

Level 2: Doing the Analysis

Statistical modelling for clinical data in R

Organisation: Neudata · *#ClearDataClearImpact*

> **Status:** outline. Materials will be developed after Level 1.

1. Why Level 2

Level 1 shows the whole research journey. The instructor runs the analyses and participants read and interpret the output. Level 2 is for participants who want to run the analyses themselves.

Level 2 is a statistics course taught in R. It covers the regression models that clinical researchers meet most often, one family per session: linear regression for continuous outcomes, logistic regression for binary outcomes, Poisson, negative binomial and zero-inflated models for counts, and Kaplan-Meier and Cox models for time to an event. For each model we ask the same questions:

- Which questions and outcomes is this model for?
- What should I look at in the data before fitting it (exploratory data analysis, EDA)?
- What do the parameters mean, in the units of the data?
- What does the model assume, and how do I check it?
- What do I do when an assumption fails?
- How do I interpret and report the result?

Practical skills are taught only as far as the analysis needs them: importing and cleaning a dataset, describing it in a Table 1, handling missing data simply, and writing clear R scripts. Building fully reproducible projects (folder structures, version control, Quarto reports) and producing figures that meet journal file requirements are Level 3 topics.

By the end of Level 2, participants will be able to:

1. Import, clean and describe a clinical dataset in an R script, and produce a Table 1.
2. Describe missing data and explain the difference between a complete-case analysis and multiple imputation.
3. Choose a suitable regression model from the type of outcome and the research question.
4. Explore the data before modelling, using summaries and plots suited to the analysis problem.
5. Fit, check, interpret and report linear, logistic, count (Poisson, negative binomial, zero-inflated) and Cox models, including confounding and interaction.
6. Analyse a new question with their own R scripts and write a Results section from the output.

2. Who it is for and prerequisites

Participants	8 to 12 health professionals (doctors, nurses, etc.) from the client department who want to analyse their own data.
---------------------	---

Prerequisites	(1) Level 1 completed. (2) The free Neudata course <i>Clinical Data Analysis in R</i> , or equivalent basics: RStudio, objects and data frames, importing a CSV, basic tidyverse verbs, a simple plot. A short readiness check (30 minutes, self-marked) is sent before the course.
Equipment	Own laptop with R, RStudio and the course packages installed before Session 1 (installation guide and a 30-minute online set-up clinic). A second screen helps: one for the shared screen, one for RStudio.
Commitment	3 hours per session plus about 2 to 3 hours of homework per week; agreement from the line manager for protected time.

3. Format

Delivery	Online, live. Participants join from home and code on their own laptops while the instructor shares the screen.
Sessions	8 sessions of 3 hours over about 4 to 6 weeks (for example two sessions a week, or one a week with a one-week break mid-course). The count and length are unchanged from the first outline; the content is reorganised around statistical models.
Session rhythm	20 to 30 min concepts. About 2 hours of live coding and exercises: the instructor codes, participants follow, then practise on their own. 20 min recap, questions and homework briefing. One 15-minute break.
Working together	The whole class works as one group; there are no breakout rooms. Exercises are done individually on each participant's laptop. Participants post answers and output in the chat, and the instructor (or a volunteer sharing their screen) goes through them with the class. Help with code errors is given in the chat or by a participant sharing their screen.
Structure of each modelling session	(1) The question and the outcome type. (2) EDA for this problem : the summaries and plots that should come before this model. (3) The model and what its parameters mean. (4) Assumptions and diagnostics, and what to do when they fail. (5) Interpretation and a reporting table, with a written sentence. Simple ggplot2 code for the EDA and result plots is given in each session; there is no separate session on figures.
Shared dataset	FLUID-ICU (<i>Course/Data/</i>): the simulated ICU sepsis cohort (raw and clean data, data dictionary, cleaning log, statistical analysis plan, R scripts) used in Level 1. Count outcomes need a count variable that FLUID-ICU does not yet have (see Section 8, open items).
Own data	Each participant may bring a de-identified dataset (with the necessary ethics or data-protection approval). Two online one-to-one slots of 30 to 45 minutes (after Sessions 2 and 5) help adapt the course scripts to their data.

Support	Shared course folder (data, starter scripts, solutions after each session), an online question channel, and an instructor office hour each week.
Main R packages	tidyverse (readr, dplyr, tidyr, ggplot2), readxl, janitor, gtsummary, flextable, naniar, mice, broom, car, performance, DHARMA, MASS, pscl, pROC, splines, survival, ggsurvfit.

