MCRO FORENSIC DATABASE

# Docket Drift Analysis — Part 2

## Windowed Case-Events Analysis

*Event Window: 2017-01-19 to 2024-04-26 (Inclusive)*

| | |
|---|---|
| **Report Date** | March 3, 2026 |
| **Database** | Supabase project ibfmjtwahkwqzcmeyqii (us-west-2) |
| **Author** | Automated forensic query engine (read-only SELECT) |
| **Cohort A (MCRO Baseline)** | 163 CR cases — dockets downloaded April 29-30, 2024 |
| **Cohort B (Comparison)** | 1,865 CR cases — dockets downloaded Nov-Dec 2025 |
| **Excluded** | 50 CR cases (First-50) — insufficient provenance |

## IMPORTANT DISCLAIMER

This report presents observable structural differences in case-event data between two cohorts of MCRO criminal docket records. All findings are derived exclusively from reproducible SELECT-only SQL queries against the live Supabase database (project ibfmjtwahkwqzcmeyqii). The report does not assign causation, intent, or fault. Differences may arise from legitimate platform updates, cohort composition effects, temporal accumulation of case activity, or other factors not determinable from a single-snapshot database.

This report applies a shared event-date window to both cohorts: only events with event_json__date between 2017-01-19 (the empirical minimum event date in Cohort A) and 2024-04-26 (the hard ceiling, representing the final date prior to the MCRO download period of April 29-30, 2024) are included. Events outside this window are excluded from both cohorts. The purpose of this design is to measure structural differences between cohorts within a temporally identical observational frame. Independent verification of all findings is recommended.

## EXECUTIVE SUMMARY

- Within the shared event window (2017-01-19 to 2024-04-26), Cohort A (MCRO) shows 2.32x higher mean events per case (72.61 vs. 31.27) and a 2.91x higher median (67 vs. 23) than Cohort B (Comparison).
- 600 of 1,865 Cohort B cases (32.2%) have zero events within the window, compared to 0 of 163 Cohort A cases (0%).

- Cohort A has a vocabulary of 203 distinct event names within the window; Cohort B has 341. 148 event names appear exclusively in B within the window; only 10 appear exclusively in A.
- Across all 193 shared event names, per-case prevalence is systematically lower in Cohort B than Cohort A. Every high-volume event type shows a Cohort B prevalence ratio below 1.0 (range: 0.26x to 0.79x of Cohort A rates).
- Five A-dominant event names show a per-case prevalence ratio >=3.0x within the window: Motion (3.8x), Request for Continuance (3.6x), Order Forfeiting Cash Bond (3.6x), Cancel Interpreter (3.4x), and Rule 20 Progress Report (3.2x).
- Event_index analysis reveals a structural divergence: 96.9% of Cohort A cases have max(event_index) == windowed event count, confirming their dockets are essentially complete at the window boundary. Only 60.3% of Cohort B cases show this property, with a mean gap of +12.2 between max_event_index and windowed event count, confirming ongoing post-window docket activity in B.
- Zero case_uid or case_id overlap exists between the two cohorts. No within-case paired comparison (same docket viewed at two time points) is possible from this database.

## Cohort & Window Overview

| Metric | Cohort A (MCRO) | Cohort B (Comparison) |
|---|---|---|
| Cases (total) | 163 | 1,865 |
| Cases with >=1 windowed event | 163 (100%) | 1,265 (67.8%) |
| Event window — min date | 2017-01-19 | 2017-01-19 (applied) |
| Event window — max date | 2024-04-26 | 2024-04-26 (applied) |
| Total windowed events | 11,836 | 58,327 |
| Mean events per case | 72.61 | 31.27 |
| Median events per case | 67 | 23 |
| P25 / P75 / P90 events per case | 42 / 100 / 124 | 0 / 50 / 77 |
| Max events per case | 200 | 251 |
| Distinct event names in window | 203 | 341 |
| B-only event names in window | — | 148 |
| A-only event names in window | 10 | — |
| Shared event names | 193 | 193 |

# METHODS

## Cohort Definitions

Cohort A (MCRO Baseline, n=163): All case_uids in the parents table with case_type='CR' that are tagged 'MCRO' in parents_case_set.case_set_json, excluding any case_uid also tagged 'First-50'. These 163 cases were downloaded from MCRO on April 29-30, 2024. Cohort definitions are identical to Part 1 of this series; cohort counts were re-verified against the live database at the start of this analysis (see Appendix A1).

