Mission by Mission

Reference edition — the complete mission-by-mission walkthrough behind the talk. Every number below was produced by the run recorded in this repository.

Hsieh-Ting Lin, MD　林協霆

Department of Medical Oncology　腫瘤內科部

Koo Foundation Sun Yat-Sen Cancer Center　和信治癌中心醫院

Disclaimer and Conflict of Interest

- ⚠ **The data in this talk are entirely synthetic**
  - Simulated from the published summaries of the IMbrave150 trial[1]
  - No real patient is represented, and no record derives from any chart
  - Nothing here is evidence about atezolizumab, bevacizumab or sorafenib
- ◎ **What the talk is actually about**
  - The *workflow* — how to instruct an agent so its output can be audited
  - The recovered hazard ratio is a teaching target, not a finding
- ✓✓ **Conflict of interest**
  - The author declares no financial or non-financial conflict of interest
  - No industry funding, no honoraria, no relationship with any AI vendor

[1] Finn RS, Qin S, Ikeda M, et al. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *N Engl J Med*. 2020;382(20):1894-1905. doi:10.1056/NEJMoa1915745

What You Will See

- **The input**
  - 10 hospital CSV exports, 3 incompatible EHR schemas, 1,800 patients
- **The output**
  - A typeset preprint PDF whose every number and every reference is checked
- **The operator**
  - One person, typing prompts. No code written by hand.
- **Two things to watch on every slide**
  - **Current step** — what the agent is being asked to do right now
  - **Good prompt** — why *that* wording, and what a bad one would cost

PART I

The Operating Rhythm

The Unit of Work Is One Mission

- ⏸ **Current step**
  - Do one mission → run its acceptance script → summarise → **stop**
  - The agent is forbidden to continue because "the next step is obvious"
- 💬 **Good prompt**
  - State the stopping condition *in the instructions file*, not per turn
  - Give the halt a literal token the agent must emit

> `Do one mission. Run verify/chNN.py. Say in three lines what we now see. Then output PAUSED and wait.`

- ⚠ Without an explicit halt the agent will finish the whole project in one turn — correct, unwatchable, and unreviewable

A Prompt That Survives Contact

- 🎯 **Five parts, in this order**
  - **Goal** — one sentence, in the user's language, not the tool's
  - **Contract** — exact output paths, exact column names, exact value domains
  - **Constraints** — the failure modes you already know about
  - **Acceptance** — the command that decides pass or fail
  - **Stop** — what to emit, and to do nothing after it
- ✓✓ **Why the contract matters most**
  - A named file is checkable; "save the results" is not
  - If the schema is not pinned, the agent invents plausible column names — and every downstream step then fails for the wrong reason

Constraints Are Cheaper Than Debugging

- ⚠ **Current step**
  - Front-load every landmine you have already stepped on
- 💬 **Good prompt** — real examples from this repo

> Never call `lifelines.plotting.add_at_risk_counts()`; this environment raises `TypeError`. Use the manual version below.
>
> Set `matplotlib.use("Agg")` *before* importing `pyplot`.
>
> Weighted Cox must pass `weights_col=` **and** `robust=True`.
>
> After every merge, `assert len(df) == expected`.

- 👁 The fourth one is the dangerous kind: omit `robust=True` and nothing errors — the point estimate is right and the confidence interval is silently wrong

PART II

Ten Hospitals, Three Dialects

Mission 0.1 — Open the Box, Change Nothing

- **Current step**
  - Copy the raw exports into a sandbox; read column names and one row
  - Write no transformation code at all
- **Good prompt**
  - Ask for an *observation*, and forbid the fix

> Copy `hospitals/` into `workspace/hospitals/` — copy, do not symlink. Print the columns and first row of H01, H02, H03. Then tell me three things that look wrong.
> **Do not start writing a converter.**

- Agents default to solving. Naming the problem first is what makes the next three missions legible to an audience

Mission 1.1 — Diagnose the Dialects

- **Current step**
  - Cluster the 10 files by column fingerprint, not by filename
  - Compare value encodings, units, and per-column missingness
- **Good prompt**
  - Force evidence over recall: `unique()`, not "typical values"
  - Supply an independent check the agent can grade itself against
  - Fan out — three subagents, one per dialect family, then merge
- **What it finds**
  - Albumin in g/L at some sites, g/dL at others — one order of magnitude
  - Three sites have **no** continuous AFP column at all

Mission 1.2 — Harmonise Into One Table

- **Current step**
  - `detect_vendor()` → `harmonise_file()` → `load_pooled()` → `pooled.csv`
- **Good prompt**
  - Hand over the canonical schema as a table: 27 names, types, domains
  - Say where the medians should land, so unit errors surface immediately

> `albumin_g_dl` — float, **g/dL**, median should be near 3.9
> `child_pugh_score` — 5 or 6 as an **integer**, not `A5`

- ⚠ **Never `dropna()`** — missingness here is structural, not random. Dropping rows would delete three hospitals and the confounding with them

Mission 1.3 — The First Confounder Appears

- **Current step**
  - Left-join hospital attributes; profile each site; sort by % treated
- **What the table shows**
  - Treatment share ranges widely across the ten sites
  - The sites that treat most are also the sites with better prognosis
- **Good prompt** — end on a question, not a task

> If the hospital decides both *which drug a patient gets* and *how that patient does*, what happens when you compare the two arms directly?

