Referee Report — Round 5

Tennis Match Simulator

Referee 2

2026-02-09

Critical Data Leakage Invalidates All Backtest Results

Verdict: Major Revisions Required

- **CRITICAL:** ATP match data uses **tournament start dates**; betting data uses **actual match dates**
- Later-round results leak into Elo ratings when predicting earlier rounds
- The existing leakage check (`validate_elo_betting.R`) is **tautological** — it checks `tourney_date` against `tourney_date`
- All reported metrics (68.6% accuracy, +9.9pp over MC, ROI figures) are unreliable

Note: The Elo implementation itself is correct. The bug is in how historical match dates are defined, not in the model.

Two Data Sources Use Different Date Semantics

Source	Column	Semantics	Brisbane 2024 Example
ATP (Sackmann)	`tourney_date`	Tournament start	All 31 matches: 2024-01-01
Betting (t-d.co.uk)	`Date`	Actual match date	R1: Dec 31, QF: Jan 3, F: Jan 4

In the code:

- `02_player_stats.R:68` — `match_date = ymd(tourney_date)`
- `05_betting_data.R:316` — `match_date = as_date(match_date)`

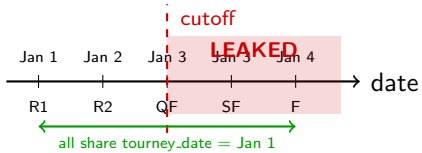
Both are stored as `match_date` in their respective data frames — same name, different semantics.

How Future Results Leak Into Predictions

Predicting a QF on Jan 3:

1. Betting data: `cutoff = Jan 3`
2. ATP filter: `tourney_date < Jan 3`
3. Tournament started Jan 1 $\rightarrow$ **passes filter**
4. **SF (Jan 3) and Final (Jan 4) included in Elo**

The model knows who won the Final before predicting the Quarterfinal.



Existing Leakage Check Tests the Wrong Thing

`validate_elo_betting.R:46--62`:

```
prior_matches <- historical_matches %>% filter(match_date < test_date)
```

match_date is **tourney_date**

```
in_history <- elo_history %>% filter(match_date >= test_date) %>% nrow()
```

also **tourney_date**

Result: **Always passes.** Compares tourney_date to tourney_date.
Does NOT verify actual future matches are excluded.

Leakage Disproportionately Inflates Elo Accuracy

Why Elo is more affected than MC:

- Each leaked match changes Elo by up to ± 32 points (K-factor)
- Effect compounds across multiple leaked rounds
- Players who advance get rating boosts from future wins
- Those players *did win* their earlier matches $\rightarrow$ spurious accuracy

Why MC is less affected:

- Serve/return stats aggregated over 9 years (2015+)
- A few extra matches barely change averages
- No per-match compounding effect
- Leakage has minimal impact on MC predictions

Tournament Type	Matches	Max Gap	R1+R2 (% of draw)
Grand Slam	$\sim$127	14 days	75%
Masters 1000	$\sim$55–95	7–9 days	75%
ATP 500 / 250	$\sim$31	6 days	75%

All Reported Metrics Are Based on Leaked Data

Metric	Reported Value	Status
Elo Accuracy	68.6%	Unreliable — leakage inflated
MC Accuracy	58.7%	Less affected but same bug
Elo – MC Gap	+9.9 pp	Unreliable — differential leakage
Brier Score (Elo)	0.2029	Unreliable
Log Loss (Elo)	0.5913	Unreliable
All ROI / Edge figures	Various	Unreliable

Conservative estimate: Leakage inflates Elo accuracy by 1.3–2.2 percentage points. The true Elo–MC gap may be 8–9 pp instead of 9.9 pp — or could be smaller if the Elo model's calibration is also affected.

True performance can only be determined after fixing the date alignment.

Elo Implementation Itself Is Correct (Modulo Dates)

- ✓ Standard Elo formula
- ✓ Per-player K-factors (fixed R3)
- ✓ Surface-specific blending
- ✓ Alphabetical ordering (no calibration bias)
- ✓ 14 unit tests passing
- ✓ History tracking (fixed R4)

- ? K=32 not validated for tennis
- ? Scale factor 400 not validated
- ? No calibration analysis published
- ? No decay for inactive players
- ? Surface Elo starts at 1500

The Elo module is well-structured, well-tested code. The issue is upstream: the date field fed into it is semantically wrong.

Edge Analysis: Methodology Sound, Inputs Compromised

What is correct:

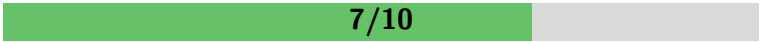
- Edge = model_prob − implied_prob (includes vig — correct for betting decisions)
- Fractional Kelly criterion with 5% max bet cap
- Bootstrap CI for ROI (1,000 resamples)
- Four baseline strategies compared

What is problematic:

- All edge calculations use leaked model probabilities
- No out-of-sample test completed (H2 2024 data unavailable)
- Closing vs. opening odds not formally verified from source documentation
- `validate_elo_betting.R` results not reported in any correspondence round

Bottom line: The edge analysis framework is well-designed. But the inputs (model probabilities) are contaminated by leakage, making all ROI figures uninterpretable.

Replication Readiness: 7/10 (decreased from 8/10)



- ✓ Folder structure
- ✓ Relative paths
- ✓ Variable naming
- ✓ Script naming / ordering
- ✓ Dependencies (`renv.lock`)
- ✓ Random seeds set
- ✓ Master script exists

- × Date alignment broken (results unreliable)
- × Comparison scripts not in master pipeline
- × In-text statistics manually entered

Decreased from 8/10 because biased results are not meaningfully replicable.

Recommendations (Priority Order)

1. **Fix date alignment** — Replace `tourney_date` with actual match dates
   - Preferred: Join ATP matches with betting data by player names + tournament
   - Alternative: Infer from `round` column (R1 = day 1, R2 = day 2, etc.)
   - Stopgap: Use 14-day buffer (`cutoff - 14d`) to bound leakage
2. **Fix leakage validation** — Test against actual match dates, not tournament dates
3. **Re-run all backtests** with corrected dates; update CLAUDE.md
4. **Quantify leakage impact** — Compare accuracy before/after fix
5. Report Elo calibration (predicted vs. actual win rates by bin)
6. Add McNemar test for model comparison significance
7. K-factor sensitivity analysis (K = 16, 20, 24, 32, 40)