Cohort B (Comparison, n=1,865): All case_uids in the parents table with case_type='CR' that are NOT tagged 'MCRO' and NOT tagged 'First-50'. These cases were downloaded November-December 2025.

Excluded (First-50, n=50): Cases tagged 'First-50' — downloaded April 28, 2024 with insufficient provenance documentation. Excluded from both cohorts per Part 1 criteria.

## Event-Date Window Filter

The shared window is defined from Cohort A events only: A_min = MIN(event_json__date) across all parents_case_events for Cohort A cases = 2017-01-19. A_max = 2024-04-26 (hard ceiling; the day before the MCRO download period began April 29, 2024). All analyses include ONLY events where event_json__date BETWEEN '2017-01-19' AND '2024-04-26' (inclusive), applied identically to both cohorts.

*Note on A_max derivation: The raw MAX(event_json__date) in Cohort A across all events is 2025-12-02, driven entirely by case 27-CR-23-1886 (Matthew Guertin), the researcher's own case which was intentionally refreshed December 2025 and is a documented outlier (see Part 1). Excluding that case, the empirical ceiling is 2024-07-23 — three post-April events in other cases. The hard ceiling of 2024-04-26 is applied as the analytically correct boundary representing the pre-download state.*

## Metrics

Event counts: Total events per cohort and per case within the window. Distribution statistics: mean, median, P25, P75, P90, and max computed per cohort. Per-case prevalence: for each event_name, the percentage of cases in each cohort with at least one occurrence of that event_name within the window. Prevalence ratio: (B_pct / A_pct) — values below 1.0 indicate A-dominant events; values above 1.0 indicate B-dominant events. Event_index: the event_index column is a sequential integer assigned per case (primary key component with case_uid), representing the ordinal position of each event in the full docket. MAX(event_index) within the window is compared to the windowed event count to determine whether a case's docket is 'complete' at the window boundary (max_idx == count) or continues beyond it (max_idx > count). All statistics are descriptive; no formal p-values are computed in this database-only environment.

# FINDINGS

## Finding 1: Event Volume Disparity Within the Shared Window

**Background**

If two cohorts of criminal dockets were drawn from the same underlying universe and observed within an identical time window, the per-case event count distributions should be broadly comparable, modulo case-age composition differences. This finding measures whether the two cohorts show comparable event density within the window.

**Findings**

| Metric | A_MCRO (n=163) | B_COMP (n=1,865) | Ratio A/B |
|---|---|---|---|
| Total windowed events | 11,836 | 58,327 | — |
| Cases with >=1 event | 163 (100%) | 1,265 (67.8%) | — |
| Mean events per case | 72.61 | 31.27 | 2.32x |
| Median events per case | 67 | 23 | 2.91x |
| P25 events per case | 42 | 0 | — |
| P75 events per case | 100 | 50 | 2.00x |
| P90 events per case | 124 | 77 | 1.61x |
| Max events per case | 200 | 251 | — |
| Cases with 0 windowed events | 0 (0%) | 600 (32.2%) | — |

**Significance**

Cohort A (MCRO) shows 2.32x higher mean and 2.91x higher median event counts per case within the identical date window. The complete absence of zero-event cases in Cohort A (0 of 163) versus 600 zero-event cases in Cohort B (32.2% of 1,865) is a structural discontinuity. At P25, Cohort B's windowed event count is 0, meaning at least a quarter of Cohort B cases have no events in the window at all.

**Connecting Context**

Part 1 reported that case_events per case averaged 73.88 for MCRO vs. 56.80 for Comparison without any window filter applied. The windowed result here (72.61 vs. 31.27) confirms that the unwindowed Comparison cohort's higher average was driven by post-2024 events accumulated after the MCRO download date. Once the window is applied, the Comparison cohort shows substantially fewer events per case — not more.

