أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

AI-Driven Workflow Optimization and Prediction in CMS Glance

Final Report · 28 June – 23 August 2026

Ammar Bunajmah, **Muhammad Umair Aslam**

ammarxd1390@gmail.com, muhammad.umair.aslam@cern.ch

Abstract

Every CMS physics analysis clears a linear chain of seventeen review stages before publication, and a delay in that chain is normally noticed only once it has already happened. This report measures where analyses wait, tests what can be predicted about that wait, and states plainly what the record cannot support. Across 3,147 analyses and 30,080 executed transitions from 2012 to July 2026, mean time at stage 212 Preapproved is 284 days against 85 at the next slowest stage, and that one stage carries roughly 53 per cent of all prediction error. Eleven predictors spanning five model families were scored on the same visits and corrected for whether each estimates a typical or an average duration; none beats a per-stage constant, and between 96 and 100 per cent of every model's error is systematic rather than random. Two capabilities survive: a corrected occupancy map, and an ordering of which analyses advance soonest. The binding constraint is observation rather than modelling, and three data requests follow from it.

Date: 23 August 2026

1 Introduction

CMS administers the review of its physics analyses through a linear chain of seventeen stages, tracked in a platform called Glance. Nothing becomes a published paper until it has cleared every one of them. With more than three thousand analyses on record and seventeen steps each, no coordinator can read the list and say which analyses are quietly falling behind, so a stalled analysis surfaces when somebody thinks to ask about it.

This project set out to answer three questions in order. Where do analyses actually wait? How much of that wait can be predicted in advance? And where prediction fails, what is the reason: the models, or the record they are fitted to?

The answers are, respectively, that the waiting is concentrated almost entirely in one stage; that individual waits cannot currently be predicted beyond a per-stage constant; and that the limit is observational. That third answer is the substantive result of the work, and it is reached by four independent routes rather than asserted.

The contributions are as follows.

- A forensic reconstruction of nineteen years of workflow history from three tables whose relationship to one another is not what it appears (Section 2.1), including a correction that takes one stage's apparent queue from 734 analyses to the 27 actually present.
- A like-for-like comparison of eleven predictors across five model families, controlled for both the population each can score and the quantity each estimates (Section 3.4). None beats a per-stage constant.
- Five independent tests of whether the chain's graph structure carries information (Section 3.2), all null or negative, with a diagnosis of why.
- A descriptive account of the Preapproval slowdown as a scheduling signature rather than a queueing one (Sections 3.6 and 3.7), and the data request that would test it.
- A statement of what is deployable today and what is not (Section 4.3), with an evaluation protocol that rejects four otherwise clean wins (Section 4.2).

Throughout, a result is reported only if it survives a rolling-origin backtest at four cut-off dates. That rule is stricter than the one this work started with, and Section 4.2 explains why it was adopted mid-project.

2 Methodology

2.1 The record and what it actually contains

The workflow history is held in three tables, and the relationship between them decides what any measurement can mean. Table 1 states it.

A side channel sits outside all three and shapes every occupancy figure. The stage recorded against an analysis disagrees with the position reconstructed from its movement history for 51.5 per cent of analyses. Certain actions, notably direct jumps to a terminal stage and kick-back reversions to an earlier one, update the analysis record without writing any transition or event, and carry no usable timestamp. The observed chain is therefore genuinely clean, with zero backward or skipping moves in the log, while 385 kick-back reversions and 1,237 off-chain resolutions exist only in the side channel. Occupancy computed from the log alone overstates live queues badly: stage 206 appeared to hold 734 analyses where the live snapshot holds 27.

The correction is to cap each silently resolved or reverted analysis at its last observed event. This reproduces the live snapshot essentially exactly, and two independent implementations agree

Table 1 The three sources and what each one records.

Source	What it records
The itinerary table	**Not a move log.** Sixteen edge rows are written for every analysis at creation, all carrying the creation timestamp, with zero exceptions. A row existing does not mean the move happened.
The event log	The real record of movement. A transition occurred if and only if it carries events, and its time is the first event of that transition. The extract holds 30,080 executed transitions over 3,147 analyses.
The tracker table	One row per analysis carrying its current stage. Its modification timestamp is unreliable in both directions, with a median 47 days before the analysis's last observed event, and is never used as an action time.

on all 12,886 stage-weeks. It is used for the occupancy map. Tested as a prediction feature it made the predictor worse, 14,097 to 14,481 against 12,393 without it, so it is deliberately absent from every model reported here. Truthful and predictive are not the same property.