- The answer to that question is the entire next section — and the audience arrives at it before you say it

[9] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

PART III

The Unadjusted Answer Is Lying

Mission 2.1 — Ask the Naive Question First

- **Current step**
  - Fit a Cox model on treatment alone. One covariate. Nothing else.
  - Result: **HR 0.505 (95% CI 0.43–0.60)**
- **Good prompt**
  - Pin the output format so the number is machine-readable later

> `Write naive_hr.json as {"n":…, "hr":…, "ci_low":…, "ci_high":…, "p":…}.`
> `Then tell me in one sentence what you would conclude from this number alone.`

- ⚠ This is the slide the audience believes. Let them — it is the only way the correction lands

Mission 2.2 — Prove It: The Arms Differ at Baseline

- ⚖ **Current step**
  - Standardised mean difference for all 11 covariates, by arm
  - Seven covariates exceed |SMD| 0.15 before any adjustment
- 💬 **Good prompt**
  - Write the formula out; do not let the agent pick a variant

```
SMD = (mean_t − mean_c) / sqrt((var_t + var_c) / 2)
```

- 👁 **Then ask for direction, not just magnitude**
  - Which covariates are imbalanced, which way do they lean, and does that bias the naive HR up or down?
  - "Report the SMDs" gets a table; this gets a diagnosis

[10] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107. doi:10.1002/sim.3697

Mission 2.3 — Propensity-Score Matching

- ⚖ **Current step**
  - Logistic PS on the same 11 covariates → 1:1 nearest neighbour, no replacement
  - Caliper = 0.2 × SD of **logit(PS)**[3] = 0.114
  - **706 pairs, 1,412 patients — 256 treated patients find no match**
- 💬 **Good prompt** — three clauses that decide reproducibility

> The caliper is 0.2 × SD of **logit(PS)**, not of PS.
> Sort treated by `logit_ps` ascending **before** matching.
> Report how many treated were left unmatched.

- ⚠ Greedy matching depends on order. Unstated order means a different cohort every run — and an unauditable manuscript

[2] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786 · [3] Austin PC. Optimal caliper widths for propensity-score matching. *Pharm Stat*. 2011;10(2):150-161. doi:10.1002/pst.433

Mission 2.4 — Did the Matching Actually Work?

- **Current step**
  - Love plot: |SMD| before vs after, reference line at 0.10
  - Propensity-score overlap, before and after
  - Largest remaining |SMD| after matching: **0.044**
- **Good prompt**
  - Specify the plot, not the impression: axes, reference line, pairing
  - Then ask the question the plot exists to answer

> `Is there any region of the propensity score where one arm has no patients at all?`

- ⚠ Do **not** ask for balance p-values — they measure sample size, not balance

PART IV

Survival, and Whether It Holds

Mission 3.1 — Kaplan–Meier, Log-Rank, Cox

- **Current step**
  - On the matched cohort: **OS HR 0.578 (0.48–0.70)**, p < 0.001
  - 12-month survival 71.1% vs 55.7%; median 21.3 vs 13.5 months
- **The point of the whole exercise**
  - The published randomised trial reported **HR 0.58**[1]
  - Naive 0.505 → adjusted 0.578. The gap was confounding, and it is recoverable
- **Good prompt**
  - Ask it to show that raw percentages mislead *before* it draws the curve — censoring becomes visible instead of asserted

[1] Finn RS, Qin S, Ikeda M, et al. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *N Engl J Med*. 2020;382(20):1894-1905. doi:10.1056/NEJMoa1915745

Mission 4 — Subgroups Test Consistency, Not Winners

- **Current step**
  - 13 pre-specified subgroups, one univariable Cox each, then a forest plot
  - All hazard ratios fall between 0.51 and 0.63; none crosses 1.0
- **Good prompt** — the third question is the teaching one

> `ECOG 0 gives 0.61 and ECOG 1 gives 0.54.`
> **`May I say that ECOG 1 patients benefit more?`**

- **The answer is no**, and the agent should say why: overlapping intervals, no interaction test, subgroups pre-specified for consistency
- An agent asked "which subgroup benefits most" will happily answer

The result survived one analysis. Now try to break it.