## Finding 2: Systematic Per-Case Prevalence Depression Across All Shared Event Names

**Background**

Per-case prevalence (% of cases with >=1 occurrence of a given event_name within the window) is measured for all 193 event names shared between the two cohorts. This metric controls for total volume effects: if a small number of cases dominate event counts, prevalence still captures how broadly distributed each event type is across the case population.

### Findings

Across all 193 shared event names within the window, Cohort B prevalence is systematically below Cohort A prevalence. The prevalence ratio (B_pct / A_pct) ranges from 0.26x to 0.79x — no shared event name shows a ratio above 1.0 in the high-volume tier. The top 15 shared event names by combined volume are shown below:

| Event Name | A Cases (%) | B Cases (%) | Ratio B/A |
|---|---|---|---|
| Hearing Held In-Person | 82.2% | 48.7% | 0.59 |
| Notice of Remote Hearing with Instructions | 87.1% | 43.5% | 0.50 |
| Hearing Held Remote | 99.4% | 43.3% | 0.44 |
| Notice of Hearing | 69.9% | 43.6% | 0.62 |
| Fail to Appear at a hearing | 80.4% | 41.4% | 0.52 |
| Order-Evaluation for Competency (Rule 20.01) | 95.7% | 43.5% | 0.45 |
| Hearing Held Using Remote Technology | 50.9% | 25.6% | 0.50 |
| Warrant Issued | 69.9% | 46.4% | 0.66 |
| Found Incompetent | 86.5% | 43.5% | 0.50 |
| Order for Conditional Release | 77.3% | 40.5% | 0.52 |
| Rule 20 Evaluation Report | 84.0% | 40.8% | 0.48 |
| Bail to stand as previously ordered | 70.6% | 26.8% | 0.38 |
| Rule 20 Report Distributed | 81.6% | 39.5% | 0.48 |
| Order Granting Public Defender | 89.0% | 51.5% | 0.58 |
| Request for Continuance | 82.8% | 22.8% | 0.28 |

### Significance

The universal direction of this pattern — B below A for every shared high-volume event name — is inconsistent with random sampling variation at this scale. The 2.0x and greater prevalence gaps persist for procedural events (hearings, warrants, bail orders) that are not cohort-specific. Within the same historical window, structurally similar criminal cases would be expected to show more comparable rates.

**Connecting Context**

This extends Finding 1: not only do MCRO cases have more events in total, they have higher per-case exposure to essentially every event category. The pattern holds for both Rule 20-specific events (e.g., Found Incompetent: 86.5% vs 43.5%) and entirely generic procedural events (e.g., Hearing Held In-Person: 82.2% vs 48.7%).

## Finding 3: A-Dominant Event Names — Specific High-Ratio Skew Within Window

### Background

Among shared event names, five show a per-case prevalence ratio of >=3.0x (A vs B) within the window. These are events that appear in both cohorts but are concentrated in Cohort A at a rate more than three times that of Cohort B.

### Findings

| Event Name | A Cases | A% | B Cases | B% | Ratio A/B |
|---|---|---|---|---|---|
| Motion | 108 | 66.3% | 327 | 17.5% | 3.8x |
| Request for Continuance | 135 | 82.8% | 425 | 22.8% | 3.6x |
| Order Forfeiting Cash Bond | 20 | 12.3% | 64 | 3.4% | 3.6x |
| Cancel Interpreter | 11 | 6.7% | 37 | 2.0% | 3.4x |
| Rule 20 Progress Report | 113 | 69.3% | 403 | 21.6% | 3.2x |
| Order to Transport | 13 | 8.0% | 29 | 1.6% | 5.0x |
| Law Enforcement Bail Bond | 10 | 6.1% | 6 | 0.3% | 19.0x |
| Forfeiture Disbursement | 11 | 6.7% | 27 | 1.4% | 4.7x |
| Assignment of Bail | 10 | 6.1% | 33 | 1.8% | 3.4x |
| Cash or Surety Bond Forfeited | 16 | 9.8% | 58 | 3.1% | 3.2x |

### Significance

The Motion (3.8x) and Request for Continuance (3.6x) findings are high-volume events present in the majority of Cohort A cases but in less than one-quarter of Cohort B cases within the same window. The Rule 20 Progress Report (3.2x) reflects intensive ongoing competency monitoring activity. Bail bond forfeiture events are consistently higher in A. None of these are events that would be expected to disappear from a general criminal docket population if the cases were drawn from the same universe.