2.2 Reconstruction conventions

Four conventions apply to every figure in this report.

Measurement starts in 2012, because the share of moves resolved through the side channel is era-bound: 22 to 46 per cent in 2009 to 2011, and at or below 4.2 per cent afterwards. Thirteen of the seventeen stages carry a measurable wait; three are degenerate because creation fires them under one shared timestamp, and the exit from Published is never written. Negative intervals of two days or less are clamped to zero and larger ones masked, and waits still running at extraction are excluded rather than zero-filled. Because 35.7 per cent of timestamps are date-only, daily resolution is the honest floor and event-level continuous-time modelling is unavailable.

One further property governs what any deployed method may use. Elapsed time in stage is recorded for 99.2 per cent of 2012 entrants but only 14.8 per cent of the current live population, so a method that relies on it is fitted on history it will not have when scoring. Separately, the descriptive figures in this report were recomputed from the raw extracts by a self-contained script sharing no code with the modelling pipeline; the two agree exactly on transition counts and stage ordering, and within half a percentage point on the day-of-week and monthly shares.

2.3 Predictors evaluated

Two prediction targets were addressed. The first is the aggregate stage-week view: the mean wait at each stage in each week. The second is per-analysis: how long one analysis will remain at one stage.

Against the first, the ladder runs from a frozen per-stage historical mean and a textbook queueing treatment [1], through tuned classical recency methods, to spatiotemporal graph networks with and without cross-stage message passing. Against the second, it runs from age-conditional Kaplan-Meier survival curves [2], through per-analysis survival and competing-risks heads [3], to per-visit gradient boosting [4], visit-sequence attention and recurrent models, and block-based hazard heads. Eleven predictors were eventually scored on a common per-visit target. Appendix A lists all of them with the quantity each estimates and the verdict each received.

2.4 How claims were tested

Every experiment ran a fixed chronological split with three random restarts, followed by a rolling-origin backtest at four cut-off dates covering the last two fifths of the timeline, each with a 90-week test window. Three rules were fixed in advance of each run, and they decide how every result below is stated.

1. **The bar is the incumbent, not a weak baseline.** A margin quoted against a textbook queueing formula or a naive last value inflates it by a factor of three or more, so comparisons are made against the strongest thing already available.
2. **A difference counts only if it exceeds its own uncertainty.** Restart-to-restart deviation understates that badly: the deviation of a model-versus-model difference, resampled over analyses, is 212.3 against restart deviations of 17 to 36 for identical configurations.
3. **A result that reverses in the recent era is worse than a flat null.** A configuration that wins on the fixed split and at older cut-offs, then loses at the most recent one, is not reported as a result. Four otherwise clean wins failed this test.

3 Results

3.1 Where analyses wait

Mean time at Preapproval is 284 days against 85 days at the next slowest stage, a ratio of roughly three to one that holds between 2.8 and 3.3 across the several reasonable ways of counting a visit. The distribution at every stage is heavily right-skewed: taken across all thirteen modelled stages the median sits 28.3 days, or 44.6 per cent, below the mean. That gap is not a detail. It is larger than any modelling difference reported anywhere in this work, and Section 3.4 shows that it accounts for most of what appeared to be model skill.

