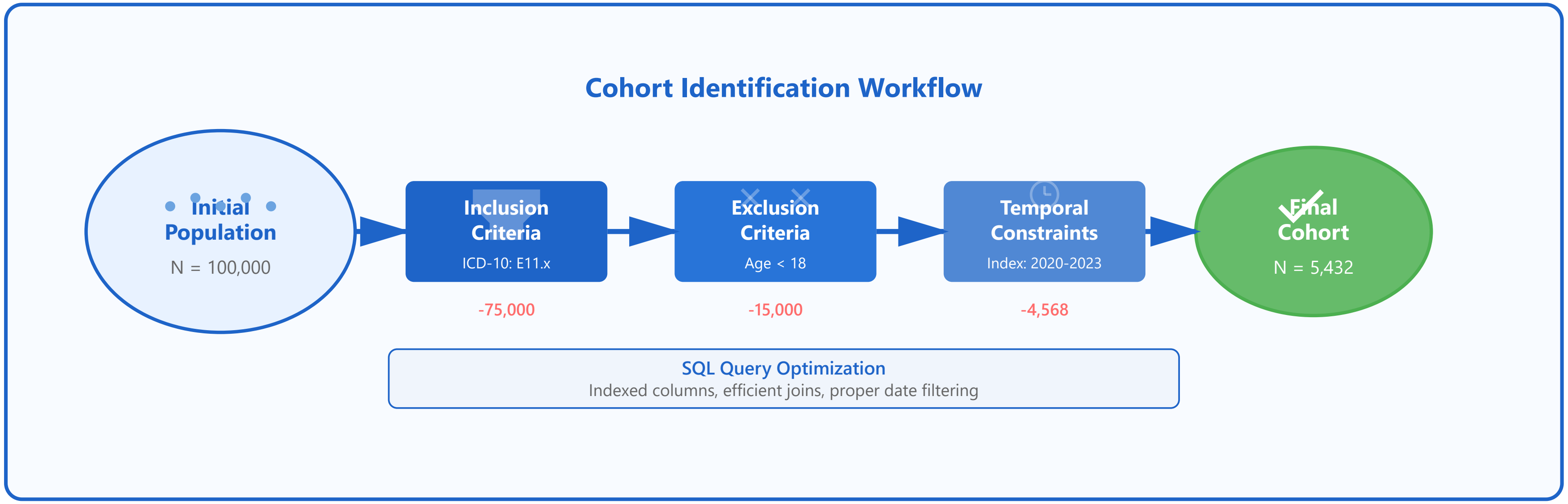
Cohort Identification



✅ Inclusion Criteria

- Age range (18-65 years)
- Primary diagnosis codes
- Minimum encounter count
- Medication exposures
- Lab value thresholds

❌ Exclusion Criteria

- Competing diagnoses
- Prior treatments
- Missing key data
- Insufficient follow-up
- Pregnancy or nursing

⏰ Temporal Logic

🔧 Implementation Tools

- Index date definition
- Washout periods (180 days)
- Follow-up windows
- Event sequence ordering
- Censoring rules

- OHDSI ATLAS interface
- SQL query builders
- Cohort validation metrics
- Attrition diagrams
- Sample size calculations

Detailed Explanations & Examples

✅ 1. Inclusion Criteria: Defining Your Study Population

> Inclusion criteria define the characteristics that patients must possess to be eligible for your study cohort. These criteria should be carefully selected based on your research question and should be specific, measurable, and clinically meaningful. Proper inclusion criteria ensure that you capture the target population while maintaining study validity.