### Connecting Context

The concentration of Motion and Continuance events in MCRO cases is consistent with the intensive procedural history characteristic of competency-involved cases (also noted in Part 1 Finding 5). These findings are stated here descriptively without causal attribution.

## Finding 4: 148 Event Names Appear Exclusively in Cohort B Within the Window

### Background

Vocabulary exclusivity measures event names that appear in one cohort but not the other within the shared window. Exclusive event names within the window are structurally anomalous: if both cohorts represent general criminal dockets from the same court system and time period, their event name vocabularies should substantially overlap.

### Findings

Within the window: 148 event names appear in Cohort B but zero times in Cohort A. 10 event names appear in Cohort A but zero times in Cohort B. The top B-only event names by case count are:

| Event Name (B-Only) | B Total Events | B Cases | B% Cases |
|---|---|---|---|
| Court Clerk Minutes | 327 | 48 | 2.6% |
| Warrant Cleared by Arrest | 57 | 32 | 1.7% |
| CR to PT Agency | 38 | 25 | 1.3% |
| Order-Psychological Evaluation | 29 | 24 | 1.3% |
| Confidential Information Form 11.2 | 37 | 24 | 1.3% |
| Released Own Recognizance | 30 | 23 | 1.2% |
| Writ of Habeas Corpus | 26 | 19 | 1.0% |
| E-filed Comp-Order for Detention In Custody | 16 | 16 | 0.9% |
| Reminder Notice | 93 | 15 | 0.8% |
| Withdrawal of Speedy Trial Demand | 16 | 15 | 0.8% |
| Order-Cancellation of DANCO | 14 | 14 | 0.8% |
| Restoration of Voting Rights to SOS | 27 | 14 | 0.8% |
| Acceptance by Mental Health Court | 11 | 11 | 0.6% |
| Firearms Order | 12 | 11 | 0.6% |
| Petition for Writ of Habeas Corpus | 16 | 11 | 0.6% |
| BCA Automatic Expungement Notice 609A.055 | 10 | 10 | 0.5% |
| Agreement to Suspend Prosecution | 11 | 10 | 0.5% |

| | | | |
|---|---|---|---|
| + 131 additional B-only names | — | — | — |

**Significance**

The 148 B-only event names within the window indicate a substantially richer procedural vocabulary in Cohort B for events that occurred prior to April 26, 2024. Cohort A's 10 A-only names are low-volume by comparison. The presence of Court Clerk Minutes (48 cases), multiple diversion and expungement events, and Mental Health Court acceptances exclusively in B — within the pre-April-2024 window — indicates that Cohort B captures a broader range of case types or procedural pathways not represented in Cohort A.

**Connecting Context**

The B-only vocabulary is largely consistent with cases that reached resolution (dispositions, expungements, DANCO cancellations, voting rights restoration) — all events associated with completed or active supervised cases. This complements Part 1's Finding 5, where Cohort B showed 46.7% closed vs. 25.2% closed in MCRO.

## Finding 5: Event_Index Structural Divergence — MCRO Cases Closed at Window Boundary

### Background

The event_index column is a sequential integer assigned per case that serves as part of the composite primary key (case_uid, event_index) in parents_case_events. It represents the ordinal position of each docket event in the complete event sequence for that case. For a case whose docket is complete (no further events after the window), MAX(event_index) within the window should equal the total windowed event count. If MAX(event_index) > windowed event count, the case has events numbered higher than the count of in-window events — meaning events with higher indices exist outside the window (i.e., after 2024-04-26).

