KBO / NPB / CPBL → MLB Pitcher Translator (Shiny App)

Last updated: October 27, 2026 · Live at `leogentile.shinyapps.io/pitcher-translator`

What this is

An interactive R Shiny app that answers one question: when a pitcher moves from South Korea's KBO, Japan's NPB or Taiwan's CPBL to MLB, what should we expect from him?

- It lets you browse every KBO, NPB, CPBL and MLB pitcher since 1995, with each person's seasons in every league on one page.
- It projects a pitcher's first MLB season with seven models and honest uncertainty ranges (CPBL pitchers through the two-way league model; see section 8).
- It estimates a KBO/NPB pitcher's odds of reaching MLB and of sticking there.
- It finds his closest MLB statistical comparisons and shows how real pitchers who made the same move actually did.
- An Article tab explains all of it in plain language for non-technical readers, with a technical appendix.

Like every Shiny app, it has two pieces. The UI is what you see and click. The server is the logic that reacts to your clicks and draws the output. The line `shinyApp(ui, server)` joins the two and runs the app.

How to load the app

First time on a computer: install the packages once. In the RStudio Console, run:

```
install.packages(c("shiny", "DT", "ggplot2", "ggrepel", "survival",
"survminer", "lme4", "plotly", "rstanarm", "ranger", "gbm"))
```

The last three are optional. Without them, the app still runs but hides the Bayesian and machine-learning projections.

Every time:

1. Open RStudio.
2. File → Open Project → choose `pitcher-translator.Rproj` in the pitcher-translator folder. This sets the working folder so the app can find `data/`, `logos/` and `reports/`.
3. Open `app.R` from the Files pane, then click **Run App** at the top of the editor. Typing `shiny::runApp()` in the Console does the same.

4. The app opens in a window or browser tab after 10–20 seconds while it loads the data and models.
5. To stop it, press the red stop sign in the Console, or press Esc in the Console.

You don't need to run any other file first. All the data and fitted models are already saved in `data/`, and the app loads them at startup; nothing is refit when you click. Run the scripts in `scripts/` only when adding a new season (see section 14).

Deploying to the web: with the project open, run `rsconnect::deployApp(appName = "pitcher-translator", account = "leogentile")` in the Console and type Y if asked. If R says a namespace "is already loaded", use Session → Restart R and run it again. `.rscignore` keeps backups, scripts and raw pulls out of the upload.

What changed since the last version of this document

- **Taiwan's CPBL (new).** Every CPBL pitcher since the league began in 1990 (1,244 pitchers, 3,308 seasons) is in the app: in Explorer, Profile and Quadrants with MLB comps, and in the Scouting Report and Next Up with an MLB projection from the two-way league model. All 10 CPBL franchises have logos, including era logos for renamed teams (e.g. EDA Rhinos, Lamigo Monkeys).
- **Cross-league linking by stat lines (new).** Each CPBL pitcher is linked to MLB, KBO and NPB by matching his actual season lines (batters faced per season) from his Baseball-Reference register page, in both directions. 448 pitchers are linked this way. The same check found and merged two people who had been split in two (Marty Janzen, MLB ↔ KBO; Cristopher Mercedes, KBO ↔ NPB).
- **Two-way league model expanded.** It now learns from every pitcher who pitched in two or more of MLB, KBO, NPB and CPBL: 7,040 seasons by 1,103 pitchers (was 4,955 by 726).
- **Tabs reorganized.** Nine tabs became a Scouting Report, Next Up, an Explore menu (Explorer, Profile, Quadrants), a Research menu (Translations, Survival, Accuracy) and an Article. Projection, Comp Finder and the career odds now live together on one Scouting Report page, with a CSV download.
- **Article tab (new).** A plain-language write-up for front-office and general readers ("Will He Translate?"), with live charts and a technical appendix whose numbers are read from the fitted models at startup.
- **US minor-league file expanded.** It now covers every affiliated level (AAA to Rookie): 557 lines for 334 departures. The career-odds models were refit on it.
- **Project reorganized.** Data in `data/` (raw pulls in `data/raw/`), five rebuild scripts in `scripts/`, the two R Markdown reports in `reports/`, logos in `logos/`. `README.md` at the top of the folder is the full reference.

