

## **Glossary of Key Terms**

---

### **Bayes' theorem**

A formula relating conditional probabilities, used here to link diagnostic test sensitivity/specificity to positive predictive value.

### **Censoring**

In survival analysis, the condition where the exact event time is unknown because the event had not occurred (or was not observed) by the end of follow-up.

### **Choropleth map**

A map that shades administrative units by the value of an indicator, such as disease case rate.

### **Confounding**

A distortion of an exposure-outcome association caused by a third variable associated with both.

### **DHIS2**

The District Health Information Software platform underlying Nigeria's National Health Management Information System (NHMIS).

### **Design effect**

The factor by which a complex survey design (clustering, stratification) inflates variance relative to simple random sampling.

**Hazard ratio**

In survival analysis, the multiplicative effect of a covariate on the instantaneous risk of an event, estimated by a Cox model.

**MAR / MCAR / MNAR**

Missing at random / missing completely at random / missing not at random — mechanisms describing why data are missing.

**Odds ratio**

The exponentiated coefficient from a logistic regression, interpreted as a multiplicative change in the odds of an outcome.

**Overdispersion**

A situation where a count outcome's variance exceeds its mean, violating the Poisson assumption and motivating negative binomial regression.

**R0 (basic reproduction number)**

The expected number of secondary infections caused by one infectious individual in a fully susceptible population.

**Respondent-driven sampling (RDS)**

A peer-referral chain sampling method for populations without a sampling frame, requiring specialised network-based estimators.

**SEIR model**

A compartmental infectious disease model dividing a population into Susceptible, Exposed, Infectious, and Recovered states.

**Test positivity rate (TPR)**

The proportion of diagnostic tests performed that return a positive result.

**Attack rate**

The proportion of an at-risk population that becomes a case during a defined outbreak period.

**Basic reproduction number ( $R_0$ )**

The expected number of secondary infections produced by one infectious individual in an otherwise fully susceptible population.

**Case fatality ratio (CFR)**

The proportion of confirmed cases that die from a given disease, ideally calculated using only cases with a resolved outcome.

**Cohen's  $d$**

A standardised effect size expressing the difference between two group means in units of pooled standard deviation.

**Cross-validation**

A model validation technique that repeatedly trains and tests a model on different partitions (folds) of the data to estimate out-of-sample performance.

**Data dictionary**

A structured document defining every variable in a dataset, including its meaning, type, allowed range, and source.

**DBSCAN**

A density-based clustering algorithm that groups closely-packed points and labels sparse, isolated points as noise rather than forcing them into a cluster.

**Effective reproduction number ( $R_t$ )**

The time-varying average number of secondary cases produced per infectious case at a given point during an ongoing outbreak.

**Empirical Bayes smoothing**

A statistical technique that shrinks an unstable small-area estimate toward a more stable wider-area average, in proportion to its uncertainty.

**Epidemic curve (epicurve)**

A histogram of case counts by date of symptom onset, used to infer the likely mode of transmission in an outbreak.

**k-anonymity**

A privacy standard requiring that every combination of quasi-identifying variables in a dataset is shared by at least  $k$  records.

**LASSO regression**

A regularised regression method that penalises the size of coefficients and can shrink some coefficients exactly to zero, performing automatic variable selection.

**Moran's I**

A statistic testing whether the spatial distribution of a variable is more clustered than would be expected under complete spatial randomness.

**Multiple imputation**

A missing-data method that generates several plausible completed datasets, analyses each, and pools the results to reflect imputation uncertainty.

**Poisson regression**

A generalised linear model for count outcomes, assuming the outcome's variance equals its mean.

**Positive predictive value (PPV)**

The probability that a person who tests positive on a diagnostic test truly has the disease, which depends heavily on disease prevalence.

**Sensitivity and specificity**

Fixed properties of a diagnostic test: sensitivity is the probability of a positive result given true disease; specificity is the probability of a negative result given true absence of disease.

**Survival analysis**

The family of statistical methods for time-to-event data that correctly incorporates censored (incomplete) observations.

**Transfer learning**

A deep learning technique that re-uses a neural network already trained on a large general dataset, retraining only its final layers on a smaller, task-specific dataset.

**Change-point detection**

A statistical technique for identifying the point(s) in a time series at which its underlying mean, variance, or trend shifts abruptly.

**Federated learning**

A machine learning approach that trains a shared model across multiple institutions' datasets without centralising the underlying raw data.

**Nowcasting**

A correction applied to the most recent period of surveillance data to account for reporting delay, producing a more accurate estimate of the current true case count.

**Quantile regression**

A regression method that models a specified quantile (rather than the mean) of an outcome's conditional distribution.

**Landmark analysis**

A survival analysis technique that restricts analysis to a fixed time point, avoiding immortal time bias from covariates measured partway through follow-up.