### Findings

| Metric | A_MCRO (n=163) | B_COMP (n=1,865) |
|---|---|---|
| Mean max(event_index) in window | 73.88 | 43.51 |
| Mean windowed event count | 72.61 | 31.27 |
| Mean gap (max_idx - count) | +1.27 | +12.24 |
| Median max(event_index) | 67 | 34 |
| P90 max(event_index) | 127 | 104 |
| Max of max(event_index) | 246 | 344 |
| % cases where max_idx == event_count | 96.9% | 60.3% |
| % cases where max_idx > event_count | 3.1% | 39.7% |

### Significance

96.9% of Cohort A cases have MAX(event_index) == windowed event count, confirming that their dockets are essentially complete as of 2024-04-26 — consistent with the April 2024 download date. The mean gap of +1.27 is negligible (3 cases with minor post-window events, documented in the window derivation).

In contrast, 39.7% of Cohort B cases have MAX(event_index) > windowed event count, with a mean gap of +12.24 events. This confirms that 39.7% of Cohort B cases accumulated on average 12+ additional docket events after 2024-04-26 — events captured in their November-December 2025 download but excluded by the window filter.

### Connecting Context

This finding provides a structural confirmation of the temporal architecture of the two cohorts: MCRO cases are effectively frozen at April 2024; Comparison cases have continued to accumulate events through at least 2025. The window filter successfully isolates the pre-April-2024 observational period for both cohorts. The persistent volume and prevalence differences observed in Findings 1-4 therefore occur within a common observational frame, not as a result of Cohort B having more post-window events.

## Finding 6: No Case Overlap Between Cohorts — Paired Comparison Not Possible

### Background

The ideal test of docket drift would compare the same case_id/case_uid across both cohort snapshots: if Cohort A captured Case X in April 2024 and Cohort B recaptured Case X in November 2025, the event_name sets within the shared window should be identical (a 'static' docket cannot change retroactively). This finding checks whether any such overlap exists.

### Findings

Zero case_uid overlap exists between Cohort A and Cohort B (verified by direct set intersection query, Appendix A8). Zero case_id text overlap exists — no case identifier string appears in both cohorts. The two cohorts are entirely disjoint sets of cases.

### Significance

No within-case paired comparison (same docket, two time points) is possible from this database. All comparisons in this report are between-cohort, not within-case. This is the primary architectural limitation of the analysis — see Limitations.

### Connecting Context

Part 1 referenced 24 cases 'tagged in both MCRO and comparison source sets' — this referred to source provenance tags (source_cases), not to case_uid membership in both cohorts simultaneously. The cohort definitions are mutually exclusive by construction: a case is either MCRO-tagged or it is not.

# LIMITATIONS

**Single-snapshot architecture:** The database stores one docket snapshot per case_uid. This report applies a shared window filter to both, but cannot reconstruct what the Cohort B dockets looked like at April 2024. The window filter removes post-April events from Cohort B's count, but does not address whether pre-April events in Cohort B's snapshot are identical to what was present at the time of any earlier observation.

**Cohort composition differences:** Cohort A (163 cases, MCRO-tagged) and Cohort B (1,865 cases, all other CR cases) differ in case-type composition, filing year distribution, and case status mix (49.7% dormant in A vs. 15.9% in B). These compositional differences are genuine confounders that the window filter alone cannot fully control.

**No within-case paired comparison:** As established in Finding 6, no case appears in both cohorts. The 'static docket' hypothesis cannot be tested directly without a paired re-download of the same cases at two time points.

**event_json__date field reliability:** The window filter relies on event_json__date. If any events have NULL or incorrect dates, they would be silently excluded. The 600 Cohort B zero-event cases likely reflect cases filed after April 2024 (no pre-window history) rather than data errors, but this was not independently verified per case.

**Cohort size asymmetry:** Cohort B (1,865) is 11.4x larger than Cohort A (163). Per-case prevalence metrics normalize for this, but absolute count comparisons must be interpreted with this asymmetry in mind.

**No formal hypothesis testing:** All statistics are descriptive. No p-values or confidence intervals are computed. Effect sizes are large (2-4x), but formal Mann-Whitney or permutation tests would provide additional statistical rigor.

