Running a Study Without Writing Code

Ten hospital exports, one AI agent, and the part that takes longest — working out whether we can believe it.

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 are entirely synthetic**
  - Simulated from the published summaries of the IMbrave150 trial[1]
  - No real patient is represented; no record comes from any chart
  - Nothing here is evidence about atezolizumab, bevacizumab or sorafenib
- ◎ **What this talk is about**
  - How to instruct an agent so that its output can be checked
  - The hazard ratio is a teaching target, not a finding
- ✓✓ **Conflict of interest**
  - No financial or non-financial conflict; no industry funding
  - 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 an Agent Is, and Is Not

- **A chatbot** answers you
  - You paste data in, it writes text back. Nothing on your disk changes.
- **An agent** works on your machine
  - It reads your files, writes scripts, runs them, reads the error, tries again
  - It keeps going until a condition you set is met
- **So the hard part moves**
  - Not *"can it write the code"* — it can
  - But *"how do I know the answer it handed me is right"*
- That question is the whole talk

What You Will Be Looking At

THE SCREEN

- **Top** — what I typed, in plain English
- **Middle** — what it decided to do, and the commands it ran
- **Bottom** — the answer, and how long it took

THE IMPORTANT PART

- It does not answer from memory
- It **writes a file of instructions**, runs it, reads the result
- That file stays on disk
- Anyone can open it and check what was actually done

The Sentence Worth Learning to Write

⚠ WHAT MOST PEOPLE TYPE

> Analyse this data and tell me if the treatment works.

- Picks its own method
- Picks its own file names
- Different answer every run
- Nothing to check it against

✓✓ WHAT I WILL TYPE

> Fit a Cox model on treatment alone. Write `naive_hr.json` with `hr`, `ci_low`, `ci_high`. Then tell me in one sentence what you would conclude.

- Named file, named fields
- Reruns identically
- A script can grade it

Ten Hospitals, Three Dialects

What this step does

- Ten CSV exports, three incompatible record systems, into one table

> Group the ten files by their column fingerprint, not by filename. Report the actual values with `unique()`. Convert to one schema, names and units fixed. Never drop a row for missingness. Then attach each hospital's type and region.

What to watch

- Albumin is in g/L at some sites, g/dL at others — a tenfold error waiting
- **1,800 patients, 962 vs 838.** Three sites record no AFP at all

RECORDED

Ten hospitals, three dialects

Actual values found

Field	A	B	C
arm	Atezo+Bev / Sorafenib	AtezoBev / Sorafenib	A+B / SOR
sex	Female / Male	F / M	male 0/1
etiology	HBV HCV Nonviral	same	**lowercase** hbv hcv nonviral
flags	0/1	**Yes/No**	0/1
Child-Pugh	5, 6	5, 6	**A5, A6**
albumin	g/dL	**g/L** (24–52)	g/dL
bilirubin	mg/dL	**µmol/L** (3.4–54.5)	mg/dL
enroll	2018-01 monthly	monthly	**year only**

Two things worth flagging beyond naming:

–

❯

⚠ Transcript saving is off — inherited CLAUDE_CODE_CHILD_SESSION marker · restart with CLAUDE_CODE_FORCE_S…
Opus 5 high 25% [5]3% 06:10 [W]18% 08/28 04:00
-- INSERT -- ▸▸ auto mode on (shift+tab to cycle)

It built the comparison table itself and found the trap: **dialect B stores albumin in g/L and bilirubin in µmol/L**.

The Unadjusted Answer Is Lying

What this step does

- Compares the two arms with no adjustment whatsoever

> Fit a Cox model on treatment alone.
> Write `naive_hr.json`. Then tell me in one
> sentence what you would conclude from
> this number by itself.