Figure 1 Time spent at each stage, in workflow order, entries from 2012

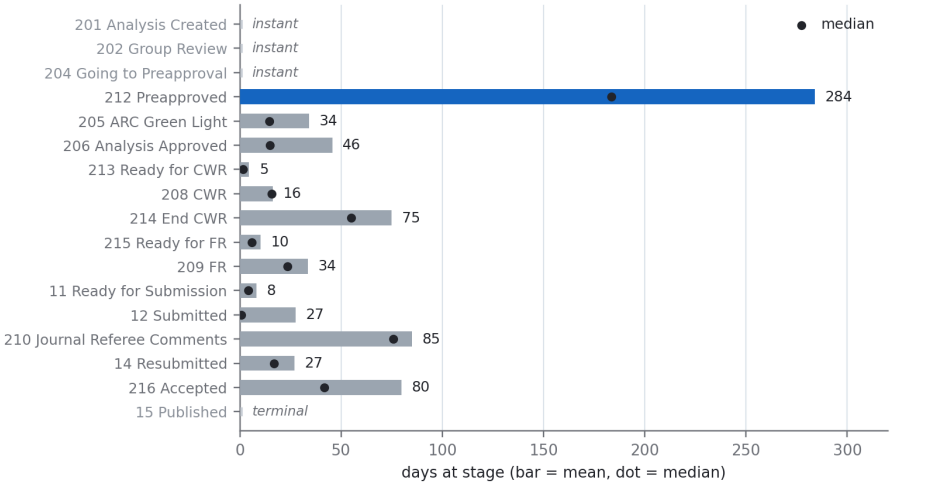


Figure 1 Mean and median days at each stage, in workflow order, computed from the event record. Preapproval is roughly three times slower than any other stage. Creation fires the first three stages under a single shared timestamp, so 99 per cent of visits to them are recorded as zero and any time actually spent there is invisible; the exit from Published is not recorded at all.

3.2 Chain structure carries no information the sequence does not already carry

The approval chain is a graph, and the obvious modelling instinct is to exploit it: let each stage exchange information with its neighbours. Five independent tests, across three architectures and two prediction targets, agree that this earns nothing (Table 2).

Table 2 Five independent tests of whether cross-stage structure helps. None found that it does.

Test	Design	Result
Symmetric message passing	Neighbour mixing against a size-matched per-stage layer, crossed with a recurrent unit against simple pooling	No mixing wins on both training and test data, 6,940 and 12,190 against 6,971 and 12,393
Directed message passing	Messages sent downstream, upstream and both ways, plus an undirected control	Best gain 2.0 against a restart deviation of 72.6, which is nothing
Analysis-level graph	Individual analyses attached as nodes to their current stage, with age and elapsed-time features	No configuration wins; informationless placeholder features score the same as real ones
Completed analyses only	Kick-back and parked analyses removed, finished work only	The no-mixing advantage is stable across eras at plus 2.10 days, interval excluding zero
Mixing inside the final models	Neighbour exchange switched on and off in the best configuration	Costs 0.83 days on one target, null on the other, worse on the slow tail, and reverses by era

A representation diagnostic contradicts the usual justification directly: measured on the internal state, the mixed representation is *less* robust to noise than the unmixed one, so the mechanism is not smoothing away error. The reason is visible in the data. The observed chain is exactly the canonical sixteen-edge line, no analysis ever occupies two stages at once across 2.39 million analysis-weeks, and analyses live for 56 to 70 weeks against an eight-week modelling window, so the structure barely changes. Passing messages along a fixed line adds a smoothing step that the per-stage temporal component already performs better.

The conclusion is not that graph methods fail in general, but that where the topology is known in advance to be a line, structure is redundant. A reviewer-to-analysis graph, where the structure is unknown and varies, would be a different object, and Section 5.2 proposes testing it as flat features before any graph is built.

3.3 The weekly view is close to its measurement floor

The median number of exits behind a weekly figure is two at every one of the thirteen stages. Each weekly value is therefore a two-sample mean of a heavy-tailed distribution and carries sampling variance no model can remove. That floor was estimated per cell by repeatedly splitting each cell's exits in half, and validated against an independent estimator with a correlation of 0.99999.

This restructures the comparison with classical methods rather than settling it. Weak classical methods do lose decisively: a textbook queueing treatment scores 20,664 and is beaten by a frozen historical mean at twelve of thirteen stages. Strong classical recency does not lose. A per-stage blend of exponential smoothing with the historical mean scores 12,186.6 against the learned model's 12,190.0, a gap of 3.4 against a measurement uncertainty of 212.3. Both saturate the same

Table 3 Roughly 93 per cent of the error reported by earlier work was never removable by any model. The captured-share point values are not robust: floor and error are dominated by the same few cells, and the interval on the 81 per cent runs from 25 to 1039 per cent.

Quantity, test window	All cells	Cells with $n \geq 2$
Error of a frozen historical mean	12,946	14,762
Estimated measurement floor	12,006	13,687
Error that is attackable at all	940	1,075
Error removed by the best model	759	1,035
Share of the attackable error captured	81%	96%

narrow band of removable error. The learned model's advantage is real only where the target is statistically resolvable, at four stages led by Preapproval, which alone contributes an advantage of 10,590 with an interval well clear of zero.

The aggregate figure is fragile. One test cell, a single week at stage 216 with two exits and a mean wait of 2,752 days, supplies 48.5 per cent of the entire benchmark error. Measuring on a four-week grid raises the median exits per cell from two to eight and grows the model's advantage from about 9 to 21.9 per cent, though that test held the lookback fixed in weeks, so grid width and lookback length remain confounded.