The tabs

1. **Scouting Report:** a one-page report for any KBO, NPB or CPBL pitcher-season. Settings: projected role, batters faced, headline model, and comp options. It shows:
   - a header with the headline projection (FIP, K%, BB%, HR%) and, for KBO/NPB, career odds;
   - every model's projection in a chart and table, with the Bayesian 80% range and each model's test error;
   - his closest MLB comps, a fingerprint radar against the closest one, and a zoomed comp map;
   - the closest past movers (real KBO/NPB → MLB moves) and how they went;
   - his full career in every league, and a CSV download of the projection.

   For CPBL seasons only the two-way league model is shown, with a note explaining why.
2. **Next Up:** ranks every qualifying pitcher in a chosen KBO, NPB or CPBL season by projected MLB FIP. Columns add the Bayesian 80% FIP range, the machine-learning FIP, the two-way FIP, "Reach MLB %" and "Stick 2+ %", and each pitcher's closest real precedent (same role). CPBL seasons are ranked by the two-way model. Click a pitcher to open his Scouting Report.
3. **Explore → Explorer:** browse one league's pitching (KBO, NPB or CPBL) by season range and minimum innings: an ERA-vs-stat scatterplot, a league trend chart and a sortable leaderboard with team logos.
4. **Explore → Profile:** a Baseball-Reference/Savant-style page for any pitcher across all four leagues: season table with logos and career rows, percentile bars, an interactive career chart, and the closest MLB seasons to a chosen season. If he made a KBO/NPB → MLB move, the header shows his before/after line.
5. **Explore → Quadrants:** a K% vs. BB% scouting plot for one league-season (any of the four leagues), with shaded "elite" zones and points colored by ERA.
6. **Research → Translations:** each real KBO/NPB → MLB move's final origin season vs. first MLB season for a chosen stat. Hover a point to see the pitcher.
7. **Research → Survival:** how long movers keep pitching in MLB. Kaplan-Meier curves split by origin league or earlier MLB experience, with a log-rank test and a robust p-value.
8. **Research → Accuracy:** every model's average error on pitchers it never saw (cross-validation or future test; MLB year 1 or years 1–2), with the luck floor, predicted-vs-actual plots and an interval check.
9. **Article:** "Will He Translate?", the plain-language write-up, with a technical appendix.

Every tab opens with an "About this tab" summary. The layout works on phones (390 px) without sideways scrolling.

Data sources

League	Seasons	Pitcher-seasons	Source	Checked against
KBO	1982–2025	7,662	nk_datasets (through 2024) + koreabaseball.com (2025)	2025: every pitcher matches the official site. 1995–2025: the list of pitchers each season matches exactly.
NPB	1936–2025	18,284	nk_datasets + npb.jp official team pages (2008–2025)	2008–2025: every season's totals match npb.jp exactly.
CPBL **(new)**	1990–2025	3,308 (1,244 pitchers)	Baseball-Reference register: every CPBL team's pitching table, pulled through a browser	2025 spot-checked against cpbl.com.tw; league wins = losses every season.
MLB	1871–2025	49,094	Lahman Baseball Database + MLB Stats API (2025)	2025: matches Lahman's 2025 release exactly (873 pitchers).
US minors (AAA to Rookie)	1996–2025	557 lines for 334 departures	MLB Stats API	Only pitchers who left KBO/NPB; matched on name and timing (used for "reaches MLB").

nk_datasets is an MIT-licensed Python package by Myungkeun Park (github.com/frankmkpark/mnk_datasets). It scrapes and standardizes data from the official league sites and the Lahman database.

Rate stats (K%, BB%, HR%) are calculated by us from raw counts (strikeouts, walks, home runs, batters faced). They are not taken pre-computed from any source. A pitcher traded mid-season has one combined row; his team is the one he threw the most innings for.

Data quality: what we found and fixed

The original problem. Yoshinobu Yamamoto's numbers looked too ordinary for a pitcher who was a top MLB starter from day one. His 2023 NPB season was missing from the data: 1.21 ERA, 169 K in 164 IP, a no-hitter.

