

The Ethical Foundations of Health Data Research

Every technique in this book operates on data that ultimately originates from real people — patients, survey respondents, research participants — and the ethical obligations that attach to health research (respect for persons, beneficence, and justice, as classically formulated in the Belmont Report and operationalised in Nigeria through the National Code of Health Research Ethics and institutional Health Research Ethics Committee review) apply with undiminished force to secondary data analysis, machine learning, and big data projects, not only to primary data collection. Informed consent, confidentiality, and the principle of minimal necessary data use should shape a data science project's design from its earliest stage, not be treated as a compliance checkbox appended at the end.

Privacy, De-identification, and the Limits of Anonymisation

Removing direct identifiers (name, address, phone number) from a dataset is necessary but not sufficient for genuine de-identification, because a combination of seemingly innocuous quasi-identifiers (age, sex, LGA of residence, rare diagnosis) can

be sufficient to re-identify an individual, particularly in a small facility or a rare-disease dataset — a risk this book’s own worked examples (dermatoglyphic studies, small-area NTD surveillance) are directly exposed to. Techniques such as k-anonymity (ensuring every combination of quasi-identifiers is shared by at least k records), aggregation to a coarser geographic or age band, and, for the most sensitive projects, formal differential privacy methods, should be considered proportionate to the sensitivity of the data and the size of the population it describes.

Algorithmic Bias and Fairness in Health Prediction

A predictive model (Chapter 11) trained on historical data inherits any bias present in how that data was generated — if historical access to a diagnostic test or a specialist referral was itself unequal across geography or socioeconomic status, a model trained on outcomes conditional on that access will reproduce and can even amplify the inequity if deployed uncritically. Auditing a health prediction model’s performance across relevant subgroups (as introduced in Section 11.5) is not an optional extra step but a core ethical obligation whenever a model’s output will influence who receives clinical attention, referral, or resources, and it should be conducted and reported alongside, not instead of, overall performance metrics.

Data Sharing, Sovereignty, and Institutional Capacity

Nigerian health data increasingly flows into international research collaborations, global health databases, and commercial cloud infrastructure, raising questions of data sovereignty (who controls, and benefits from, data collected within Nigeria) that a working data scientist should be able to discuss with collaborators and institutional research offices: appropriate material/data transfer agreements, local capacity-building commitments (co-authorship, training, technology transfer) as a condition of collaboration, and, wherever feasible, retention of a locally accessible copy of any

dataset generated from Nigerian participants or facilities. Building durable in-country capacity for the full data science pipeline described in this book — not only data collection, but analysis, modelling, and interpretation — is itself a matter of research equity, not merely of technical training.

Worked Example

A K-Anonymity Check on a Small Dataset

Consider a de-identified dataset (names removed) of 40 leishmaniasis patients from a single LGA, with quasi-identifiers age band (0-9, 10-19, 20+), sex, and ward of residence. Cross-tabulating these three quasi-identifiers reveals that one combination — female, age 0-9, Ward C — has exactly one matching record. This single record fails even 2-anonymity (the weakest reasonable standard, requiring at least two records per combination) and is at meaningful risk of re-identification by anyone in the community who knows a young girl in Ward C was recently treated for leishmaniasis. The appropriate remedy is not simply to delete that one row (which would create a different problem — silently biasing the dataset) but to coarsen the offending quasi-identifiers, for example by widening the age band or aggregating ward of residence to a larger administrative unit, until every combination is shared by at least $k = 5$ records, a commonly used minimum threshold in small-area public health data releases.