3.4 No predictor beats a per-stage constant

Predicting one analysis at one stage removes the two-sample floor entirely, since every analysis is its own observation. Eleven predictors were scored on that target. Two corrections were required before any comparison between them was meaningful, and together they account for the entire apparent difference between the best model and the simplest possible rule.

Who gets scored on what.. Two of the eleven work in 28-day blocks and therefore cannot see the 39 per cent of visits that finish inside one block. Those are the fast, easy visits. Scored on the 1,640 visits every model reaches, the strongest per-visit model rises from 30.50 to 45.48 days of error while the block-based models do not move at all, because the hard visits were all they ever had.

What each model is estimating.. Some predict a typical duration and others an average one. On a right-skewed distribution the typical value is systematically lower, and lower scores better on average-error metrics without being more correct. A standard retransformation [5], which adds no information whatsoever, converts one into the other and removes the difference.

Every model lands within a fraction of a day of a constant of its own type. The strongest per-visit model moves from 30.50 to 37.99 against a constant of 35.65; the strongest stage-week model moves from 30.56 to 35.59 against the same reference. The five-day gap that made those two look strongest was the distance between a median and a mean, not skill. Between 96 and 100 per cent of every model's error is systematic rather than restart-to-restart variation, so no amount of retraining will change it.

Two rows break the pattern, and the exception is a trade rather than a win. The models that read an analysis's own sequence of past visits are the only ones that do not collapse on the slowest tenth of visits, at plus 9.5 and plus 26.6 days against minus 44 to minus 107 for everything else. They pay for it at Preapproval, where they post the worst errors in the comparison, 193 to 224 days against 153.

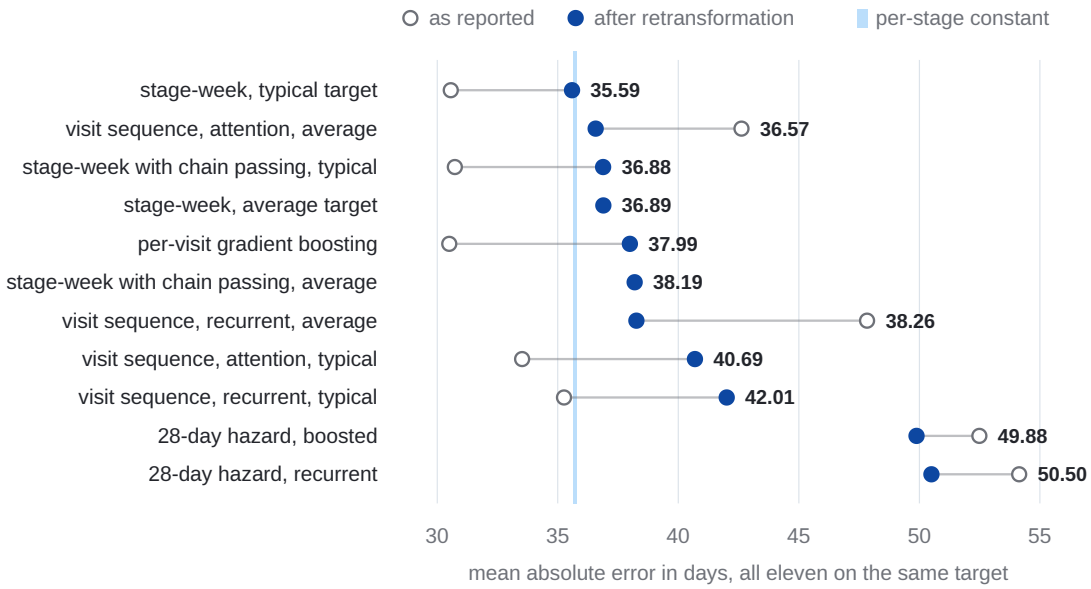


Figure 2 The eleven predictors before and after retransformation to a common target, against a per-stage constant that uses no information about the analysis at all. The hollow marker is the score as originally reported; the filled marker is the same model after the correction. Every predictor lands within a fraction of a day of the constant, or behind it.

3.5 Ranking survives where dates do not

The one property that never broke under scrutiny is ordering. In the per-analysis work, ranking accuracy held at four of four cut-offs under mean ranking and three of four under median ranking, across every weighting scheme and both architectures, while the matching point estimates collapsed twice: once as a demonstrated calibration artefact, and once by reversing, where a clean gain of plus 3.54 days on the fixed split became a loss of 28.5 days at the cut-off closest to deployment.