Part V — Robustness

Mission 5.1 — One Hundred Twenty Analytic Paths

- **Current step**
  - Data held fixed; only analytic choices vary
  - 15 covariate sets × 8 adjustment methods = **120 specifications**
  - Median HR **0.582**, IQR 0.560–0.609, range 0.480–0.716 *(this run)*
- **Good prompt**
  - Enumerate the grid explicitly — the agent must not choose the sensitivity analyses that flatter the result
- **Report the honest number**
  - Only **63%** of specifications land in 0.55–0.61 — and the range, unlike the median, moves with implementation choices
  - That fraction belongs in the Limitations, not in the abstract

[4] Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*. 2016;11(5):702-712. doi:10.1177/1745691616658637

Mission 5.2 — The Specification Curve

- **Current step**
  - 120 specifications sorted by HR, each with its interval
  - Reference lines at 0.58 (trial) and 0.505 (naive)
- **Good prompt**
  - Ask for a title that states the conclusion, so the figure can be read without you standing next to it

> `Title the figure:`
> *`Every reasonable adjustment lands near 0.58 · 120 specifications`*

- This is the slide that answers "did you just pick the analysis that worked?" — and it answers it before anyone asks

[5] Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav*. 2020;4(11):1208-1214. doi:10.1038/s41562-020-0912-z

Mission 5.3 — Change the Estimand Entirely

- **Current step**
  - Landmark 12-month mortality risk difference, TMLE with IPCW[6]
  - **RD −0.151 (−0.215 to −0.086)** — a different target, same direction
- **Good prompt**
  - Specify the censoring rule rather than letting it be inferred

> `Y = 1` if death before 12 months; `Y = 0` if alive at 12 months;
> if censored before 12 months, **`Y` is unknown** — set `Delta = 0`.

- ⚠ Left implicit, an agent will complete-case this and quietly bias the estimate. Naming the third state is the whole prompt

[6] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):Article 11. doi:10.2202/1557-4679.1043

PART VI

Writing the Paper Backwards

The Order of Writing Is the Method

- **Current step** — write the sections in this order, never another
  - **Methods** first — reconstructed by re-reading the scripts, not from memory
  - **Results** — every number pulled from a file in `workspace/`
  - **Discussion**, then **Introduction** — the intro exists to set up the discussion
  - **Conclusions**, then **Abstract**, then **Title** — last, once there is something to name
- **Good prompt**
  - "Re-read `harmonize.py`, `psm.py`, `multiverse.py`, `tmle.py`. Describe what the code **did**, not what you intended."
  - Ask for the software versions from `pip list`, not from the agent's memory
- Written in any other order, the Introduction promises a paper you did not write

Mission 6.2 — Every Number Carries Its Source

- 🎯 **Current step**
  - Draft Results reading **only** from the output files
  - Tag each subsection with the file it came from
- 💬 **Good prompt**

> `Take every number from files in workspace/. Do not copy from our conversation. End each subsection with <!-- src: … -->.`

- ⚠ **The failure this prevents**
  - An agent recalling "HR was about 0.58" writes a plausible number
  - Plausible numbers pass human review and fail the audit in Mission 6.9
  - Source tags turn a reviewer's spot check into a scripted one

Mission 6.8 — Every Reference Verified at Crossref

- **Current step**
  - Search for the real paper → resolve the DOI → confirm six bibliographic fields against the Crossref record
  - One subagent per reference; this stage is network-bound and embarrassingly parallel
- **Good prompt**
  - "Never write a DOI from memory. Search, then verify. A reference that fails verification is deleted, not guessed at."
  - Send a `mailto:` User-Agent — anonymous Crossref requests get rate-limited mid-demo
- ⚠ Fabricated citations are the single most reputationally expensive failure mode of this whole workflow

[11] Crossref REST API. Accessed August 2026. https://api.crossref.org

Mission 6.9 — The Numeric Audit

- ✓✓ **Current step**
  - A script re-reads every output file, re-extracts every number in the manuscript, and compares them
  - Any mismatch fails the build. Not a warning — a failure.
- 💬 **Good prompt**
  - The audit is not written by the agent that wrote the manuscript
  - It is fixed in advance, in the repo, and the agent is told the command

```
python verify/ch06.py --mission 6.9
```

- 🎯 This is what "auditable" means operationally: a claim of correctness that does not depend on trusting the writer

PART VII

Reviewing Its Own Work

Mission 7.1 — Four Independent Reviewers

- **Current step**
  - Four subagents dispatched in one message, blind to one another
  - Statistical · Clinical · Reporting standards (STROBE[7], RECORD[8]) · Reproducibility
- **Good prompt** — two clauses do the real work

> Each reviewer may read the manuscript and `workspace/` output files **only**. Reading the course material is forbidden.
>
> Each must raise at least three major comments. If you find none, read it again.