A. Age Range Specifications

📊 Example: Type 2 Diabetes Study

Research Question: Effectiveness of metformin in adult patients with newly diagnosed Type 2 Diabetes

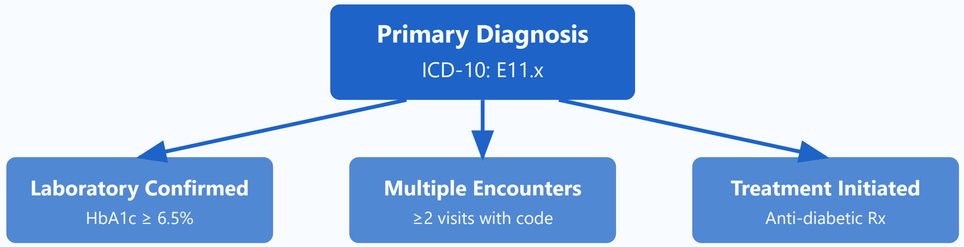
Age Inclusion: Patients aged 18-75 years at index date

Rationale: Excludes pediatric cases (Type 1 more common) and very elderly patients (different treatment protocols)

```
-- SQL Implementation SELECT person_id, birth_datetime FROM person WHERE TIMESTAMPDIFF(YEAR,
birth_datetime, index_date) BETWEEN 18 AND 75 AND condition_concept_id IN (-- Type 2 Diabetes
codes 201826, -- Type 2 diabetes mellitus 443238 -- Diabetes mellitus type 2 without
complication );
```

B. Diagnosis Code Requirements

Diagnosis Code Hierarchy



💡 Best Practice: Multiple Validation Points

Require multiple sources of evidence to confirm diagnosis:

- **Administrative codes:** At least 2 occurrences of ICD-10 E11.x on separate days
- **Laboratory confirmation:** HbA1c ≥ 6.5% or fasting glucose ≥ 126 mg/dL
- **Treatment validation:** Prescription of anti-diabetic medication within 30 days

C. Minimum Encounter Requirements

📈 Encounter Pattern Analysis

Criterion: At least 1 encounter in baseline period (365 days before index) AND at least 1 encounter in follow-up period

Purpose: Ensures continuous engagement with healthcare system and data availability

```
-- Check for continuous enrollment WITH enrollment_check AS ( SELECT person_id,
COUNT(DISTINCT visit_date) as baseline_visits, COUNT(DISTINCT CASE WHEN visit_date >
index_date THEN visit_date END) as followup_visits FROM visit_occurrence WHERE visit_date
BETWEEN index_date - INTERVAL 365 DAY AND index_date + INTERVAL 365 DAY GROUP BY person_id )
SELECT * FROM enrollment_check WHERE baseline_visits >= 1 AND followup_visits >= 1;
```

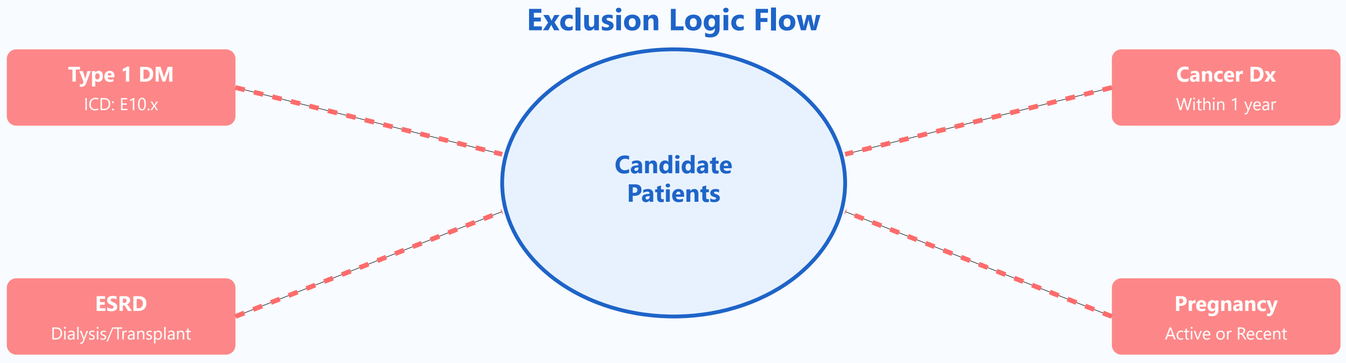
🔑 Key Takeaways for Inclusion Criteria

- Be specific and measurable - vague criteria lead to cohort heterogeneity
- Use multiple validation sources when possible (codes + labs + treatments)
- Consider clinical relevance - criteria should reflect real-world practice
- Document all code sets and versions used for reproducibility
- Balance between strict criteria (homogeneous cohort) and sample size needs