R example

```
# R: checking k-anonymity across quasi-identifier combinations
library(dplyr)

k_anon_check <- patient_data %>%
  count(age_band, sex, ward, name = "n_matching_records") %>%
  arrange(n_matching_records)

cat("Smallest group size (k):", min(k_anon_check$n_matching_records), "\n")
print(k_anon_check %>% filter(n_matching_records < 5)) # combinations failing k=5
```

Worked Example: A Subgroup Fairness Audit

Returning to the triage risk model introduced in Chapter 11: an overall AUC of 0.85 sounds strong, but a subgroup audit reveals AUC of 0.88 for patients presenting at tertiary hospitals (where the training data was concentrated) versus only 0.74 for patients presenting at primary health centres — and, more concerning still, a calibration check shows the model under-predicts risk by an average of 12 percentage points specifically for primary-health-centre patients. Deploying this model unmodified at primary health centres, where it was least well validated, would mean the facilities with typically the fewest resources for clinical judgement to compensate for a flawed score are also the ones receiving the least reliable predictions — precisely the equity concern raised in Section 16.3, made concrete with numbers rather than left as an abstract worry.

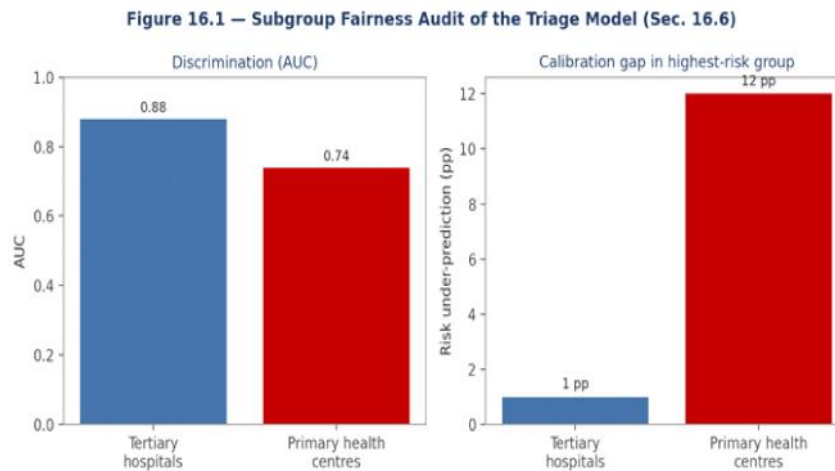


Figure 16.1 — Subgroup fairness audit of the triage model (Section 16.6).

PYTHON example

```
# Python: a subgroup AUC and calibration audit
from sklearn.metrics import roc_auc_score

for facility_type in ["tertiary", "primary"]:
    subset = triage_df[triage_df["facility_type"] == facility_type]
    auc = roc_auc_score(subset["severe_outcome"], subset["predicted_risk"])
    mean_predicted = subset["predicted_risk"].mean()
    mean_observed = subset["severe_outcome"].mean()
    print(f"{facility_type}: AUC={auc:.2f}, predicted={mean_predicted:.1%},
observed={mean_observed:.1%}")
```

Looking Forward

The chapters of this book have moved from foundational probability and statistical inference, through classical and modern regression, survival analysis, time series forecasting, machine learning, unsupervised learning, mechanistic infectious disease modelling, spatial epidemiology, and big data/AI methods, consistently anchored in Nigerian health data and health system realities. The frontier of this field continues to move — toward richer integration of routine health system data with research cohorts, toward federated learning approaches that allow model training across institutions without centralising sensitive data, and toward AI-assisted diagnostic and decision-support tools appropriately validated for low-resource settings — but the foundational discipline this book has tried to instil (rigorous attention to where data comes from, what it can and cannot support, and who is accountable for how it is used) will remain the essential and durable skill beneath whatever specific technique or tool comes next.

Community Engagement in Data Governance

Beyond formal ethical approval, meaningful community engagement — informing the communities from which health data is collected about how it will be used, and, where feasible,

involving community representatives in decisions about data sharing and secondary use — is increasingly recognised as good practice in health data governance, particularly for sensitive research involving stigmatised conditions or key populations (Chapter 7’s respondent-driven sampling contexts). This is not merely an ethical nicety: community trust, once damaged by a perceived misuse of health data, can take years to rebuild and can compromise the willingness of an entire community to participate in future essential public health research and surveillance activities.