Separating the reasons an analysis leaves a stage produced the strongest positive result of the whole investigation. A cause-specific model of the advance transition beat its age-conditional classical counterpart, with the headline cut-off confirming a ranking gain of 0.0344, interval plus 0.0225 to plus 0.0526, and an equivalent gain of 0.0402 on a fixed-horizon measure, with support at all four cut-offs [6].

The operational caveat is blunt. Where the deliverable is a triage list rather than a score, sorting by age in stage beats every model tested: precision at five of 0.364 against the best model’s 0.336 overall, and 0.372 against 0.202 at Preapproval, where random selection scores 0.180. Triage needs a sort key, and the sort key is a query.

3.6 The Preapproval slowdown is real, isolated and not compositional

Measuring the slowdown by the median of completed visits silently drops every analysis still stuck in the stage and therefore flatters recent cohorts. Replacing that with the probability of leaving within one year, estimated over all 2,109 visits including those still open, removes the bias and needs no distributional assumption. Every cohort through 2024 has full one-year follow-up.

Clearance at Preapproval falls from 0.76 for 2012 to 2016 entrants to 0.52 for 2020 to 2022, recovering to 0.56 for 2023 to 2024. Across the twelve other stages the weighted mean change over the same period is plus 0.028, and no other stage falls by more than 0.06; the largest move

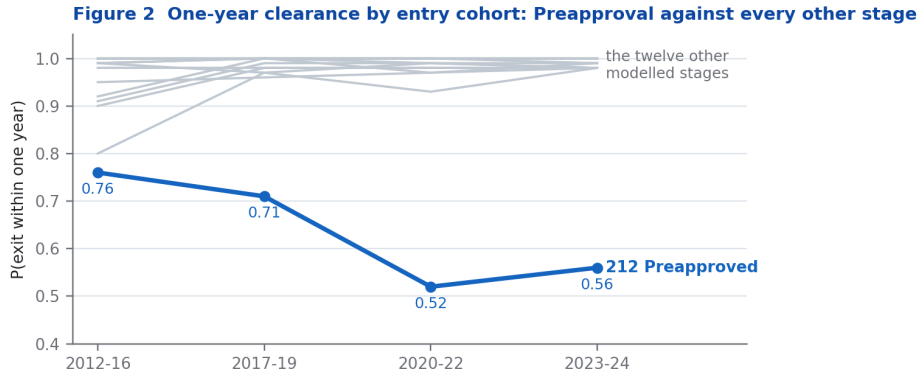


Figure 3 One-year clearance by entry cohort. Preapproval halves while the rest of the chain holds or improves. Cohorts through 2024 only; the 2025 cohort is 59 per cent resolved and is not quotable.

anywhere else is an improvement of 0.18. This is the control that decides the story. A pandemic, a run transition or a platform migration would each have depressed clearance across the whole chain, and none did. Over fourteen years the workflow as a whole became faster or held flat, and Preapproval is the sole exception.

Composition is not the explanation either. The per-group spread is wide, from a median of 90 days for one working group to 374 for another, so a change in which groups enter is the obvious candidate. Reweighting recent cohorts back to the earlier mix moves the median the wrong way, from 298.0 to 321.9 days for 2022 onwards: holding the mix fixed makes the stage look worse, so composition was masking the slowdown rather than causing it.

3.7 Preapproval behaves like a meeting, not like a queue

Three richer seasonal specifications were tested against a simple one, and all three came back null or worse: three harmonics cost a day, three harmonics crossed with stage cost 1.6 days for 76 extra columns, and evaluating the seasonal terms at the forecast date rather than the scoring date degraded accuracy by 5.0 to 7.7 days while improving calibration. A descriptive measurement taken outside the modelling harness explains why.