4. Session-by-session outline

Session 1: Preparing and describing data: cleaning, Table 1 and missing data

- **Concepts:** raw data are never edited; the cleaning log and the data dictionary; tidy data; choosing the right summary (mean and SD, median and IQR, n and %); what Table 1 is for, and why p-values do not belong in Table 1 of an RCT.
- **Working with R scripts:** one script per step, numbered (01_clean.R, 02_describe.R); a short header and comments; packages loaded at the top; reading the data from the course folder. The course folder is provided ready-made.
- **Cleaning (FLUID-ICU raw file):** import CSV and Excel; check structure and types; find duplicates; recode missing-value codes (999, -99); harmonise categories (M, Male, 1); parse dates; range and consistency checks; derive cumulative fluid balance in ml/kg; set factor levels and reference categories. Each change is written in the cleaning log.
- **Describing:** distributions and skewness (histograms, box plots); Table 1 by fluid-balance group with **gtsummary** (labels, units, the right statistic per variable type, missing counts, footnotes); export to Word with **flextable**.
- **Missing data, the basics:** count and visualise missing values (**naniar**); compare patients with and without lactate; MCAR, MAR and MNAR in one slide each; what a complete-case analysis assumes; what multiple imputation does, shown once with the result from `Course/Data/R/05_missing.R` (lactate: complete-case OR 1.39 vs imputed OR 1.45). Participants do not run imputation themselves in Level 2. Sensitivity analyses are mentioned as the next step and taught in Level 3.
- **FLUID-ICU:** the clean file matches the course clean file; Table 1 matches the Level 1 results pack.
- **Deliverable:** 01_clean.R, 02_describe.R, a cleaning log and Table 1 as a Word table.

Session 2: Linear regression in depth

- **Opening map:** which model for which outcome (continuous: linear; binary: logistic; count: Poisson or negative binomial; time to event: Cox). This map is used again in every session.
- **EDA for a continuous outcome:** histogram and Q-Q plot of the outcome; box plots of the outcome by each categorical predictor; scatter plots with a smoother for each continuous predictor; checking for skewness, outliers and floor or ceiling effects.
- **The model:** the model as a line; what the intercept and slope mean in the units of the data; continuous and categorical predictors; centring; adjusted coefficients; confidence intervals and p-values; R^2 .
- **Confounding:** choosing adjustment variables from the question and a simple causal diagram (DAG), not from p-values; crude vs adjusted estimates.

- **Assumptions and diagnostics:** linearity, independence, constant variance, normal residuals; residual and Q-Q plots (`performance::check_model`); influential points; multicollinearity (VIF).
- **When assumptions fail:** log-transforming a skewed outcome and back-transforming to a ratio of geometric means; non-linear terms with splines; when a non-parametric comparison is enough.
- **Interaction:** fitting, interpreting and plotting an interaction term.
- **Reporting:** a regression table with `gtsummary`; the sentence "each 10 ml/kg higher fluid balance was associated with ...".
- **FLUID-ICU:** ICU length of stay in survivors (Q5: adjusted ratio of geometric means 1.24); ventilator-free days (Q3) as an example of an outcome that a linear model fits badly (many zeros from deaths, many 28s), which we come back to in Session 5.
- **Deliverable:** `03_linear.R`, a diagnostics plot and a reporting table with a written interpretation. *One-to-one slot 1 follows this session.*

Session 3: Logistic regression in depth

- **EDA for a binary outcome:** outcome proportions by each categorical predictor (with 95% CIs); box plots of continuous predictors by outcome; the proportion with the outcome across groups (deciles) of a continuous predictor, to see whether the relationship looks linear on the log-odds scale; checking that every cell has enough events.
- **The model:** odds and odds ratios; the logit link; what the coefficients mean; ORs for continuous predictors per meaningful unit; crude vs adjusted ORs (confounding by indication made visible).
- **Model building:** events per variable and overfitting; why stepwise selection misleads; confounding vs interaction (effect modification) and how to test and report an interaction.
- **Assumptions and checks:** linearity in the logit (splines compared with categories); separation and sparse data; influential observations; residuals with **DHARMA**.
- **Performance basics:** discrimination (ROC curve, AUC with **pROC**) and calibration (observed vs predicted by risk group); why these matter more for prediction than for explanation.
- **Ordinal outcomes, introduced:** proportional-odds logistic regression for an ordered outcome such as AKI stage 0 to 3 (`MASS::polr`), and its key assumption.
- **Reporting:** a table of crude and adjusted ORs with 95% CIs; a forest plot of the ORs; STROBE wording.
- **FLUID-ICU:** positive fluid balance and 28-day mortality (Q1: crude OR 2.49, adjusted OR 1.62), with the heart-failure interaction; fluid balance and AKI (Q2).
- **Deliverable:** `04_logistic.R`, a crude vs adjusted OR table, a forest plot and a short interpretation.