- NPB 2023 had only 261 pitchers, versus 352 in 2022.
- The four 2024 NPB arrivals all showed a suspicious two-year gap: Yamamoto, Shota Imanaga, Yuki Matsui and Yariel Rodriguez.

- The likely cause: the scraper only saw pitchers still on a team's current roster page, so players who had just left for MLB were dropped.

The fix.

- NPB: 2023 and 2024 were rebuilt from npb.jp's official team pages, and 2025 was added. Every NPB season from 2008 to 2025 now matches npb.jp's official totals exactly. Rodriguez's last NPB season is 2022: he did pitch in 2023, but not in NPB, so his gap is real.
- KBO: 2025 was re-pulled in full from the official site and checked pitcher by pitcher. The site's English pages label outs as "IP", so the update script checks that 27 × ER ÷ outs reproduces each pitcher's ERA.
- MLB: 2025 was added from MLB's own Stats API, mapped to Lahman's IDs, and matches Lahman's 2025 release exactly.
- CPBL: B-Ref writes Taiwanese given names as two words; they are hyphenated the MLB way ("Chun-Hsiang Chu"). Names B-Ref gets wrong are fixed by hand in `update_seasons.R`.

Linking one person across leagues. Every pitcher has a `person_id` in `player_id_crosswalk.csv`, which holds 12,714 people. Pitchers are never linked by name alone when an ID or stat line is available.

- NPB ↔ MLB links come from the Chadwick Bureau player register, which stores many players' NPB IDs next to their MLB IDs.
- CPBL ↔ everyone (new): each CPBL pitcher's Baseball-Reference register page lists his seasons in every league. Those seasons are matched to our files by batters faced each season (all seasons agree, or at least half of those with 100+ batters faced). If the lines show two crosswalk rows are one person, they are merged. Only one CPBL link rests on a name alone (Chien-Cheng Kuo, NPB 1990–91, which B-Ref doesn't list).
- Remaining KBO and some NPB links use spelling-tolerant name matching ("Ohno"/"Ono", "Wei-Yin"/"Weiyin"), requiring a unique name in both leagues and careers that line up in time.
- Result: 546 NPB+MLB people, 278 KBO+MLB, 105 KBO+NPB, 75 in all three. CPBL: 350 linked to MLB, 143 to NPB, 94 to KBO; 795 CPBL-only.
- Each link records how it was made, so name-only links can be reviewed. This matters because names collide: two different KBO pitchers named Tae-Hun Kim both pitched in 2025.

The audit. Every file is checked for one row per pitcher per season; innings, ERA and rates consistent with the raw counts; league wins equal losses; one person per ID; transitions consistent with the season files; and cross-validation groups never splitting a pitcher.

Derived datasets (built by our own scripts)

- `transition_pairs_pitching.csv`: one row per move, 256 moves (190 from NPB, 66 from KBO), comparing the final KBO/NPB season with the first MLB season after it.
  - 75 are first-time MLB arrivals; 181 had pitched in MLB before (e.g. Miles Mikolas, Erick Fedde). `prior_mlb_seasons` tells them apart.
  - A move is an MLB season whose most recent earlier season was in KBO or NPB, at most 3 seasons before. CPBL → MLB moves are not included (too few; see section 8).
  - `origin3_*` columns hold the 3-season blend of K%, BB% and HR% (section 2).
- `transition_full_career.csv`: every MLB season in the stint after each move (728 rows), with a starter/reliever role (innings per appearance ≥ 3 = starter). A stint ends when the pitcher goes back to Asia or misses two straight MLB seasons.
- `survival_pitchers.csv`: one row per move, with how many MLB seasons the stint lasted. 72 stints ended with a return to Asia; 57 are still going (censored).
- `us_minors_after_asia.csv`: affiliated US minor-league lines for pitchers in the 1–3 seasons after they left KBO/NPB.
- `league_env.csv`: league averages for every league-season in all four leagues, plus each MLB season's FIP constant (2025: 3.27).
- Saved models, fit once and loaded by the app: `component_models.rds` (mixed models), `bayes_models.rds`, `ml_models.rds`, `two_way_models.rds`, `career_odds.rds`, `model_comparison.rds` (all test results).

