Mamadou S. Diallo, Ph.D.

Statistician · Population Health Methodologist · RWE Platform Architect

Hoboken, NJ, USA

msdiallo@samplics.org | +1 917 251 0020 | msdiallo.com | linkedin.com/in/MamadouSDiallo

PROFESSIONAL SUMMARY

Ph.D. statistician with 20 years generating population-level health evidence from complex real-world data — observational study design, causal inference, multi-source data integration, and small area estimation — across health surveillance, immunization, HIV epidemiology, and chronic disease.

Lead survey statistician on national HIV biomarker surveys (PHIA, 8 countries with CDC/PEPFAR), the **California Health Insurance** and **Health Coverage** surveys for Covered California at NORC, the **California Health Care Foundation** survey programs, the WHO/UNICEF immunization coverage estimates (WUENIC) for 195 countries, and several WHO/UNICEF National Vaccination Coverage Surveys in low- and middle-income countries. Senior contributor on the National Health and Nutrition Examination Surveys (NHANES) and designed the NHANES 2012 compositing strategy at Westat. Author of **Samplics** (265K+ downloads, JOSS) and architect of the open-source **svy** ecosystem.

Sole architect and developer of **svyLab** — a multi-tenant analytics platform with architectural sandbox isolation, lifecycle and classification governance, AI-assisted analysis with full provenance, and a 327-test invariant suite. Built deliberately to model the reproducibility, auditability, and tenant-safety properties that regulated RWE work requires.

Currently extending the **svy** ecosystem into pharmacoepidemiology (ISPE 2026 poster, Milan) and seeking to apply 20 years of population-based observational research and platform engineering to industry RWE. Fluent in English and French.

CORE COMPETENCIES

Causal Inference & Observational Methods

- Survey-weighted causal inference: IPTW, stabilized weights, doubly robust estimation, propensity-score methods (in active development for ISPE 2026)
- Population-based observational study design: complex multi-stage probability samples, biomarker cohorts, longitudinal surveillance studies (PHIA, NHANES, CCHS, SLCDC)
- Small area estimation: Fay–Herriot, unit-level skew-normal, spatial Bayesian models for subnational disease burden and prevalence
- Endpoint definition and harmonization: led indicator design for the global MICS immunization module (100+ countries) and for WUENIC across 195 countries

Real-World Data Integration & Evidence Generation

- Multi-source health-data integration at scale: harmonized administrative reporting (DHIS2), household-survey microdata (DHS, MICS, PHIA, NHANES), and programme surveillance across 195 countries under GATHER reporting standards
- Fit-for-purpose data assessment: source selection, quality triage, discrepancy resolution between administrative and survey signals
- Population estimation under uncertainty: ML-based census-independent denominators (Scientific Reports 2022); immunization target-population modeling
- Subnational evidence generation: built *HILDX* producing county-level health estimates across all 102 Illinois counties (NORC)

Platform & Reproducibility Engineering

- Architected and built *svyLab* solo: multi-tenant SaaS with architectural sandbox isolation (cross-org reads return 404 by design), fail-closed unified auth middleware, dataset lifecycle and classification authority model with audit logs
- AI-with-provenance: every *svy-agents* output persists as a structured artifact with model, version, prompt, context hash, tokens, cost, and timestamp — no silent AI-authored data modifications
- Reproducibility primitives: immutable dataset versions, versioned survey designs, analysis lineage with rerun-against-current-version, soft delete with selective restore, Typst-based audit reporting
- Test-invariant discipline: 327 passing tests including structural CI checks (every API route has auth + access guard), cross-org sandbox enforcement, existence-leak prevention, quota correctness

Leadership & Stakeholder Engagement

- Mentored statisticians, data scientists, and analysts across 20+ countries — fostering technical depth in study design, programming, and methodological judgment
- Presented methodology and findings to CDC, WHO, PEPFAR, USAID, Gavi, BMGF, and national governments — translating rigorous methods for technical and non-technical audiences
- Co-authored international standards (WHO Vaccination Coverage Cluster Surveys Reference Manual, 2018). Guest Editor, JSSAM 2025 Special Issue
- Bilingual operations: English and French (fluent in both speaking and writing)

Programming Python (Samplics, svy, Polars, JAX, Litestar) · R (survey, sae, tidyverse, Shiny) · SQL · SAS · Stata · Rust/PyO3

Methods Causal inference (IPTW, doubly robust) · Bayesian hierarchical · Survey design · SAE (Fay–Herriot, unit-level, spatial) · GLMs · Bootstrap · ML