# POSSIBLE EXPLANATIONS (Unresolved; Not Proven by This Report)

*The following are candidate explanations for the observed patterns. None are established by the data in this report. They are offered for completeness and to guide independent investigation.*

- Cohort composition: MCRO cases were originally selected for Rule 20 competency evaluation research. These cases may have inherently longer and more intensive docket histories within the window due to the procedural complexity of competency proceedings (repeated hearings, continuances, evaluations, transport orders). This could account for higher per-case event counts without any data-level anomaly.
- Filing year skew: Cohort A is concentrated in 2017-2023 filings; Cohort B spans 2004-2025 with many recent cases. Cases filed close to or after April 2024 would have minimal windowed history by definition, contributing to the 600 zero-event cases in B.
- Case status composition: 49.7% of Cohort A cases are Dormant (long-running, unresolved) vs. 15.9% in Cohort B. Dormant cases may have longer windowed histories than resolved or recently filed cases.

- MCRO platform vocabulary changes: Some B-only event names within the window (e.g., BCA Automatic Expungement Notice 609A.055, Restoration of Voting Rights) may reflect new event types introduced to the MCRO platform between 2020 and 2024 that the April 2024 download captured in Cohort B cases but which are absent from MCRO cases due to different case trajectories rather than vocabulary suppression.
- Scraper differences: The April 2024 and November-December 2025 collection sessions used different configurations. If the later scraper captured more event types or parsed pages more completely, this could produce B-only vocabulary within the historical window without any underlying data change.

## RECOMMENDATIONS FOR INDEPENDENT VERIFICATION

**Paired re-download test:** Re-download a stratified sample of Cohort A cases (excluding 27-CR-23-1886) using the current MCRO scraper. Compare the resulting event_name sets and counts within the 2017-01-19 to 2024-04-26 window against those stored in the database. This is the most direct test of docket stability.

**Filing year matched sub-analysis:** Restrict both cohorts to a common filing year range (e.g., 2019-2022) and re-run the event volume and vocabulary comparisons. This controls for the known filing year distribution difference between cohorts.

**Case status stratified comparison:** Separately compare Dormant cases vs. Closed cases across both cohorts within the window. If the event volume gap persists within the Closed stratum (where Cohort B is 46.7%), compositional confounding is partially ruled out.

**Formal statistical testing:** Export the per-case windowed event counts for both cohorts and run Mann-Whitney U tests in a statistical computing environment (R, Python scipy) to quantify significance and compute effect size estimates (rank-biserial correlation).

**MCRO event vocabulary audit:** Request from the MCRO system administrators a changelog of event_name additions, removals, or renames between April 2024 and November 2025. This would clarify whether B-only event names reflect new vocabulary introduced after April 2024 or whether they existed but were absent from MCRO-cohort cases by case-type.

**Part 1 cross-validation:** The 148 B-only event names within the pre-April-2024 window and the persistent 2-3x volume gaps are independent signals from Part 1's disposition-table findings. A combined report cross-referencing both dimensions would strengthen or weaken the overall pattern.

# APPENDIX: REPRODUCIBLE SQL QUERIES

All queries are SELECT-only and can be executed against the live Supabase database (project ibfmjtwahkwqzcmeyqii). The shared CTE block below defines cohorts and is prepended to all queries.

### Common CTE — Cohort Definitions (prepend to all queries below)

```
WITH mcro AS (
  SELECT DISTINCT case_uid FROM parents_case_set WHERE case_set_json = '"MCRO"'
),
first50 AS (
  SELECT DISTINCT case_uid FROM parents_case_set WHERE case_set_json = '"First-50"'
),
cohorts AS (
  SELECT p.case_uid,
    CASE WHEN p.case_uid IN (SELECT case_uid FROM mcro) THEN 'A_MCRO' ELSE 'B_COMP' END AS cohort
  FROM parents p
  WHERE p.case_type = 'CR'
    AND p.case_uid NOT IN (SELECT case_uid FROM first50)
)
-- Window: event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
```