Limitations to keep in mind

- **CPBL projections** come from one model (two-way), with no Bayesian range or career odds, and can't be scored directly because only 24 pitchers have gone straight from the CPBL to MLB since 1995. Treat them as an informed starting point with more uncertainty than KBO/NPB projections.
- **Name-based links:** most KBO cross-league links, some NPB ones, and the US minor-league matches are name-based. They are spot-checked but not guaranteed. CPBL names haven't been fully checked against cpbl.com.tw.
- **Older data:** NPB before 2008 can't be checked against npb.jp and has no team information. KBO stat lines before 2025 come from nk_datasets.
- **"Survival" is a proxy:** it means "still appearing in MLB pitching stats," not roster transactions.
- **Season cutoff:** the app starts at 1995 (`EARLIEST_SEASON`), the start of the modern wave with Hideo Nomo.
- **Small sample:** only about 130 pitchers have enough MLB innings to fit the translation models. That is why the intervals matter.

- **No batted-ball data:** xFIP and SIERA would be steadier targets than FIP, but they need fly-ball / ground-ball counts that aren't available for Asian leagues here.

Still open

- A slow name check of every CPBL pitcher against cpbl.com.tw.
- Adding the 2026 seasons once they're final.
- Survival by profile (Cox regression on K-heavy vs. command-heavy pitchers), and a head-to-head two-pitcher view.

How the statistics work

1. What we predict, and why FIP

Every projection model predicts three rates for a pitcher's first MLB season after a move:

- K% = strikeouts ÷ batters faced
- BB% = walks ÷ batters faced
- HR% = home runs ÷ batters faced

These are the parts of pitching a pitcher controls most, and they stabilize fastest. They are combined into FIP (Fielding Independent Pitching):

$$FIP = (13 \cdot HR + 3 \cdot BB - 2 \cdot K) \div IP + C$$

- C is that MLB season's constant, set so league-average FIP equals league-average ERA: C = league ERA − (13·lgHR + 3·lgBB − 2·lgK) ÷ lgIP. From our own MLB totals: 3.40 in 2023, 3.31 in 2024, 3.27 in 2025.
- Since the models predict rates per batter, innings are converted with the league's batters faced per inning (4.25 in 2025): FIP = (13·HR% + 3·BB% − 2·K%) × (BF per IP) + C.

We use FIP instead of ERA or WAR because ERA depends on defense, sequencing and luck, and WAR also depends on playing time, park and role.

2. The 3-season origin blend

A single season can be a fluke, especially for relievers with 50 innings. The fitted models use a blend of the last three seasons in the same league, weighted toward the most recent (like the Marcel projection system):

$$\text{blended K\%} = (5 \cdot SO_{Y} + 4 \cdot SO_{Y-1} + 3 \cdot SO_{Y-2}) \div (5 \cdot BF_{Y} + 4 \cdot BF_{Y-1} + 3 \cdot BF_{Y-2})$$

The same formula applies to BB% and HR%. On the same cross-validation folds, the blend cut the average FIP miss:

Model	**Final season only**	**3-season blend**
Boosted trees	0.689	0.643
Random forest	0.701	0.662
Mixed model	0.714	0.713

(Side-by-side test from the October 7 version; the current scores are in section 12.)

3. League-season adjustment (comp space)

A 3.00 ERA means something different in a high-offense KBO season than in a low-offense NPB season. For the comparison space, each stat (ERA, K%, BB%, HR%) is turned into a z-score within its own league and season:

z = (pitcher's stat − league-season mean) ÷ league-season standard deviation

- The mean and standard deviation come from pitchers with 30+ IP, the same pitchers that make up the comp pool.
- League-seasons with fewer than 8 such pitchers fall back to that league's all-years mean and standard deviation.
- CPBL seasons get z-scores the same way, against CPBL league-season averages.

4. PCA + nearest neighbors (comps, precedents)

- The four z-scores for every KBO/NPB/MLB pitcher-season with 30+ IP are rotated into four principal components with `prcomp()`. CPBL seasons are projected into that same space with `predict()`, so adding the CPBL doesn't move anyone else's comps; CPBL percentiles use their own league's distribution.
- Distance is straight-line (Euclidean) distance across all four components. The map draws only the first two (about 72% of the information), so a comp can look a little farther away on the map than it is.
- Matching is within role: starters with starters, relievers with relievers. Yamamoto 2023's closest precedent is Masahiro Tanaka 2013.
- The Scouting Report and Profile return the closest MLB pitcher-seasons, excluding his own. Next Up and the Scouting Report also find the closest origin season of a real past move and report how that pitcher's MLB stint went.