Session 4: Count outcomes (I): Poisson and negative binomial regression

- **Count outcomes in clinical research:** number of infections, readmissions, falls, organ-support days, exacerbations. Counts and rates; why linear and logistic models are the wrong tools.
- **EDA for a count outcome:** a bar chart of the count distribution (how many zeros, how long the tail); mean vs variance overall and by group (a first check for overdispersion); counts against follow-up time; event rates by group.
- **Poisson regression:** the log link; rate ratios and how to interpret them; the offset for different follow-up times (for example days in ICU), and when an offset is needed.

- **Assumptions and checks:** mean equal to variance; independence; overdispersion tests and residual checks (`DHARMA`, `performance::check_overdispersion`); what overdispersion does to standard errors and p-values.
- **When Poisson fails:** negative binomial regression (`MASS::glm.nb`); robust (`sandwich`) standard errors as an alternative; comparing Poisson and negative binomial fits (AIC, residual plots, observed vs predicted counts).
- **Reporting:** a table of rate ratios with 95% CIs; the sentence "the rate of ... was X times higher ...".
- **FLUID-ICU:** a simulated count outcome (proposed: number of ICU-acquired infections, with ICU stay as the offset; see open items).
- **Deliverable:** `05_count.R`, a Poisson vs negative binomial comparison and a rate-ratio table with interpretation.

Session 5: Count outcomes (II): excess zeros and zero-inflated models

- **Where extra zeros come from:** two kinds of zero (patients who could never have the event, and patients who could but did not); why a Poisson or negative binomial model then underestimates the number of zeros.
- **EDA for excess zeros:** observed zeros vs the number expected under a Poisson or negative binomial fit; the count distribution split by a likely "structural zero" group; a rootogram of observed vs fitted counts.
- **Zero-inflated models:** zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) with `pscl` (`zeroinfl`); the two parts of the model (the chance of a structural zero, and the count among the rest) and how to interpret each.
- **Hurdle models:** how they differ from zero-inflated models, and how to choose between them from the clinical meaning of a zero.
- **Model choice and checks:** comparing Poisson, negative binomial, ZIP, ZINB and hurdle fits (AIC, observed vs predicted zeros, rootograms); why a model should be chosen from the data-generating story first and fit statistics second.
- **Bounded counts:** ventilator-free days is a count of days out of 28, with zeros for deaths. We discuss why it is hard to model (zeros, a ceiling at 28, death as a competing event) and the options used in the literature: a hurdle or zero-inflated model, or a zero-inflated binomial (or beta-binomial) model for "days out of 28". The zero-inflated binomial model is shown, not practised.
- **Reporting:** what to report for a two-part model; plain-language interpretation.
- **FLUID-ICU:** ventilator-free days (Q3) reanalysed with a hurdle model and compared with the linear model from Session 2; a simulated zero-heavy count (proposed: days on renal replacement therapy by day 28; see open items).
- **Deliverable:** `06_zero_inflated.R` and a short comparison of models with one paragraph of interpretation.

Session 6: Survival analysis: Kaplan-Meier and the Cox model

- **Time-to-event data:** censoring, and why ignoring it gives the wrong answer; the survival and hazard functions in plain words; the analysis data layout (time and event status).

- **EDA for time-to-event data:** follow-up time and events by group; when censoring happens and whether it differs between groups; Kaplan-Meier curves by each key categorical predictor, with numbers at risk (`ggsurvfit`).
- **Kaplan-Meier and the log-rank test:** survival at a fixed time with a 95% CI; median survival; what the log-rank test compares.
- **The Cox model:** the hazard ratio and what it means (and does not mean); crude and adjusted HRs; continuous predictors per meaningful unit.
- **The proportional-hazards assumption:** what it means; checking it with Schoenfeld residuals (`cox.zph`) and log-minus-log plots; what to do when it fails (stratification; splitting follow-up time, introduced). Time-varying covariates and competing risks are mentioned and taught in Level 3 (Module D).
- **Other checks:** functional form of continuous predictors; influential observations.
- **Reporting:** a Kaplan-Meier figure with numbers at risk; a table of crude and adjusted HRs with 95% CIs; the PH check stated in the Methods.
- **FLUID-ICU:** positive fluid balance and 28-day survival (Q1: Kaplan-Meier 81.6% vs 64.1% at day 28; adjusted HR 1.46; PH global test $p = 0.10$), compared with the logistic result from Session 3.
- **Deliverable:** `07_survival.R`, a Kaplan-Meier figure, an HR table with the PH check and a short interpretation. *One-to-one slot 2 follows this session.*

Session 7: Analysis project (I): choosing and running the right model

- **Set-up:** each participant receives a different analysis question with a short brief (the question, the outcome, the dataset, the expected model family). Questions are spread across the model types taught: linear, logistic, count (Poisson, negative binomial or zero-inflated) and Cox. Data are FLUID-ICU, a second simulated dataset prepared for the course, or the participant's own data if the instructor agrees.
- **In class:** the instructor works one short question with the whole class from start to finish (question, EDA, model choice, checks, interpretation) as a model of the workflow. Participants then work individually on their own question in R scripts, with the instructor answering questions in the chat and by screen share. Mid-session, a few participants share their screen to show their EDA and model choice to the class.
- **Workflow checklist:** clean and describe the data; Table 1; EDA suited to the outcome; choose the model and the adjustment variables; check assumptions; interpret; save tables and plots.
- **Deliverable (homework):** a set of R scripts that run from the data to all tables and plots, and a short note on the model chosen and the checks done.