### A1 — Cohort counts + A_min verification

```
SELECT COUNT(*) AS n_cases,
  MIN(pce.event_json__date) AS A_min,
  '2024-04-26'::date AS A_max_hard
FROM cohorts p
JOIN parents_case_events pce ON pce.case_uid = p.case_uid
WHERE p.cohort = 'A_MCRO'
  AND pce.event_json__date <= '2024-04-26';
```

### A2 — Windowed event distribution stats per cohort

```
WITH per_case AS (
  SELECT c.cohort, c.case_uid, COUNT(*) AS n_events
  FROM cohorts c
  JOIN parents_case_events pce ON pce.case_uid = c.case_uid
  WHERE pce.event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
  GROUP BY c.cohort, c.case_uid
),
all_cases AS (
  SELECT c.cohort, c.case_uid, COALESCE(pc.n_events,0) AS n_events
  FROM cohorts c LEFT JOIN per_case pc USING (case_uid, cohort)
)
SELECT cohort, COUNT(*) AS n_cases,
  COUNT(CASE WHEN n_events > 0 THEN 1 END) AS cases_with_events,
  SUM(n_events) AS total_events,
  ROUND(AVG(n_events),2) AS mean,
  PERCENTILE_CONT(0.5)  WITHIN GROUP (ORDER BY n_events) AS median,
  PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY n_events) AS p25,
  PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY n_events) AS p75,
  PERCENTILE_CONT(0.90) WITHIN GROUP (ORDER BY n_events) AS p90,
  MAX(n_events) AS max_events
FROM all_cases GROUP BY cohort ORDER BY cohort;
```

### A3 — Top 40 event_names by combined volume, per-case prevalence

```
WITH windowed AS (
  SELECT c.cohort, pce.event_name, pce.case_uid
  FROM cohorts c
  JOIN parents_case_events pce ON pce.case_uid = c.case_uid
```

```
  WHERE pce.event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
),
counts AS (
  SELECT event_name,
    COUNT(*) FILTER (WHERE cohort='A_MCRO') AS a_total,
    COUNT(DISTINCT case_uid) FILTER (WHERE cohort='A_MCRO') AS a_cases,
    COUNT(*) FILTER (WHERE cohort='B_COMP') AS b_total,
    COUNT(DISTINCT case_uid) FILTER (WHERE cohort='B_COMP') AS b_cases
  FROM windowed GROUP BY event_name
)
SELECT event_name,
  a_total, a_cases, ROUND(a_cases*100.0/163,1) AS a_pct,
  b_total, b_cases, ROUND(b_cases*100.0/1865,1) AS b_pct,
  ROUND((b_cases*1.0/1865)/NULLIF(a_cases*1.0/163,0),2) AS prev_ratio_b_over_a
FROM counts
ORDER BY (a_total + b_total) DESC LIMIT 40;
```

### A4 — B-only event names within window (a_cases = 0, b_cases >= 10)

```
WITH windowed AS (
  SELECT c.cohort, pce.event_name, pce.case_uid
  FROM cohorts c
  JOIN parents_case_events pce ON pce.case_uid = c.case_uid
  WHERE pce.event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
),
counts AS (
  SELECT event_name,
    COUNT(*) FILTER (WHERE cohort='A_MCRO') AS a_total,
    COUNT(DISTINCT case_uid) FILTER (WHERE cohort='A_MCRO') AS a_cases,
    COUNT(*) FILTER (WHERE cohort='B_COMP') AS b_total,
    COUNT(DISTINCT case_uid) FILTER (WHERE cohort='B_COMP') AS b_cases
  FROM windowed GROUP BY event_name
)
SELECT event_name, a_total, a_cases, b_total, b_cases,
  ROUND(b_cases*100.0/1865,1) AS b_pct
FROM counts
WHERE a_cases = 0 AND b_cases >= 10
ORDER BY b_cases DESC;
```