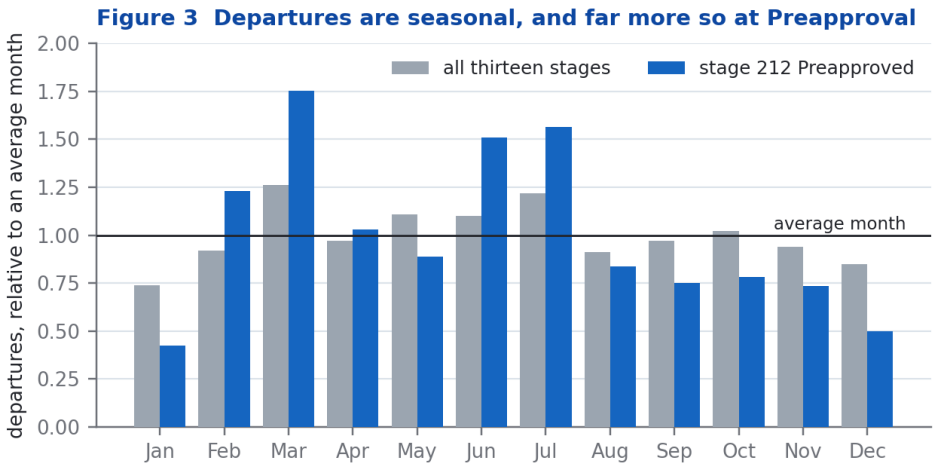


Figure 4 Departures relative to an average month, computed from the event record. The chain-wide profile swings by a factor of 1.7; Preapproval swings by 4.1 between January and March.

Departures from Preapproval are not spread across time. Of 1,928 logged departures since

2012 over 1,279 distinct dates, 34 dates carry five or more simultaneous departures. Drawing the same departures uniformly across the 4,866 active dates yields 0.29 such dates on average and never more than three in two thousand simulations. The largest single day is 25 departures, on 22 July 2016.

Friday carries 18.3 per cent of departures chain-wide, but 26.2 per cent at Preapproval and 25.7 per cent at ARC Green Light, which is 9.1 and 8.6 standard errors above baseline. No other stage deviates, and those two are precisely the stages at which a committee has to convene. On a 30-day hazard controlling for age in stage, the December exit hazard at Preapproval in the first period is 0.00 and the June hazard is 2.42 times baseline. The effect lives in the hazard rather than in the entry snapshot: adding entry month to the predictor moves error from 28.54 to 28.46 days, which is nothing, while the pattern among visits still open at day 30 is undiminished.

The generating process is a step function, not a sine wave: nothing for weeks, then a batch clears on a single day, at the stages where a committee sits, in the months that align with the conference year. No number of harmonics can represent that shape, which is a sufficient explanation for three consecutive seasonal nulls and the strongest argument for obtaining the committee calendar. The link to specific conferences remains a hypothesis until real meeting dates are available.

3.8 The binding constraint is observation, not modelling

Several independent lines of evidence converge on the same limit, and Table 4 groups them because together they define what the present record can and cannot support.

Table 4 Six structural limits of the present record, each established independently.

Constraint	Evidence
Nothing is recorded between the start and end of a wait	Preapproval has exactly one transition in and one out, always from the same predecessor, for 100 per cent of analyses. Within-stage ranking accuracy is 0.5154, a coin flip; a target-conference flag lifts it to about 0.57 and nine further candidate fields add 0.003 on top.
Analysis attributes are a present-day snapshot	Every analysis carries its 2026 values regardless of when its wait occurred. Reading-institution count looked like the strongest predictor until it was split by cohort, where the association falls from 0.32 to 0.13 for later entrants: accumulation, not prediction. Working group is the only attribute with durable signal.
The field the triage sort key needs is largely missing live	Age in stage covers 99.2 per cent of 2012 entrants and 14.8 per cent of the live population. A model built to fill the gap did not ship: on a live holdout of 67 analyses the raw days-since-last-event heuristic matched or beat it.
Repeat visits cannot be represented, and the tail is truncated	Reconstruction yields one record per analysis and stage, so kick-back loops cannot be expressed as second visits. A block-based method over 32 periods of 28 days cannot represent a wait beyond 896 days, which is 4.47 per cent of visits at Preapproval and the longest ones there.
The training pool is about to become biased	Only completed waits are trainable. The resulting optimism is at most 0.8 percentage points today, but the 2023 to 2025 cohort already shows a gap of 0.041, and 112 of the 159 excluded waits began in 2025 or later.

One audit-trail check makes the practical consequence concrete. The worst apparent case of staleness in the live population resolves to a genuine two-stage kick-back that has been untouched

for about 14.6 years since 2011. A raw staleness sort would place decade-old dead analyses above genuinely stuck recent ones, so any deployed triage list needs an abandonment cap or a separate bucket for closure review.

4 Discussion

4.1 One finding, measured four ways

Four separate lines of work answered the same underlying question the same way: the information a coordinator wants is not in the record. Structure carries nothing, because the topology was known in advance. The weekly view is mostly sampling accident. Per-visit predictions are indistinguishable from a per-stage constant once the comparison is fair. And inside the stage that dominates the error, the record holds two timestamps and nothing else across waits running to hundreds of days.