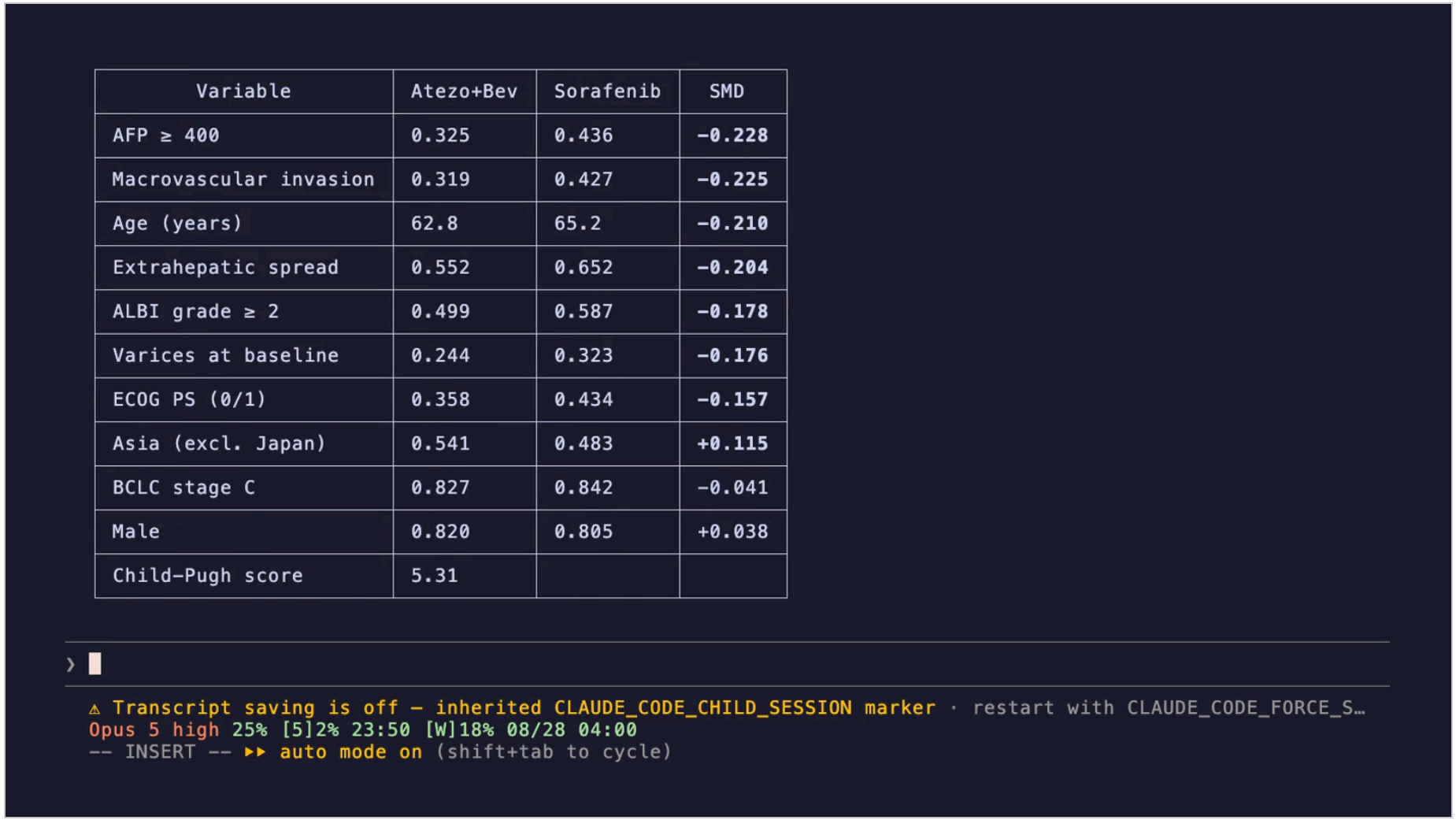
What to watch

- **HR 0.505** — the drug looks better than the trial ever claimed
- Then: **8 of 11 baseline factors differ**, every one favouring the treated arm
- Believe the first number for thirty seconds. That is the point

RECORDED

The unadjusted answer

Variable	Atezo+Bev	Sorafenib	SMD
AFP ≥ 400	0.325	0.436	−0.228
Macrovascular invasion	0.319	0.427	−0.225
Age (years)	62.8	65.2	−0.210
Extrahepatic spread	0.552	0.652	−0.204
ALBI grade ≥ 2	0.499	0.587	−0.178
Varices at baseline	0.244	0.323	−0.176
ECOG PS (0/1)	0.358	0.434	−0.157
Asia (excl. Japan)	0.541	0.483	+0.115
BCLC stage C	0.827	0.842	−0.041
Male	0.820	0.805	+0.038
Child–Pugh score	5.31		



HR 0.505 — then the balance table. Every serious imbalance leans the same way, toward the treated arm.

Make the Two Arms Comparable

⚖ What this step does

- Pairs each treated patient with an untreated patient of similar profile

> Match 1:1 on the propensity score, nearest neighbour, no replacement. Take the treated in **ascending** score order. Do not force a match outside the caliper. Report how many treated were left unmatched.

👁 What to watch

- **706 pairs. 256 treated patients found no match** — that is honesty, not failure
- **HR 0.578.** The randomised trial reported **0.58**[1]

[1] Finn RS, et al. *N Engl J Med*. 2020;382(20):1894-1905. doi:10.1056/NEJMoa1915745

RECORDED

Matching, and what it costs

missingness in any of them, so all 1800 patients enter. Fitted PS range 0.156–0.848; SD of the logit PS = 0.5710, so the **caliper = 0.1142** on the logit scale.