❌ 2. Exclusion Criteria: Refining Your Study Population

Exclusion criteria remove patients who meet inclusion criteria but have characteristics that could confound results, introduce bias, or make them ineligible for the study. These criteria help ensure internal validity and reduce heterogeneity in treatment effects. Exclusions should be justified scientifically and documented transparently.

A. Competing or Confounding Diagnoses



Exclusion Logic Flow

Candidate Patients

Type 1 DM
ICD: E10.x

Cancer Dx
Within 1 year

ESRD
Dialysis/Transplant

Pregnancy
Active or Recent

🚫 Example: Type 2 Diabetes Study Exclusions

Excluded Diagnoses:

- **Type 1 Diabetes (E10.x):** Different pathophysiology and treatment approach
- **Gestational Diabetes (O24.4x):** Temporary condition with different natural history
- **Secondary Diabetes (E13.x):** Due to pancreatic disease, medications, or other causes
- **Active Malignancy:** Confounds mortality and treatment adherence outcomes
- **End-Stage Renal Disease:** Alters drug metabolism and glucose control

```
-- Exclude competing diagnoses SELECT p.person_id FROM cohort_candidates p WHERE NOT EXISTS (
SELECT 1 FROM condition_occurrence co WHERE co.person_id = p.person_id AND
co.condition_concept_id IN ( 201254, -- Type 1 diabetes 4058243, -- Gestational diabetes
4130162, -- Secondary diabetes 4030518 -- Active malignant neoplasm ) AND
co.condition_start_date <= p.index_date );
```

B. Prior Treatment Exposure

Treatment-Naive Cohort Example

Scenario: Study of first-line metformin effectiveness

Exclusions:

- Any prescription of anti-diabetic medications in the 365 days before index date
- Includes: Metformin, Sulfonylureas, DPP-4 inhibitors, GLP-1 agonists, SGLT2 inhibitors, Insulin
- Ensures true "new user" design for causal inference

> **Why This Matters:** Prior treatment exposure can create "depletion of susceptibles" bias. Patients who previously tolerated a drug are systematically different from treatment-naive patients, leading to biased treatment effect estimates.

C. Data Quality Exclusions

Data Completeness Requirements

Complete Data: Include ✓	100% required variables
Partial Data: Review	Missing non-critical variables
Exclude ✗	Missing key covariates (age, sex, baseline labs)

Examples:

- Missing birth date → Exclude
- Missing baseline HbA1c → Exclude*
- Missing race/ethnicity → Include†
- Missing smoking status → Impute

**If required for outcome, †Use "Unknown" category*

D. Insufficient Follow-up Time

⏱️ Follow-up Duration Requirements

Minimum Follow-up Rule: At least 365 days of observation post-index OR until outcome event occurs

Reasons for Exclusion:

- **Death within 30 days:** May indicate acute illness unrelated to chronic disease
- **Loss to follow-up:** No encounters for >180 days and no documented outcome
- **Insurance disenrollment:** Cannot observe outcomes during gap periods

```
-- Ensure minimum follow-up duration WITH followup_check AS ( SELECT p.person_id,
p.index_date, MIN(death.death_date) as death_date, MIN(outcome.outcome_date) as outcome_date,
MAX(obs.observation_period_end_date) as obs_end FROM cohort_candidates p LEFT JOIN death ON
p.person_id = death.person_id LEFT JOIN outcomes outcome ON p.person_id = outcome.person_id
LEFT JOIN observation_period obs ON p.person_id = obs.person_id GROUP BY p.person_id,
p.index_date ) SELECT * FROM followup_check WHERE ( -- Either 365 days of follow-up
DATEDIFF(obs_end, index_date) >= 365 -- OR outcome occurred before 365 days OR (outcome_date
IS NOT NULL AND outcome_date <= index_date + INTERVAL 365 DAY) ) AND (death_date IS NULL OR
death_date > index_date + INTERVAL 30 DAY);
```