5. Model A: negative-binomial mixed model (GLMM)

Why this model. Strikeouts, walks and home runs are counts out of a number of chances (batters faced), so the right tool is a count model with an offset. We use the negative binomial rather than the Poisson because the data are overdispersed (residual deviance ÷ df ≈ 5): Var = $\mu + \mu^2/\theta$.

The model, fit once each for strikeouts, walks and home runs:

$$\log E[\text{count}_{ij}] = \log(\text{BFP}_{ij}) + \beta_0 + \beta_1 \cdot \text{origin K\%} + \beta_2 \cdot \text{origin BB\%} + \beta_3 \cdot \text{origin HR\%} + \beta_4 \cdot \text{NPB} + \beta_5 \cdot \text{years since MLB debut} + \beta_6 \cdot \text{starter} + \beta_7 \cdot \text{prior MLB} + u_i$$

- Origin K%, BB% and HR% are the 3-season blends; log(BFP) is the offset, turning a count model into a rate model.
- $u_i \sim \text{Normal}(0, \sigma^2)$ is a random intercept per pitcher.

- Role (starter vs. reliever) and prior MLB time are covariates.
- Fitting data: every MLB season with 30+ IP after a move whose origin season had 20+ IP: 394 MLB seasons from 130 pitchers, fit with `lme4::glmer.nb`.

Projection and its interval (delta method). For a new pitcher we use the fixed effects only (u = 0). $\eta = x'\beta + \log(BFP)$; $Var(\eta) = x'Vx$, with V the covariance matrix of β; the 95% interval is $\exp(\eta \pm 1.96 \cdot SE)$. This interval describes the average pitcher with this profile, not one pitcher's season, which is why the Bayesian model exists.

6. Model B: Bayesian mixed model (prediction intervals)

Same formula as Model A, fit with `rstanarm::stan_glmer.nb` (4 chains × 2,000 iterations, default weakly informative priors), giving 4,000 posterior draws. Each pitcher's random intercept is pulled toward the group average ("borrowing strength"), which is what lets a 130-pitcher dataset produce stable estimates.

For each of the 4,000 draws:

1. Take that draw's β, σ and θ.
2. Draw a new pitcher effect $u \sim Normal(0, \sigma^2)$ (pitcher-to-pitcher differences).
3. Compute $\mu = BFP \cdot \exp(x'\beta + u)$.
4. Draw a season's count from the negative binomial with mean μ (one-season luck).
5. Divide by BFP to get a rate.

The middle 80% and 95% of the 4,000 rates are the prediction intervals; FIP is computed draw by draw, giving FIP intervals and P(FIP < 4.00). K, BB and HR come from three separate models, so their draws are treated as independent, which makes the FIP interval slightly wider than a joint model's.

7. Models C and D: machine learning

- Features (16): origin K%, BB%, HR% (3-season blends), final-season ERA, IP and batters faced; prior MLB seasons; years between seasons; the origin league's averages that season; MLB's averages the season before the move; NPB and starter indicators.
- Training data: 156 moves with 20+ origin IP and 50+ MLB year-1 batters faced, each weighted by MLB batters faced.
- Random forest (`ranger`): 1,000 trees, minimum 5 moves per leaf.
- Boosted trees (`gbm`): 400 depth-2 trees, learning rate 0.01, 70% subsampling.
- In the app, a new projection assumes the pitcher arrives the next season into the latest MLB environment (2025).

8. Model E: two-way league model (expanded; the CPBL model)

The other models learn only from MLB seasons after KBO/NPB → MLB moves. But many more pitchers have pitched in more than one league, in every direction: Americans who went from MLB to Asia, Taiwanese pitchers who went to Japan, Korean and Japanese veterans who finished in Taiwan. Every one of those seasons says something about the gap between leagues.