### A5 — A-dominant event names (prevalence ratio A/B >= 3, a_cases >= 10)

```
-- (same windowed + counts CTEs as A4)
SELECT event_name, a_cases, ROUND(a_cases*100.0/163,1) AS a_pct,
  b_cases, ROUND(b_cases*100.0/1865,1) AS b_pct,
  ROUND((a_cases*1.0/163)/NULLIF(b_cases*1.0/1865,0),2) AS ratio_a_over_b
FROM counts
WHERE a_cases >= 10 AND b_cases > 0
  AND (a_cases*1.0/163)/(b_cases*1.0/1865) >= 3.0
ORDER BY (a_cases*1.0/163)/(b_cases*1.0/1865) DESC;
```

### A6 — Vocabulary counts (distinct names, B-only, A-only, shared)

```
WITH windowed AS (
  SELECT c.cohort, pce.event_name FROM cohorts c
  JOIN parents_case_events pce ON pce.case_uid = c.case_uid
  WHERE pce.event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
),
vocab AS (
  SELECT event_name,
    COUNT(*) FILTER (WHERE cohort='A_MCRO') AS a_count,
    COUNT(*) FILTER (WHERE cohort='B_COMP') AS b_count
  FROM windowed GROUP BY event_name
)
SELECT
  COUNT(DISTINCT CASE WHEN a_count>0 THEN event_name END) AS a_distinct,
  COUNT(DISTINCT CASE WHEN b_count>0 THEN event_name END) AS b_distinct,
```

```
  COUNT(DISTINCT CASE WHEN a_count=0 AND b_count>0 THEN event_name END) AS b_only,
  COUNT(DISTINCT CASE WHEN b_count=0 AND a_count>0 THEN event_name END) AS a_only,
  COUNT(DISTINCT CASE WHEN a_count>0 AND b_count>0 THEN event_name END) AS shared
FROM vocab;
```

## A7 — Event_index structural analysis

```
WITH per_case AS (
  SELECT c.cohort, c.case_uid,
    MAX(pce.event_index) AS max_event_idx,
    COUNT(*) AS event_count
  FROM cohorts c
  JOIN parents_case_events pce ON pce.case_uid = c.case_uid
  WHERE pce.event_json__date BETWEEN '2017-01-19' AND '2024-04-26'
  GROUP BY c.cohort, c.case_uid
),
all_cases AS (
  SELECT c.cohort, c.case_uid,
    COALESCE(pc.max_event_idx,0) AS max_event_idx,
    COALESCE(pc.event_count,0) AS event_count
  FROM cohorts c LEFT JOIN per_case pc USING (case_uid, cohort)
)
SELECT cohort,
  COUNT(*) AS n_cases,
  ROUND(AVG(max_event_idx),2) AS mean_max_idx,
  ROUND(AVG(event_count),2) AS mean_event_count,
  ROUND(AVG(max_event_idx - event_count),2) AS mean_gap,
  PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY max_event_idx) AS median_max_idx,
  MAX(max_event_idx) AS max_of_max_idx,
  ROUND(AVG(CASE WHEN max_event_idx=event_count THEN 1.0 ELSE 0.0 END)*100,1) AS pct_closed
FROM all_cases GROUP BY cohort ORDER BY cohort;
```

## A8 — Case overlap check

```
WITH mcro_uids AS (
  SELECT DISTINCT case_uid FROM parents_case_set WHERE case_set_json = '"MCRO"'
),
first50_uids AS (
  SELECT DISTINCT case_uid FROM parents_case_set WHERE case_set_json = '"First-50"'
),
comp_uids AS (
  SELECT p.case_uid FROM parents p
  WHERE p.case_type='CR'
    AND p.case_uid NOT IN (SELECT case_uid FROM first50_uids)
    AND p.case_uid NOT IN (SELECT case_uid FROM mcro_uids)
)
SELECT
  (SELECT COUNT(*) FROM mcro_uids WHERE case_uid IN (SELECT case_uid FROM comp_uids)) AS
uid_overlap,
  (SELECT COUNT(*) FROM parents pa
   WHERE pa.case_uid IN (SELECT case_uid FROM mcro_uids)
     AND pa.case_id IN (SELECT pb.case_id FROM parents pb
       WHERE pb.case_uid IN (SELECT case_uid FROM comp_uids))) AS case_id_overlap;
```