These are not four failures but one finding measured four ways, and it reads operationally: further modelling on the present record has a low ceiling, and the productive investment is in observation. Even a coarse indicator of whether an analysis saw any activity in a given month would exceed what is available now.

4.2 Why nothing is reported without an era check

A specific failure pattern recurred often enough to be worth stating as a result in itself. A configuration wins on a fixed chronological split, wins at older cut-offs, and then loses at the most recent one, which is the period that overlaps deployment.

It occurred four times: with the analysis-level graph; with the per-analysis point estimate, where a confirmed gain of plus 3.54 days became a loss of 28.5 days; with the kick-back arm of the cause-specific work, where a fixed-split gain of 0.2455 reversed to a ranking loss of 0.2047; and across the seasonal sweep, where every configuration lost at the final cut-off.

The rule adopted after the second occurrence is that a reversal of this kind is worse than a flat null, because it looks like a result. Evaluated on a single split, at least four claims in this work would have shipped that a rolling backtest shows to be unsupported. That is the strongest methodological recommendation this report can make to any successor work on the same system.

4.3 What is supportable today

One warning attaches to any interface built on the first four rows. Every predictor examined, including the constant, is systematically optimistic for exactly the analyses that are in trouble: the slowest tenth of visits is under-predicted by 44 to 107 days, and the slowest working groups at Preapproval by up to 300 days.

> **Attention**
>
> A screen that displays a predicted date without displaying that fact will mislead its users in the one situation where they most need not to be misled.

Table 5 Deployment status of each capability examined.

Capability	Status	Basis
Corrected occupancy map of the chain	Supportable now	Two independent implementations agree on all 12,886 stage-weeks and reproduce the live snapshot
Triage list of stragglers per stage	Supportable now, as a query	Sorting by age in stage beats every model at nearly every threshold; needs an abandonment cap
Alerting on Preapproval at weekly resolution	Supportable, with the floor quoted alongside	The average-target model wins the stage at four of four cut-offs; the advantage is real but the aggregate metric is fragile
Ranking which analyses advance soonest	Supportable as an ordering, not a date	Cause-specific advance model beats its classical counterpart, supported at all four cut-offs
Quoting a predicted date per analysis	Not supportable	No predictor beats a per-stage constant once population and estimand are controlled
Any cross-stage structural mechanism	Not supportable	Five independent tests, all null or negative

5 Limitations and Future Work

5.1 Limitations

Every result is conditioned on a single extract taken in July 2026. Analysis attributes carry no version history, so any group or count result inherits a snapshot caveat.

Visit tables built by different reconstructions differ by about 2 per cent, 18,049 against 18,410 visits, most likely in the handling of zero-length visits. Numbers are internally consistent within each; cross-quoting between them carries that difference.

The measurement floor was computed for the average-target formulation only, so verdicts specific to the typical-value formulation remain unresolved, and the comparison between the strongest classical blend and the learned model exists only on the fixed split.

The committee-calendar account is supported by descriptive evidence and by three modelling nulls that it explains. It has not been tested against real meeting dates, because those dates are not yet available.

5.2 Next moves

Three of the items below are requests for data, and they are the ones that matter. The others follow from what has already been measured.

Treat the search over analysis attributes as closed on the present record.. Five architectures and three further specification changes have returned null or worse against properly chosen baselines. A sixth is not the productive next step, and saying so with measurements attached is more useful than continuing.

Ship the descriptive layer and the triage sort key, with the optimism warning attached.. The occupancy map and an age-in-stage ordering are supportable today and require no model. Add an abandonment cap so decade-old kick-backs do not crowd out recent stragglers.

Obtain the review committee and conference calendar.. Meeting dates, and the deadline behind the existing target-conference flag. This has the highest expected value of any outstanding item: the ingestion path is already built, three independent descriptive signatures support the mechanism it would test, and a past deadline carries no snapshot problem.

Extract the identity of who acts on an analysis.. Nothing in the present pipeline knows this. Reviewer load is a different quantity from the stage occupancy that has twice returned null. Three flat features suffice for a first test, and a pseudonymous identifier is enough, so no names need be requested.

Run a fill-rate query on the free-text comment field.. One statement grouped by year and stage. If the fill rate at Preapproval after 2022 is below about 30 per cent, predictive use of the text is dead on arrival and only retrospective attribution survives. No free text exists anywhere in the current extract, so this cannot be answered from what is already held.