Session 8: Analysis project (II): writing the Results section and sharing it

- **Writing Results:** from output to text; effect sizes with 95% CIs before p-values; calibrated language; numbers that match between text, tables and figures; the STROBE (or CONSORT) items for Results; the statistical-analysis paragraph for the Methods.
- **In class:** the instructor writes the Results paragraph for the worked question with the class. Participants then draft their own Results section individually.
- **Sharing back:** in the last hour, several participants (one for each model type where possible) present their question, model choice, checks and main result to the class in 5 minutes. The instructor and the class comment on the choice of model, the checks and the wording.

- **Deliverable:** a Results section (about 300 to 500 words) with Table 1, one main results table and one figure, plus the R scripts that produced them.

5. Per-session deliverables (summary)

Session	Deliverable
1	Cleaning and description scripts, cleaning log, Table 1 (Word)
2	Linear model script, diagnostics plot, reporting table and interpretation
3	Logistic model script, crude vs adjusted OR table, forest plot, interpretation
4	Count model script, Poisson vs negative binomial comparison, rate-ratio table
5	Zero-inflated or hurdle model script and model comparison with interpretation
6	Survival script, Kaplan-Meier figure, HR table with PH check, interpretation
7	R scripts for the individual analysis question and a note on model choice and checks
8	Results section with Table 1, a results table and a figure, plus the scripts

Each deliverable is submitted to the course folder before the next session; instructors give short written feedback.

6. Final project (optional)

Within **4 weeks after Session 8**, participants who want to go further may submit an analysis of **their own data** (preferred, with approval) or of a second course question with a different model type from their Session 7 to 8 question:

1. **R scripts** that run from the clean data to all tables and plots without errors, with a short analysis plan written before the main analysis.
2. A **draft Results section** (about 500 to 800 words) with Table 1, one main results table and one figure, plus the statistical-analysis paragraph for the Methods.

A fully reproducible project with a Quarto report is not required; that is taught in Level 3.

7. Assessment

Component	Weight	How it is assessed
Session deliverables (Sessions 1 to 6)	40%	Completed and running (pass or fail per deliverable, with short feedback)
Analysis project: scripts and model (Session 7)	30%	The scripts run; the model suits the question and outcome; EDA done; assumptions checked; interpretation correct

Component	Weight	How it is assessed
Analysis project: Results section (Session 8)	30%	Effect sizes with 95% CIs; tables and text agree; guideline items covered; calibrated language

Marks are used for feedback; there is no ranking.

8. Materials to be developed and open items

Material	Content
Slides (EN + VN)	8 short concept decks (20 to 30 slides each) with speaker notes; each modelling deck follows the same five parts (question, EDA, model, assumptions, reporting)
R scripts	Starter scripts (with gaps) and complete instructor scripts for each session, built from the FLUID-ICU scripts in <code>Course/Data/R/</code> ; ready-made ggplot2 code for the EDA and result plots of each model
Exercises	One exercise sheet per session (core and stretch tasks) on FLUID-ICU, with a "your data" task
Solutions	Full worked solutions and output, released after each session
Analysis-project briefs	8 to 12 individual questions across the model types (linear, logistic, count, zero-inflated, Cox), each with data and a marking guide
Checklists	Model-choice map; assumptions checklist per model; Results-section checklist (STROBE or CONSORT)
Set-up guide	Installation guide, package list and readiness check

Open items

- **FLUID-ICU needs count variables.** The dataset has no true count outcome. Proposed additions to the generator (`Course/Data/R/00_generate_fluid_icu.R`), the data dictionary and `key_results.md`: (1) `icu_infections_n`, number of ICU-acquired infections during the ICU stay, overdispersed, with `icu_los_days` as the offset (Session 4); (2) `rrt_days_28`, days on renal replacement therapy by day 28, zero for patients without RRT (Session 5; links to the existing `rrt` variable). Alternatives: number of ICU readmissions within 90 days, or organ-support days. Until then, Session 5 can use ventilator-free days, and Session 4 needs a small separate simulated count dataset.
- **A second simulated dataset** may be needed for the analysis-project questions, so that each participant has a different question and not all use FLUID-ICU.