- ⚠ Give a reviewer the answer key and you get an answer check. Let it off with "looks good" and you got no review at all

[7] von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. *Lancet*. 2007;370(9596):1453-1457. doi:10.1016/S0140-6736(07)61602-X · [8] Benchimol EI, Smeeth L, Guttmann A, et al. The RECORD statement. *PLoS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885

Mission 7.2 — Triage, Respond, Actually Edit

- **Current step**
  - Merge and de-duplicate into `review_log.csv`
  - Each comment gets ACCEPT / PARTIAL / **REJECT** — and a file that changed
- **Good prompt**
  - "Every REJECT needs a technical reason. 'We disagree' is not one."
  - "`action_taken` may not be left blank."
  - "Re-run the numeric audit immediately after editing."
- ⚠ Left unconstrained, an agent writes a courteous response letter and changes nothing. The edited file is the deliverable, not the letter

Mission 7.4 — Removing the Machine Voice

- **Current step**
  - Deterministic pattern check first, report-only, count the hits
  - Back up, rewrite, then diff — and re-run the numeric audit
- **Good prompt**
  - Calibrate on the author's real writing samples, not on "sound natural"
  - Forbid new claims: rewriting may change words, never numbers
- **The order matters**
  - Humanise **after** the audit passes, then audit again
  - A rewriter with no backup and no diff is an unreviewable change

PART VIII

Shipping It

Mission 8 — From Markdown to a Submittable File

- **Current step**
  - One build pipeline → PDF and DOCX in parallel; they share no dependency
  - Bibliography rendered from the verified `refs.bib`
- **Good prompt**
  - Name the flag that fails silently

> `pandoc` must be called with `--citeproc`. Without it there is no error — the raw `[@key]` markers are typeset into the PDF.

- **Then verify the artefact, not the command**
  - Extract text from the built PDF and assert no `[@` survives
  - "The build exited zero" is not evidence that the build is correct

What CI Does in This Repo

- **Current step**
  - Push to `main` touching `slides/**` → the deck is re-rendered to PDF
  - The PDF is uploaded as a build artefact on every push
  - Push a tag `slides-v*` → a GitHub Release is cut with the PDF attached
- **Why bother for a talk**
  - The version being projected is the version in the repository
  - A slide corrected at 23:00 the night before cannot diverge from the file
- Same principle as the numeric audit: make the machine hold the invariant, not the human

PART IX

What to Take Home

Six Prompt Patterns Worth Stealing

- **Pin the contract** — exact filenames, exact columns, exact value domains
- **Front-load known landmines** — every constraint is a debugging session you do not have on stage
- **Make acceptance external** — a script you did not write this turn decides pass or fail
- **Ask for evidence, not recall** — `unique()`, `pip list`, the output file; never "as I remember"
- **Name the stopping point** — an explicit halt token, stated once in the instructions
- **End on a question** — the last line of a good prompt is often what the agent should *ask*, not do

Four Ways It Goes Wrong

- ⚠ **The agent solves too early**
  - Ask it to describe the problem and forbid the fix in the same breath
- ⚠ **The number is remembered, not read**
  - Every figure re-derived from a file; source tags; a scripted audit
- ⚠ **The reviewer has seen the answer key**
  - Reviewer subagents get the manuscript and the outputs, nothing else
- ⚠ **The citation looks perfect**
  - Search, resolve, verify six fields at Crossref. Unverified means deleted

The agent did not make the analysis trustworthy.
The acceptance scripts did.

Every mission in this talk is reproducible from a public repository:
github.com/htlin222/IMBrave150-mock

References

1. Finn RS, Qin S, Ikeda M, et al. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *N Engl J Med*. 2020;382(20):1894-1905. doi:10.1056/NEJMoa1915745
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786
3. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150-161. doi:10.1002/pst.433
4. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*. 2016;11(5):702-712. doi:10.1177/1745691616658637
5. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav*. 2020;4(11):1208-1214. doi:10.1038/s41562-020-0912-z

References (continued)

6. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):Article 11. doi:10.2202/1557-4679.1043

7. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453-1457. doi:10.1016/S0140-6736(07)61602-X

8. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885

9. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

10. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107. doi:10.1002/sim.3697

11. Crossref REST API documentation. Crossref. Accessed August 25, 2026. https://api.crossref.org

Thank You

Repository, mission guide and acceptance scripts:
github.com/htlin222/IMBrave150-mock

Hsieh-Ting Lin, MD　林協霆

Department of Medical Oncology　腫瘤內科部

Koo Foundation Sun Yat-Sen Cancer Center　和信治癌中心醫院