly-labelled estimates, not averaging quantities that measure different things.
- Practical data access requires reshaping wide DHIS2 exports into tidy format and correctly merging multi-file survey microdata on shared identifiers.

Mini-Project

Mini-Project 2: A One-Page Data Source Comparison Brief

Produce a one-page technical brief comparing two Nigerian health data sources for a health indicator of your choice, suitable for a State Ministry of Health technical working group.

- Step 1.** Select an indicator (e.g., malaria TPR, DPT3 coverage, stunting prevalence) and two data sources that both report on it.
- Step 2.** For each source, state the unit of observation, frequency, and one known limitation, using Section 2.1's table as a template.
- Step 3.** Obtain or simulate a numeric estimate from each source for the same time period and geography.
- Step 4.** Explain, in two to three sentences, any difference between the two estimates in terms of the sources' distinct data-generating processes (not just 'sampling error').
- Step 5.** Conclude with one clear, actionable recommendation for which source (or combination) the technical working group should rely on, and why.

Deliverable: *A one-page written brief with at least one comparison table.*

End-of-Chapter Exercises

1. For an indicator you work with regularly, identify which of the six data sources in Section 2.1 it comes from, and name one limitation specific to that source.
2. Explain, in your own words, why an HMIS-derived test positivity rate might be biased upward relative to a register-verified value.
3. You are asked to estimate malaria prevalence in a state. List the sources you would triangulate and what each would contribute.

4. In the worked example of Section 2.7, explain why averaging the HMIS, register-audit, and community-survey figures would produce a meaningless number.
5. A DHIS2 export arrives with one column per month (Jan_2024, Feb_2024, ...). Describe, in words, the reshaping operation needed before this data can be plotted as a time series (see Chapter 10).
6. Using the R code in Section 2.8 as a guide, write the equivalent Python (pandas) code to reshape a wide DHIS2 export into long format using `pandas.melt()`.
7. Explain why NDHS individual and household recode files must be merged on cluster and household identifiers rather than simply concatenated (stacked) into one file.
8. You are told a dataset request will take ‘6-8 weeks for approval.’ Using Section 2.9, explain how this should change your project timeline and what you could do productively while waiting.
9. Identify one Nigerian health data source not explicitly listed in Section 2.1’s table, and describe its unit of observation and update frequency.
10. Explain why an NCDC/SORMAS dataset would typically require a more restrictive access process than an NDHS dataset.
11. In the Section 2.10 case study, calculate the percentage difference between the highest (2.00 per 1,000) and lowest (1.43 per 1,000) rate estimates relative to the lowest estimate.
12. Propose one practical step a State Ministry of Health could take to reduce the kind of denominator uncertainty illustrated in Section 2.10 for a rapidly changing LGA.
13. Explain why simply always using the most recent NPC projection, without further investigation, would not fully

solve the denominator problem illustrated in this case study.

14. If you had to choose just one of the three denominators in Section 2.10 to report as a headline figure, which would you choose and why, given the LGA's displacement context?
15. Describe how you would flag, in a technical report, that a reported rate is sensitive to denominator choice, without simply omitting the rate altogether.
16. Using the summary table in Section 2.11, identify which source you would recommend for a question about population-representative HIV prevalence, and justify your choice.
17. Which source in Section 2.11's table would you recommend for near-real-time monitoring of a suspected cholera outbreak, and why?
18. Explain why 'primary strength' and 'primary caution' are both necessary parts of a useful quick-reference table, rather than strength alone.
19. Propose one additional column you would add to the table in Section 2.11 to make it more useful for project planning.
20. If you had to choose only one source from Section 2.11's table to fund and strengthen with limited resources, which would you choose for a state focused on malaria elimination, and why?
21. Using Section 2.12, explain why two systems using different facility identifiers can make even a simple data merge difficult, beyond any denominator issue.
22. What is HL7 FHIR, according to Section 2.12, and what problem is it intended to solve?
23. Why does the chapter describe interoperability gaps as a 'structural,' rather than a one-off data entry, problem?

24. Propose one practical workaround a research team could use today, while waiting for full interoperability standards adoption, to merge two systems with mismatched facility names.
25. Explain the connection between Section 2.12's interoperability problem and Section 4.7's fuzzy-matching technique.
26. Explain why this chapter treats 'which data source to use' as a methodological decision on par with choosing a statistical model.
27. Summarise, in your own words, the single biggest practical challenge a Nigerian health data scientist faces when combining two or more of this chapter's data sources.
28. Propose one improvement to Nigeria's health data ecosystem (realistic or aspirational) that would most reduce the kinds of problems illustrated in this chapter's case studies.
29. Why does this chapter emphasise triangulation over simply trusting whichever single source is easiest to access?
30. If you could add one new nationally-representative data source to Nigeria's health data ecosystem, what would it measure, and why?
31. Explain, using Section 2.14, how an indicator's metadata definition changing partway through a multi-year period could create an artefactual trend jump.
32. What specific check would you perform before trusting a five-year DHIS2 trend line for a single indicator?
33. Why might a system administrator have a legitimate reason to change an indicator's formula, even though doing so complicates trend analysis?

34. Propose a way to flag, in a published trend chart, that an indicator's definition changed partway through the period shown.
35. How does the metadata-versioning problem in Section 2.14 relate to the data dictionary concept introduced in Section 4.8?