Change when a prediction is made before changing what makes it.. Scoring every 30 days while a wait is still running, rather than once on entry, is the change that would let a model see the calendar position it currently cannot represent. It also matches the question a coordinator actually asks: not how long this will take, but when this one, sitting for 90 days already, will move.

Re-check the training pool on the next extract.. The optimism from training only on completed waits is 0.8 percentage points today and rising; the correction is already available.

> **Note**
>
> The evidence narrows the Preapproval slowdown to something specific to that stage: committee size or turnover, meeting cadence, a change in what preapproval requires of an analysis, or a change in the criteria for entering the stage. If any of those has a date attached, that date is worth more than another model.

6 Conclusion

The waiting in the CMS analysis approval chain is concentrated in one stage, and that stage is not slow in the way a queue is slow. Its departures arrive in batches, on Fridays, in the months that align with the conference year, which is the signature of work waiting for a committee rather than work waiting for capacity.

Nothing learned from the present record predicts an individual wait better than quoting the stage's usual duration. That result held across eleven predictors and five model families once two corrections were applied, and 96 to 100 per cent of the remaining error is systematic, so it is not a training problem. The reason is visible in the record itself: between the moment an analysis enters the slowest stage and the moment it leaves, nothing is written down.

Two things do survive. Occupancy can be reported correctly once the side channel is accounted for, and analyses can be ordered by which will advance soonest even though none of them can be given a date. Both ship without a model.

The useful output of a system like this is therefore not a date. It is an ordering, plus an honest account of what the record cannot tell you, and a short list of the observations that would change the answer.

References

[1] John D. C. Little. “A Proof for the Queuing Formula: $L = \lambda W$”. In: *Operations Research* 9.3 (1961), pp. 383–387.

[2] Edward L. Kaplan and Paul Meier. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[3] Jason P. Fine and Robert J. Gray. “A Proportional Hazards Model for the Subdistribution of a Competing Risk”. In: *Journal of the American Statistical Association* 94.446 (1999), pp. 496–509.

[4] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[5] Naihua Duan. “Smearing Estimate: A Nonparametric Retransformation Method”. In: *Journal of the American Statistical Association* 78.383 (1983), pp. 605–610.

[6] Frank E. Harrell et al. “Evaluating the Yield of Medical Tests”. In: *Journal of the American Medical Association* 247.18 (1982), pp. 2543–2546.

A Predictors evaluated

Every predictor built and scored, with the quantity it estimates and the verdict that survived the era check.

Table 6 The eleven scored predictors and the classical ladder beneath them.

Predictor	Estimates	Verdict
Frozen per-stage historical mean	Weekly mean wait per stage	The anchor. Error 12,946, and deceptively strong
Textbook queueing formula	Weekly mean wait per stage	Loses to the frozen mean at twelve of thirteen stages
Rolling mean, exponential smoothing, damped trend, autoregression	Weekly mean wait per stage	Unweighted recency is harmful; the tuned methods reach 12,374 to 12,396
Smoothing blended toward the historical mean	Weekly mean wait per stage	12,186.6, statistically indistinguishable from the learned model
Stage-week network with neighbour mixing	Weekly wait per stage	Superseded. Error 12,393
Stage-week network without mixing, average target	Weekly mean wait per stage	Best error at 12,190 and wins Preapproval at four of four cut-offs
Stage-week network without mixing, typical target	Weekly typical wait per stage	Best average error at about 34.3, but loses Preapproval in every configuration
Per-analysis survival models, feedforward and recurrent	Days remaining in stage	The feedforward win proved to be a calibration artefact; recurrent point estimates reverse by era while ranking holds at four of four cut-offs
Cause-specific model, advance transition	Which exit, and when	The one headline claim that strengthens under the era check
Cause-specific model, kickback transition	Which exit, and when	Reverses completely at the era check; 93 per cent of events fall in 2024 to 2026
Per-visit gradient boosting, two heads	Exit reason, then days at one visit	The reason head fails cleanly, below chance and losing to a stage-prior baseline; the duration head is era-stable, but its margin is a median-versus-mean artefact
Visit-sequence networks, attention and recurrent	Days at one visit, from history	History adds nothing over a snapshot; the only models that hold the slow tail
Block-based hazard heads	Days at one visit	Structurally blind to 39 per cent of visits; furthest behind