Extended Case Study

Well-Intentioned Data Sharing Request, Reconsidered

A genuinely difficult, realistic ethics scenario worth working through in detail: a well-regarded international research consortium requests access to a Nigerian institution’s de-identified dermatoglyphic (fingerprint pattern) research dataset, comparing children with a specific congenital condition against unaffected controls, for inclusion in a larger, multi-country meta-analysis that could meaningfully advance understanding of the condition’s early biomarkers. The request is well-intentioned, scientifically legitimate, and could benefit the broader research community, including future Nigerian patients who might benefit from earlier or more accurate diagnosis as a result. It is also, on the specific details of this dataset, higher-risk than most typical secondary data sharing requests: the condition in question is rare enough, and the study’s recruitment sites specific enough, that even a carefully de-identified dataset carries real re-identification risk of the kind discussed in Section 16.2, and the original informed consent obtained from participants’ families, some years earlier, did not explicitly anticipate international data sharing as a possible future use.

Working through this scenario using the framework of this chapter, several considerations must be weighed together rather

than in isolation: whether the original consent language is broad enough to reasonably cover this new use, or whether a fresh ethical review and, ideally, a renewed consent process would be required; whether additional data transformation (coarsening age bands, aggregating recruitment site information beyond what the international meta-analysis strictly needs) could reduce re-identification risk to an acceptable level without compromising the scientific value of the shared data; and whether the collaboration agreement includes the kind of local capacity-building and co-authorship provisions discussed in Section 16.4, ensuring the Nigerian research team and institution receive appropriate recognition and benefit rather than simply serving as a data source for an externally-led publication. The eventual resolution in a case of this kind — often involving referral back to the original institutional Health Research Ethics Committee for a specific ruling on whether the original consent covers this new use, and negotiation of an explicit data use and co-authorship agreement before any data transfer occurs — illustrates that responsible data governance in practice is rarely a simple yes/no gate, but a structured process of working through genuine, substantive tensions between scientific value, participant protection, and research equity.

Distinguishing Data Ethics Review from Data Protection Compliance

It is worth explicitly distinguishing two related but distinct obligations that a Nigerian health data science project must satisfy: research ethics review (obtained from an institutional Health Research Ethics Committee, assessing whether the research is scientifically justified and ethically conducted with respect to participants) and data protection compliance (governed in Nigeria by the Nigeria Data Protection Act 2023 and overseen by the Nigeria Data Protection Commission, concerned specifically with the lawful processing, storage, and transfer of personal data

regardless of research purpose). A project can satisfy one without automatically satisfying the other — a well-designed study with full ethical approval could still process personal data in a way that falls short of data protection requirements (for example, transferring identifiable data internationally without an adequate legal basis) — and a mature institutional data governance process should treat these as two separate, both-necessary checks, rather than assuming ethical approval alone covers all relevant legal and regulatory obligations.

Whistleblowing and Reporting Data Misuse

This chapter has focused primarily on how to design and conduct a data science project ethically from the outset; it is equally important to briefly address what a data scientist should do upon discovering that data has already been misused — for example, finding that a colleague has shared identifiable patient data outside an agreed data use agreement, or that a dataset's findings have been selectively reported to serve a predetermined political or programmatic narrative rather than the actual evidence. Most Nigerian research institutions and international collaborations maintain a formal research integrity or ethics reporting channel specifically for this purpose, and using it — rather than remaining silent out of loyalty to a colleague or discomfort with confrontation, or conversely making an unverified public accusation — is both an individual ethical obligation and, in the long run, a protection for the broader credibility of the data science profession and the communities whose data and trust are ultimately at stake; a data scientist who is aware of this reporting pathway before they ever need it is considerably better positioned to act appropriately and promptly if the need does arise.

