

Course/Data: the FLUID-ICU course dataset

Level 1 uses it for analysis and manuscript writing in Weeks 5-8. Level 2 reuses it hands-on.

> **SIMULATED DATA: NOT REAL PATIENTS.** FLUID-ICU is an invented study. Every value was produced by a computer program with a fixed random seed. It does not describe any real patient, hospital, ICU or country. Any manuscript written from it must say that the data are simulated and must not be submitted as real research.

FLUID-ICU is a simulated multicentre prospective cohort of **620 adults with sepsis** admitted to 4 ICUs ("ICU A" to "ICU D") from January 2023 to December 2024. It carries the course's running example into Sessions 5-8: *is a positive cumulative fluid balance by day 3 associated with higher 28-day mortality?* The data were built so that the classic lessons show up in the results: confounding by indication, missing data, a skewed outcome, and an interaction worth discussing. Clinical ranges follow the two course papers (Boulet 2024; Hyun 2023).

What is in this folder

File / folder	What it is	Used in
fluid_icu_raw.csv, fluid_icu_raw.xlsx	The data as the 4 ICUs "sent" it: 623 rows with realistic errors	S5 cleaning demo
fluid_icu_clean.csv, fluid_icu_clean.rds	Analysis-ready data: 620 patients, 32 variables (.rds keeps the R factor levels)	All analyses
data_dictionary.md, data_dictionary.csv	Every variable: label, type, units/codes, allowed range, missing code, notes	S5
cleaning_log.md	Every problem found in the raw file and what was done, in plain language	S5
SAP_FLUID-ICU.md	The statistical analysis plan, written "before the analysis" (14 sections of the course SAP template)	S6, S7
key_results.md	Every number the slides and handouts quote, with its source script	Instructors
results_packs/Q1 ... Q5	One pack per group question: results_pack_QN.md, tables/ (CSV + Word) and figures/ (PNG)	S5-S8 groups
results_packs/Results_Pack_QN.docx	The same pack as one Word document, for printing	S5-S8 groups
R/	The scripts (below)	Instructor demos
python/make_slide_figure_s.py	Draws the slide figures in English and Vietnamese into Course/Resources/figures/{en,vi}/data/	Deck build
output/	What the demo scripts write: tables/, figures/, slide_data/ and fluid_icu_clean_from_raw.csv	Demos

File / folder	What it is	Used in
FLUID-ICU.Rproj	Open this in RStudio: it sets the working folder to <code>Course/Data</code>	Demos

The five group questions (blueprint section 1)

Group	Question	Main method	Pack
1	Positive fluid balance by day 3 and 28-day mortality (primary)	Logistic regression (crude vs adjusted), Kaplan-Meier, Cox, interaction with heart failure	Q1
2	Cumulative fluid balance and acute kidney injury	Logistic regression, balance per 10 ml/kg	Q2
3	Cumulative fluid balance and ventilator-free days	Mann-Whitney; linear regression	Q3
4	Admission lactate and 28-day mortality; effect of missing lactate	Complete case vs multiple imputation (mice, $m = 20$)	Q4
5	Fluid balance and ICU length of stay in survivors	Median (IQR); Mann-Whitney; linear regression on log scale	Q5

How to run the scripts

Open `FLUID-ICU.Rproj` in RStudio (or set the working directory to `Course/Data`) and run the scripts **in order**. From a terminal:

```
cd Course/Data
Rscript R/00_generate_fluid_icu.R # (course team only) simulate; writes raw + clean files
Rscript R/01_clean.R             # S5 demo: raw -> clean; confirms identical to the clean file
Rscript R/02_describe.R          # S5 demo: Table 1 + histogram, boxplot, bar chart
Rscript R/03_compare.R           # S6 demo: t-test, Mann-Whitney, chi-square
Rscript R/04_models.R            # S6 demo: crude vs adjusted OR, forest plot, KM, Cox, interaction
Rscript R/05_missing.R           # S6 demo: missing lactate, complete case vs multiple imputation
Rscript R/06_results_packs.R     # (course team) results packs, key_results.md, slide data (~1 min)
python python/make_slide_figures.py # (course team) slide figures EN + VN (needs Deck_and_Build/figures)
```

Participants never need to run anything: they read the printed outputs and their results pack. The demo scripts are written to be run **block by block** in RStudio (Ctrl+Enter), with a comment before each block saying what to point out. Scripts 01-05 only read the data files and write to `output/`. Script 00 overwrites the data files (with the same result every time, because the seeds are fixed).

R packages needed

R 4.3 or later (built and tested with R 4.6.1).

```
install.packages(c("tidyverse", "janitor", "writexl", "readxl", "gtsummary", "gt",
                  "flextable", "officer", "broom", "survival", "survminer", "mice"))
```

Python (slide figures only): Python 3.10+ with `matplotlib`, `numpy`, `pandas`; fonts Century Gothic (English) and Segoe UI (Vietnamese).

Things to know before teaching

- **Exact reproduction.** `01_clean.R` ends by checking that its output equals `fluid_icu_clean.csv` value for value and byte for byte. If it prints `FALSE`, the raw or clean file has been edited by hand.
- **Numbers.** `key_results.md` lists all quoted numbers. If you re-run `00_generate_fluid_icu.R` on another R version, random numbers could in principle differ. In that case re-run 01-06 and the Python script, and check `key_results.md` again.
- **Simulated means simulated.** The fluid "effect" was put into the data on purpose (mainly in heart failure) together with strong confounding by severity. The data cannot tell anyone whether fluid harms real patients.