Real-World Data Sources NHANES · PHIA · DHS/MICS · CCHS · DHIS2 · ACS · NCVS · BRFSS (familiar)

Standards & Frameworks GATHER · AAPOR · STROBE (familiar) · WHO survey methodology · Reproducible-analytics governance

Currently studying ICH E9(R1) · FDA RWE framework guidance · ISPE Good Pharmacoepidemiology Practices

PROFESSIONAL EXPERIENCE

Samplics LLC

Hoboken, NJ, USA

Founder & Principal Statistician

May 2022 – Present

svyLab — Multi-tenant analytics platform (sole architect):

- Designed and built a multi-tenant SaaS platform (Python/Litestar + PostgreSQL + DuckDB + Redis) with **architectural sandbox isolation** — cross-organization reads return 404 at every endpoint by design, not by access control alone. Fail-closed unified session-or-bearer auth middleware; CI-enforced structural invariant that every API route carries `require_login` + per-resource access guard.
- **Lifecycle & classification governance:** implemented an authority model where dataset maturity (`in_production` → `team_finalized` → `org_validated` → `archived`) and access classification (public / restricted) are orthogonal axes governed by separate authorities, with full audit logs for every transition. Designed deliberately to prevent the silent-authorization-expansion bugs that leak sensitive data on promotion.
- **AI with provenance (svy-agents):** every natural-language analysis request produces a translated, editable, auditable `svy` code call. Each output persists with model, version, prompt, context hash, tokens, cost, and

timestamp. The principle — "AI makes methodology accessible without hiding it" — is enforced architecturally: no silent AI-authored data modifications; every transformation traceable to a user decision.
- **Reproducibility primitives:** immutable dataset versions, versioned survey designs with full replicate-weight support, analysis lineage with rerun-against-current-version, soft delete with 30-day selective restore, Typst-based audit reporting.
- 327 passing tests including structural CI checks, cross-org sandbox enforcement, existence-leak contracts, analysis hardening, cache invariants, quota correctness, and lifecycle/classification authority.

svy ecosystem — Open-source survey statistics for Python:

- **Samplics** (265K+ downloads, JOSS 2021): production-grade Python library for sample selection, weighting, estimation, and small area estimation.
- **svy** (Beta): next-generation survey design and inference (Taylor linearization; replicate variance — BRR, jackknife, bootstrap), built with Rust/PyO3 and Polars for performance on large datasets.
- **svy-sae** (Beta): JAX-powered Fay–Herriot and unit-level small area estimation for subnational disease burden and prevalence.
- **svy-io** (Alpha): high-speed I/O for SAS, SPSS, and Stata files — preserving variable labels, value labels, and user-defined missing codes.
- **svy-causal** (in development for ISPE 2026 poster, Milan): survey-weighted IPTW and doubly robust estimation, validated against NHANES.

Consulting & capacity building:

- **National Bureau of Statistics, Tanzania** (SADC/World Bank, 2026): conducted a national statistical-capacity diagnostic, gap analysis, and prioritized reform plan; delivered a 5-day national training on multi-stage design, weighting, calibration, and variance estimation.
- **INSBU Burundi / EAC Secretariat** (2025): 2-week national training on small area estimation for granular poverty estimates.
- **FeeLoST**: designed and deployed a DHIS2-integrated platform for Ethiopia's Ministry of Health — automated routine administrative health-data ingestion, multi-level data-quality validation, and structured feedback workflows between facilities and administrative units.
- **McGovern Dole Program — The Gambia (2023, 2026) and Togo (2024)** (CRS/USDA): led end-to-end design of national probability surveys for program evaluation in education and nutrition — sample design, CAPI instrument development, weighting, analysis, reporting.

NORC at the University of Chicago

Remote, USA

Principal Data Scientist

Feb 2024 – Feb 2026

- **Healthy Illinois Analytics Platform (HILDX):** Led development of a Python-based analytics platform producing **sub-county health estimates** across all 102 Illinois counties — integrating administrative data (American Community Survey) with survey microdata using small area estimation. Implemented disclosure control, reproducible workflows, and secure data governance.
- **Covered California — California Health Insurance and Health Coverage Surveys:** Lead statistician for the monthly and annual surveys conducted by NORC for Covered California, the state's health insurance marketplace. Sampling, weighting, and estimation for cohorts of Medi-Cal renewal recipients, ACA marketplace enrollees, and health-coverage transitions — directly comparable to RWD cohorts in healthcare access, coverage, and utilization research.
- **California Health Care Foundation survey programs:** Lead statistician for multiple CHCF-funded studies on Californians' health care access, costs, medical debt, and program experiences with Medi-Cal and Covered California.