Data Retention and Secure Disposal Policies

Health data governance (Section 4.5, and this chapter's broader treatment) must address not only how data is collected, stored,

and shared, but how and when it is eventually disposed of. A written data retention policy — specifying how long identifiable data will be kept before being either securely destroyed or irreversibly de-identified, consistent with the original consent obtained from participants and any applicable institutional or regulatory requirements — protects both the research participants (whose data does not persist indefinitely beyond its original justified purpose) and the institution (which is not left holding an ever-growing, ever-riskier archive of identifiable sensitive data with no defined end point), and should be established at a project's outset alongside its ethical approval, not improvised years later when a dataset's continued retention is questioned.

Auditing an Inherited Dataset for Undocumented Ethical Debt

A researcher who inherits a dataset originally collected by a previous team, sometimes years earlier, faces a specific challenge this chapter calls undocumented ethical debt: the original ethical approval, consent language, and any data use restrictions may be poorly documented, ambiguous about permitted secondary uses, or simply unavailable. Before conducting any new analysis on an inherited dataset, a responsible researcher should attempt to locate and review the original ethical approval and consent documentation, confirm whether the planned new analysis falls within its originally approved scope, and, where this cannot be confirmed with reasonable confidence, seek fresh ethical guidance rather than assuming that physical possession of a dataset implies unrestricted permission to analyse and publish from it.

A General-Purpose k-Anonymity Checker Function

The k-anonymity check in Section 16.5 is written for one specific set of three quasi-identifiers; a general-purpose version that accepts any list of quasi-identifying columns can be reused across every future dataset a research team handles, rather than being

rewritten from scratch each time, directly supporting the personal code library habit recommended in Section 17.12.

PYTHON example

```
# Python: a reusable k-anonymity checker for any set of quasi-identifiers
def check_k_anonymity(df, quasi_identifiers, k=5):
    group_sizes = df.groupby(quasi_identifiers).size()
    failing = group_sizes[group_sizes < k]
    return {"min_group_size": int(group_sizes.min()), "n_failing_groups": len(failing),
           "passes_k_anonymity": len(failing) == 0, "failing_groups": failing}

result = check_k_anonymity(patient_df, ["age_band", "sex", "ward"], k=5)
if not result["passes_k_anonymity"]:
    print(f"{result['n_failing_groups']} group(s) fail k=5 anonymity - coarsen a quasi-
    identifier")
```

Chapter Summary

- Research ethics principles (respect for persons, beneficence, justice) apply fully to secondary analysis and machine learning, not only primary data collection.
- De-identification requires attention to quasi-identifiers and re-identification risk (k-anonymity), not just removal of direct identifiers.
- Predictive models can inherit and amplify historical inequities in access to care; subgroup performance auditing is an ethical obligation, not an optional extra.
- Data sovereignty, fair collaboration agreements, and local capacity-building are matters of research equity in international health data collaborations.

The durable skill of this book is disciplined attention to data provenance and accountability, which persists across whatever specific technique comes next

Mini-Project**Mini-Project 16: A Privacy and Fairness Audit**

Conduct both a k -anonymity check and a subgroup fairness audit on a dataset and/or model of your choice.

Step 1. Select a de-identified dataset with at least three plausible quasi-identifiers (e.g., age band, sex, geographic unit).

Step 2. Compute the smallest group size across all quasi-identifier combinations (Section 16.5) and identify any combinations that fail a chosen k threshold.

Step 3. Propose and apply a specific coarsening strategy (e.g., wider age bands, larger geographic units) to bring all combinations up to your chosen k , and report the trade-off in information lost.

Step 4. If you have a predictive model (from a previous chapter's mini-project or a new one), compute its performance (AUC and calibration) separately for at least two meaningful subgroups.

Step 5. Write a one-page ethics and equity memo summarising both audits and your recommendations for safe release or deployment.