The data: every season (30+ batters faced, 1960 on) of every pitcher who has pitched in two or more of MLB, KBO, NPB and CPBL: **7,040 seasons by 1,103 pitchers.**

The model, a binomial mixed model fit once each for K, BB and HR:

$\text{logit}(\text{rate}_{ij}) = \text{logit}(\text{league-season average}_j) + \beta_{\text{league}} + \beta_{\text{league}\times\text{era}} + \beta_{\text{starter}} + u_i$

- Every season is measured against its own league-season average (the offset), so β_{league} is a pure league gap: how the same pitcher's rate compares with his league in KBO, NPB or CPBL vs. in MLB. It can drift over time (the era term).
- $u_i \sim \text{Normal}(0, \sigma^2)$ is the pitcher's own talent, the same in every league. Fit with `lme4::glmer` (binomial).

The learned gaps (odds relative to league average vs. MLB, 2010 era; above 1 = the league flatters that stat):

	K%	**BB%**	**HR%**
Same pitcher in KBO vs. MLB	×1.37	×0.87	×0.74
Same pitcher in NPB vs. MLB	×1.15	×0.97	×0.81
Same pitcher in CPBL vs. MLB	×1.47	×0.84	×0.73

In words: CPBL numbers flatter a pitcher a little more than KBO numbers and clearly more than NPB numbers.

How a projection is made: estimate the pitcher's talent u from his seasons in the 5 years up to the move (a short record is pulled toward average), then translate: MLB rate = $\text{logit}^{-1}(\text{logit}(\text{MLB average}) + \beta_{\text{MLB era}} + \beta_{\text{starter}} + u)$. For a CPBL pitcher this is the whole projection.

Why it's the only CPBL model: only 24 pitchers have gone straight from the CPBL to MLB since 1995, too few to fit or test the other models. The two-way model is scored on the KBO/NPB moves (cross-validation trains it on every cross-league pitcher except the ones being tested): its FIP error is 0.71, tied with the mixed model, and its K% error is the lowest of any model.

9. Baselines

- No change: MLB rate = final-season origin rate.
- League ratio: MLB rate = origin rate × (MLB league average ÷ origin league average), the classic "translation factor."

10. Career odds: will he reach MLB, and will he stick?

The projection models only learn from pitchers who actually pitched in MLB. They answer "how will he pitch if he gets a job," not "will he get a job." Two logistic regressions cover that half (KBO/NPB only).

Reach. Of pitchers who left KBO/NPB for US pro ball, who pitched in MLB within 3 seasons? A departure is a 20+ inning KBO/NPB season (1995–2022) with no KBO/NPB season the next year; "went to US pro ball" means an MLB game or an affiliated minor-league game (AAA to Rookie) within 3 seasons. 315 pitchers went; 181 (57%) reached MLB.

Stick. Of moves that reached MLB (debut through 2023), which stints lasted 2+ seasons? 188 moves; 122 (65%) did.

Inputs, all known at the end of the final KBO/NPB season: K%, BB% and HR% minus the league-season average; innings (log); starter; NPB or KBO; earlier MLB experience; number of KBO/NPB seasons (log); era. Tested with grouped cross-validation, 5 folds × 20 repeats.

Question	Pitchers	Typical rate	Test AUC	Brier (model)	Brier (guess typical rate)
Reaches MLB	315	57%	0.65	0.230	0.246
Sticks 2+ seasons	188	65%	0.65	0.222	0.231

- AUC is the chance the model ranks a random success above a random failure (0.5 = coin flip, 1 = perfect). The Brier score is the average squared miss of the probabilities (lower = better).
- Both beat guessing the typical rate, but modestly. Strikeouts relative to league are the clearest signal in both.
- Shrinkage: held-out calibration slopes were 0.74 (reach) and 0.60 (stick), so the coefficients are multiplied by those factors and the intercept is re-centered on the typical rate. The app presents the odds as nudges from the typical rate, not verdicts.

11. How the models were tested

A model is only ever scored on moves it did not see while being fit.