Matching

1:1 greedy nearest neighbour on the logit PS, without replacement, treated processed in ascending PS order, no match allowed beyond the caliper.

	n
Treated (Atezo+Bev)	962
Controls (Sorafenib)	838
Matched pairs	**706**
Treated left unmatched	**256**

At least 124 treated could never match (fewer controls than treated); the other ~132 losses are caliper failures. They are not a random subset — the matched treated have a mean PS of 0.518 vs 0.567 across all treated, i.e. the dropped patients are the high-propensity treated, for whom no comparable Sorafenib patient exists. The matched cohort therefore estimates an effect in a restricted region of overlap, not the full ATT.

›

⚠ Transcript saving is off — inherited CLAUDE_CODE_CHILD_SESSION marker · restart with CLAUDE_CODE_FORCE_S…
Opus 5 high 25% [5]3% 06:10 [W]18% 08/28 04:00
-- INSERT -- ▸▸ auto mode on (shift+tab to cycle)

706 pairs, 256 treated left unmatched. It separates the 124 who could never match from the 132 caliper failures.

Try to Break Your Own Result

🔀 What this step does

- Re-runs the same data through every defensible analysis, not the one that worked

> Hold the data fixed. Vary only the analytic choices: 15 covariate sets × 8 adjustment methods = 120 runs. Report the median, the spread, and **how many landed outside the range I was hoping for.**

👁 What to watch

- The median lands on **0.58** every time. The **spread does not** — about a third of the runs fall outside 0.55–0.61
- Nothing crosses 1.0. The direction is solid; the third decimal is not

· 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

RECORDED 120 analytic paths

The one result you should discard

clin-liverperf under 1:2 matching gives HR 0.822 — the extreme of the whole range, and it is an artifact. Those small sets contain only coarse discrete covariates, so the propensity score takes very few distinct values:

Covariate set	Distinct PS values	Unique controls used	Max reuse of one control
full11	1717	640/838	18
clin-minimal	202	277/838	30
clin-liverperf	29	48/838	99
clin-tumour	24	44/838	212

In clin-tumour, five control patients supply 45% of the entire matched control weight. These are not 24 informative specifications; they are matching on a near-degenerate score. Excluding them, the honest figure is **median 0.578, IQR 0.565–0.598, range 0.528–0.672, 18.8% outside.**

Choices I had to make

- **The full 11** are age, sex, etiology, ECOG PS, Child-Pugh score, ALBI grade, BCLC stage, AFP ≥400,

›

⚠ Transcript saving is off — inherited CLAUDE_CODE_CHILD_SESSION marker · restart with CLAUDE_CODE_FORCE_S…
Opus 5 high 26% [5]3% 23:50 [W]18% 08/28 04:00
-- INSERT -- ▸▸ auto mode on (shift+tab to cycle)

It found its own outlier: the 0.822 run matches on a near-degenerate score that reuses one control **212 times**.

Let It Review Its Own Manuscript

What this step does

- Opens four reviewers at once, each blind to the other three

> Open four reviewers in parallel: statistical, clinical, reporting standards, reproducibility. Each may read the manuscript and the output files **only**. Each must raise at least three major comments. Found none? Read it again.

What to watch

- Four windows moving at once — that picture is the point
- The reproducibility reviewer re-checks **every number against its source file**

RECORDED Four reviewers, blind to each other

Four background agents launched from one message, each reading the manuscript and the output files, none seeing the others.

The Whole Path, on One Slide

WHAT WE JUST DID

- 10 CSVs, 3 record systems
- → 1,800 patients, one table
- → **HR 0.505** unadjusted
- → **HR 0.578** matched
- → 120 re-runs: median **0.58**
- → four reviewers, blind

THE BENCHMARK

- The randomised trial reported **HR 0.58**[1]
- The gap between 0.505 and 0.578 was confounding — and it was recoverable
- Nothing above was typed by hand except the prompts

[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

Four Habits Worth Keeping

- 🎯 **Name the file and the fields**
  - "Save the results" cannot be checked. `naive_hr.json` with `hr`, `ci_low` can
- ⚠ **Write your known traps into the instructions**
  - Every constraint you state is a debugging session you do not have on stage
- ✓✓ **Let something else decide pass or fail**
  - A script you did not write this turn, that the agent cannot edit
- ⏸ **Tell it where to stop**
  - Otherwise it finishes the whole project in one turn — correct, and unreviewable

Thank You

The data, the prompts and the checks:
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