Deliverable: *A k -anonymity check with a remediation plan, a subgroup performance audit table, and a one-page ethics and equity memo.*

End-of-Chapter Exercises

1. A dataset removes patient names but retains exact age, LGA, and a rare diagnosis. Explain the re-identification risk this creates and propose a mitigation.
2. Describe a plausible scenario in which a clinical risk-prediction model trained on Nigerian tertiary hospital data could disadvantage patients from primary health centres.
3. What provisions would you want to see in a data-sharing agreement before contributing a Nigerian patient cohort dataset to an international research collaboration?
4. Using the worked example in Section 16.5, explain why simply deleting the single at-risk record (rather than coarsening the quasi-identifiers) would itself introduce a new problem into the dataset.
5. In Section 16.6, explain why the combination of lower AUC AND under-prediction of risk at primary health centres is a more serious equity concern than either finding alone.

6. Using the R code in Section 16.5 as a guide, describe what would happen to the minimum group size (k) if 'ward' were replaced with the coarser variable 'LGA' in the grouping.
7. A funder asks you to justify, in two or three sentences, why 'the data has already had names removed' is not a sufficient answer to a privacy review question. Write that justification.
8. Using Section 16.8, explain why community trust, once damaged, could compromise future surveillance efforts beyond the immediate research project in question.
9. Propose one concrete way a research team could meaningfully engage a community before collecting sensitive health data from its members.
10. Why might community engagement be especially important for research involving a stigmatised condition or a key population, compared to a general household health survey?
11. In the Section 16.9 case study, explain why the fact that the original consent did not anticipate international data sharing matters, even though the new use is scientifically legitimate.
12. Propose one specific data transformation that could reduce re-identification risk for this dataset without eliminating its scientific value for the meta-analysis.
13. Why should the resolution of this kind of request generally involve the original institutional Health Research Ethics Committee, rather than being decided solely by the researchers holding the data?
14. What would you want included in a data use and co-authorship agreement before agreeing to share this dataset internationally?
15. Explain why 'the request is well-intentioned and scientifically valuable' is necessary but not sufficient

justification for approving a data sharing request like this one.

16. Using Section 16.10, explain why a study could have full research ethics approval but still fail to meet data protection compliance requirements.
17. Under the Nigeria Data Protection Act 2023 framework referenced in Section 16.10, why would international data transfer be treated as a distinct issue requiring its own legal basis?
18. Propose one practical step an institution could take to ensure both ethical approval and data protection compliance are checked, rather than assuming one covers the other.
19. Explain the difference in focus between a Health Research Ethics Committee's review and a Data Protection Commission's concerns, in your own words.
20. Why is it important for a data governance process to treat these as 'two separate, both-necessary checks' rather than a single combined approval step?
21. Using Section 16.11, explain why remaining silent about discovered data misuse is described as an ethical problem, not just a matter of personal discomfort.
22. Why is making an unverified public accusation also described as an inappropriate response, alongside remaining silent?
23. What is the recommended channel, according to Section 16.11, for reporting a genuine concern about data misuse?
24. Why is it useful to know about this reporting pathway 'before you ever need it,' rather than only learning about it after discovering a problem?
25. Propose a realistic (anonymised or hypothetical) scenario where you would feel obligated to use such a reporting channel, and explain why.

26. Summarise, in your own words, why this chapter treats data ethics review and formal data protection compliance as related but distinct processes (Section 16.10).
27. Explain the common thread linking the whistleblowing discussion (Section 16.11) and the reconsidered data-sharing request case study (Section 16.9).
28. Why does this chapter argue that a ‘well-intentioned’ data request can still raise legitimate ethical concerns?
29. Propose a scenario from your own institution where distinguishing ethics review from data protection compliance (Section 16.10) would matter in practice.
30. If you personally observed a colleague misusing sensitive health data, what would this chapter suggest as an appropriate first step?
31. Using Section 16.13, explain what is meant by ‘undocumented ethical debt’ in an inherited dataset.
32. What specific steps should a researcher take before analysing a dataset originally collected by a different team years earlier?
33. Why is physical possession of a dataset not, by itself, sufficient justification for analysing and publishing from it?
34. Describe a plausible scenario where a planned new analysis of an inherited dataset would fall outside its original ethical approval’s scope.
35. How does the concept of undocumented ethical debt relate to the data dictionary and metadata documentation practices introduced in Chapter 4?