- **Grouped 5-fold cross-validation, repeated twice.** Pitchers (not seasons) are split into 5 groups; each group is predicted by models fit on the other 4, so a model never "knows" the pitcher it is scored on.
- **Future test.** Fit only on MLB seasons before 2023, then score the 22 moves from 2023–2025.
- **Who is scored:** moves with 20+ origin IP and 100+ MLB year-1 batters faced (130 moves, scored twice in cross-validation).
- **Error measure:** mean absolute error, weighted by MLB batters faced.
- **Two targets:** MLB year 1, and years 1–2 combined (about 1.7× the batters faced, so less luck in the target).
- **Luck floor:** treat each scored pitcher's actual rates as his true talent, replay his season 2,000 times with binomial luck at the same batters faced, and score the replays like predictions. No model can beat this on average.
- **Interval check:** the share of real outcomes inside the Bayesian 80% interval. Honest intervals catch about 80%.

12. Results

Cross-validation, scored on MLB year 1 (lower is better):

Model	FIP error	K% error (pts)	BB% error (pts)	HR% error (pts)
Boosted trees	0.61	3.44	1.82	0.83
Random forest	0.63	3.50	1.90	0.90
Two-way league model	0.71	3.43	1.95	0.95
Bayesian mixed model	0.71	3.77	1.88	0.99
Mixed model (GLMM)	0.71	3.67	1.89	0.96
League ratio	1.25	6.17	2.31	1.25
No change	1.27	4.92	2.44	1.46
Luck floor (a perfect model)	0.45	1.66	1.16	0.71

Cross-validation, years 1–2 combined: boosted trees 0.51, random forest 0.53, Bayesian 0.55, mixed model 0.56, two-way 0.56, baselines 1.18–1.20; luck floor 0.33.

Future test (22 moves, 2023–2025): FIP errors 0.69–0.74 for the fitted models vs. 1.45–1.48 for the baselines, luck floor 0.47. Scored on years 1–2, the two-way model is best (0.55). With only 22 moves, the ranking can shift by chance.

Bayesian 80% interval coverage: K% 89%, BB% 83%, HR% 80%, close to the 80% target.

What this means:

- **Fitted vs. simple:** every fitted model is about twice as accurate as the simple rules (≈0.6–0.7 vs. ≈1.25 runs of FIP).
- **The tree models lead** by about 0.1 runs, helped most by the 3-season blend.
- **Luck is most of the remaining error.** The best model is about 0.16 runs worse than perfect, not 0.61.
- **The two-way model** ties the mixed models on FIP and has the lowest K% error, which supports using it for the CPBL.
- **The Bayesian model's value is its intervals,** not its point predictions.
- **HR is the weakest stat in every model.** Trust the K and BB projections more.

13. Survival analysis

- Kaplan-Meier curves (`survival` package) on `Surv(n_mlb_seasons, event_observed)`. Time = MLB seasons in the stint; event = the stint ending (a return to Asia, or two straight seasons without MLB). Stints with an MLB season in 2024 or 2025 are censored.
- Groups (origin league, or MLB experience before the move) are compared with a log-rank test, plus a robust p-value from a Cox model clustered by pitcher, since some pitchers moved more than once.
- About 67% of stints reach a second MLB season, 47% a third, and 24% a sixth.

14. Rebuilding for a new season

Open `pitcher-translator.Rproj` first, then run, in this order (each script's header says what it reads and writes):

1. `scripts/update_seasons.R`: set `SEASON` and `LEAGUES` at the top. Adds a finished KBO / NPB / MLB season from the official sites (each league file gets a `.bak` backup first). `"CPBL"` rebuilds the CPBL file from `data/raw/cpbl_raw_bref.csv`; a new CPBL season means adding its B-Ref team pages to that file first. (1–5 min)
2. `scripts/build_crosswalk.R`: relinks every person across all four leagues (needs internet for the Chadwick register). If it prints a "check:" line, add that link to `MANUAL` at the top. (~1 min)
3. `scripts/build_transitions.R`: rebuilds the moves, careers and survival files, including the 3-season blends. (seconds)
4. `scripts/fit_models.R`: part 1 refits and tests the seven translation models; part 2 the career-odds models. Run `FIT <- "career_odds"` first to run only part 2. (15–60 min)
5. Knit `reports/component_model_v2.Rmd` (refits the app's mixed models), then `reports/model_comparison.Rmd`.

Then Run App to check, and deploy. `README.md` has the full file-by-file reference.