Knowledge Base Document (KDB)

# Title:

Handling Data Discrepancy for "New vs Recurring" Classification in Reports using Self-Join and Aggregation Logic

# Problem Statement (Generic)

In reporting systems, when classifying records (e.g., patient encounters) as "New" or "Recurring", discrepancies can occur if the logic only considers the current context (such as the current month) instead of the full historical data.  
  
Symptoms:  
- Inconsistent counts between different reporting systems (e.g., Tableau vs OAC).  
- Incorrect bifurcation of categories like "New" and "Recurring" in visualizations.  
- Misleading trend lines or aggregated summaries.

# Root Cause

This issue typically occurs because:  
- The logic to determine if an entity (like a patient) is "Recurring" needs to look at all past records, not just the current slice of data.  
- Without accessing the full history, entities with previous records are incorrectly classified as "New".  
  
Key technical gap:  
The classification logic works on limited context (e.g., only current month's data) rather than the complete dataset.

# Solution Approach

## 1. Perform a Self-Join on the Unique Identifier

What:  
Self-join the dataset on the entity's unique identifier (e.g., empi\_id for patients).  
  
Why:  
This allows comparison against the entire dataset, capturing past occurrences.

## 2. Create Separate Calculations for "New" and "Recurring"

What:  
Create distinct calculated columns for each category: one for "New" and one for "Recurring".  
  
Why:  
It provides better control in visualizations and helps in clear bifurcation of data.

## 3. Use COUNT(DISTINCT CASE WHEN...)

What:  
Apply conditional filtering first, then aggregate:  
  
COUNT(DISTINCT CASE WHEN [Condition] THEN [Unique\_ID] END)  
  
Why:  
Ensures filtering happens at a row level before aggregation, producing accurate counts.

# Example: Case Study from "IBHC Productivity" Report

Context:  
In the "IBHC Productivity" report for Capital Region Healthcare (CPTL\_NH), discrepancies were observed between Tableau and OAC for "New" and "Recurring" encounter counts.  
  
Table Comparison:  
Tableau - Recurring Value: 25 | New Value: 4  
OAC - Recurring Value: 21 | New Value: 9  
  
Steps Taken:  
- A self-join was performed on empi\_id to pull all past encounters.  
- Separate calculations were created for "New" and "Recurring" encounters based on encounter history.  
- COUNT(DISTINCT CASE WHEN...) was used to correctly count encounters after filtering based on logic.  
- As a result, the OAC report aligned perfectly with Tableau's numbers and visual bifurcation.

# Important Best Practices

- Always check if classification depends on historical records.  
- Use self-join when historical data is required but not inherently available in the current dataset.  
- Separate category calculations (like "New" and "Recurring").  
- Use conditional filtering before aggregation (CASE inside COUNT).  
- Validate counts and trends across months, not just a single snapshot.

# Conclusion

Whenever you face a data mismatch related to "New vs Recurring" (or similar historical-dependent classification):  
- First, verify if the full historical data is being considered.  
- Second, implement a self-join on the unique identifier.  
- Third, apply the right aggregation and separate the calculations logically.  
  
By following this method, you ensure consistent, reliable reporting across systems.