UNICEF

New York, NY, USA

Global Lead for Immunization Data

May 2016 – Aug 2021

- **WUENIC — multi-source health-evidence integration at global scale:** led annual production of WHO/UNICEF immunization coverage estimates for 14 vaccines across all 195 countries. Integrated three categories of real-world data — administrative reporting (DHIS2), household-survey microdata (DHS, MICS, EPI cluster surveys), and programme surveillance (stockouts, disease incidence) — under GATHER reporting standards. Conducted fit-for-purpose assessments across data modalities, resolved discrepancies, and produced harmonized time-series estimates.
- **MICS immunization Subject Matter Expert:** shaped endpoint measurement for UNICEF's global household-survey programme (100+ countries) — module design, indicator definitions, cross-country harmonization, data-quality review.
- **Integrated Health Database (IHD):** initiated a multi-country platform combining DHS, MICS, and PHIA microdata with standardized indicators, automated quality checks, and an internal R package — defined and implemented standardized processes for data management and documentation.
- **Zero-dose strategy (Gavi/BMGF):** led geospatial analytics to identify under-vaccinated communities and guide targeted interventions. Presented study designs and results to Gavi board, BMGF, and WHO leadership — translating findings for non-technical audiences.
- **Population estimation under uncertainty:** developed ML models for vaccination target populations with uncertainty quantification — co-published in Scientific Reports (2022).
- Mentored data scientists and analysts across 20+ LMICs. Country missions to Chad, Côte d'Ivoire, Kenya, Mozambique, Senegal, Pakistan, Uganda, Zimbabwe.

Saudi Center for Opinion Polling (SCOP)

Riyadh, Saudi Arabia

Lead Statistician & COO

Aug 2021 – Jan 2024

- Designed and implemented sampling, weighting, and estimation for national-level multi-purpose surveys; built PowerBI dashboards for survey-performance monitoring and results visualization.
- Managed CATI call-center operations: questionnaire design, pre-testing, cognitive testing, and quality metrics. Developed AAPOR-compliant nonresponse standards.
- Delivered training-of-trainers sessions on survey design, statistical experiments, and analysis.

Westat

Rockville, MD, USA

Senior Statistician

Jun 2011 – Apr 2016

Designed and analyzed population-based observational studies for U.S. federal agencies and international health organizations.

Selected projects:

- **Population-based HIV Impact Assessment (PHIA):** led sampling design for **8 national biomarker surveys** (Cameroon, Côte d'Ivoire, Malawi, Namibia, Tanzania, Uganda, Zambia, Zimbabwe) in collaboration with ICAP at Columbia University and CDC/PEPFAR. Complex multi-stage probability samples for estimating HIV prevalence, incidence, viral load suppression, and ART coverage at national and subnational levels. Presented methodology at stakeholder meetings with CDC; led country missions for implementation and training. Advised on subnational SAE methods.
- **NHANES 2012 & National Youth Fitness Survey:** designed the compositing strategy for NHANES 2012 — the primary U.S. population health survey used extensively in pharmacoepidemiology. Improved weight-trimming methods for reducing the influence of extreme survey weights on estimates and variance.
- **National Crime Victimization Survey (NCVS):** developed model-based small area estimation for state-level estimates using 15 years of longitudinal survey data. Published the R package **sae2** on CRAN.
- **Mid-Term Haiti Survey (USAID):** designed a complex multi-stage household survey using satellite imagery and GPS localization. Led country missions, trained partners, supervised fieldwork.
- **Breastmilk Substitutes Compliance Assessment (Vietnam/Indonesia):** designed a three-stage sampling strategy for assessing compliance with international marketing regulations.

Statistics Canada

Ottawa, ON, Canada

Mathematical Statistician

May 2006 – May 2011

- Developed sampling and estimation methodology for Canada's **national health surveillance system** — the **Canadian Community Health Survey (CCHS)**, the Canadian counterpart of NHANES. Also worked on Census undercoverage and the Survey of Labour and Income Dynamics (SLID).
- Led full methodological development for the **Survey on Living with Chronic Disease in Canada (SLCDC)** — a population-based chronic-disease study with two-phase selection design and bootstrap variance estimation.

Robert Giffard / Laval University Research Center

Quebec, Canada

Biostatistician

May 2005 – Apr 2006

- Modeled cardiovascular-disease risk factors using the **Framingham Heart Study** — gene-environment interaction analysis with generalized linear models and sampling design optimization. Co-published in **Epidemiology** (2008).

EDUCATION

Ph.D., Statistics (2015)
Carleton University, Canada
Supervisor: J.N.K. Rao

M.Sc., Statistics (2006)
Université Laval, Canada

Licence, Mathematics (2002)
Université Claude Bernard Lyon 1, France

SELECTED PUBLICATIONS & PRESENTATIONS

Submitted / Accepted

Diallo M.S. (2026). *Design-based causal inference for pharmacoepidemiology: survey-weighted IPTW and doubly robust estimation in Python with NHANES validation.* Poster accepted, ISPE 42nd Annual Meeting, Milan, Italy.

Epidemiological & Statistical Methodology

Bureau A., Diallo M.S., Ordovas J.M. and Cupples L.A. (2008). *Efficiency of sampling designs within a cohort for estimating interaction effects between genetic and environmental risk factors.* Epidemiology, 19(1), 83–93. doi:10.1097/ede.0b013e31815c4d0e

Diallo M.S. and Rao J.N.K. (2018). *Small area estimation of complex parameters under unit-level models with skew-normal errors.* Scandinavian Journal of Statistics, 45(4):1092–1116. doi:10.1111/sjos.12336

Fay R.E. and Diallo M.S. (2015). *Developmental Estimates of Subnational Crime Rates Based on the NCVS.* Bureau of Justice Statistics, Office of Justice Programs. (Foundational work for the **sae2** CRAN package.) Link

Diallo M.S. (2021). *Samplics: A Python package for selecting, weighting and analyzing data from complex sampling designs.* Journal of Open-Source Software, 6(68), 3376. doi:10.21105/joss.03376

Population Health & Real-World Evidence

Neal I., Seth S., Watmough G. and Diallo M.S. (2022). *Census-independent population estimation using representation learning.* Scientific Reports, 12, 5185. doi:10.1038/s41598-022-08935-1

Danovaro-Holliday M.C., Gacic-Dobo M., Diallo M.S., Murphy P. and Brown D.W. (2021). *Compliance of WUENIC with GATHER criteria.* Gates Open Research, 5:77. doi:10.12688/gatesopenres.13258.1

Muhoza P., Danovaro-Holliday M.C., Diallo M.S. et al. (2021). *Routine Vaccination Coverage — Worldwide 2020.* MMWR, 70:1495–1500. (One of six annual MMWR co-authored publications, 2016–2021.)

Bruni L., Saura-Lázaro A., Montoliu A., Brotons M., Alemany L., Diallo M.S. et al. (2020). *HPV vaccination introduction worldwide and WHO/UNICEF estimates of national HPV immunization coverage 2010–2019.* Preventive Medicine, 106399. doi:10.1016/j.ypmed.2020.106399

Chaney S.C., Mechael P., Thu N.M., Diallo M.S. and Gachen C. (2021). *Every Child on the Map: Immunization Equity Using Geospatial Data.* JMIR, 23(8):e29759. doi:10.2196/29759

WHO (2018). *Vaccination Coverage Cluster Surveys: Reference Manual.* World Health Organization (WHO/IVB/18.09). (Contributing author.) Link

OPEN-SOURCE SOFTWARE

- **Samplics** (265K+ downloads, JOSS 2021): Python library for survey sampling, weighting, estimation, and small area estimation. PyPI
- **svy ecosystem** (svy / svy-sae / svy-io): next-generation Python survey statistics built with Rust/PyO3, JAX, and Polars. svylab.com/docs
- **svy-causal** (in development): survey-weighted IPTW and doubly robust estimation for pharmacoepidemiology — ISPE 2026 poster.
- **sae2** (CRAN): R package for model-based small area estimation, developed for the Bureau of Justice Statistics.
- **svyLab**: cloud-native multi-tenant analytics platform with sandbox isolation, lifecycle governance, AI-with-provenance, and 327-test invariant suite. svylab.com

PROFESSIONAL ACTIVITIES

- **Guest Editor**, Journal of Survey Statistics and Methodology (JSSAM), 2025 Special Issue: *"Survey Research on Asia, Africa, Latin America, The Caribbean, and Oceania."*
- **Reviewer** — Journal of Official Statistics; Survey Methodology
- **Member** — ISPE (2026), American Statistical Association (since 2010), AAPOR
- **Languages** — English (fluent) · French (fluent)