🔑 Key Takeaways for Exclusion Criteria

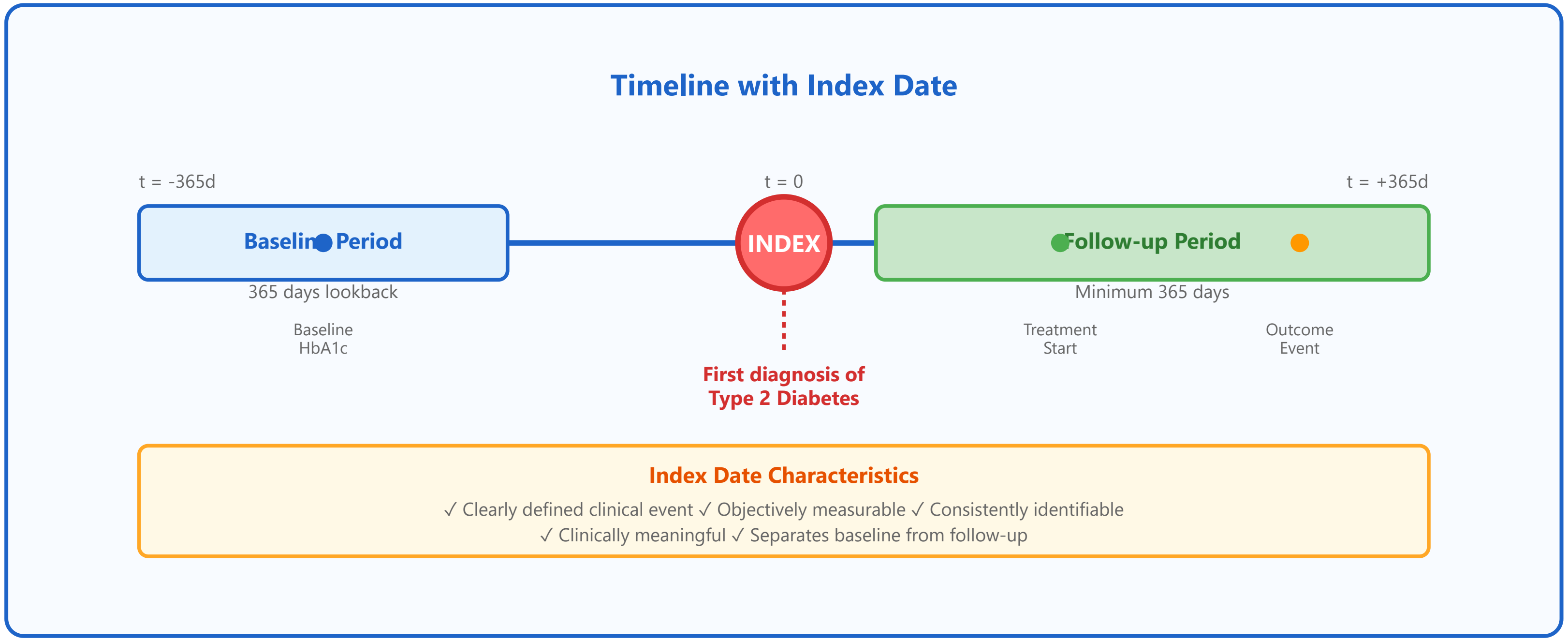
- Justify every exclusion scientifically - arbitrary exclusions reduce generalizability
- Document the order of exclusions (some are mutually exclusive)
- Report attrition at each step in a CONSORT-style diagram
- Consider sensitivity analyses with and without controversial exclusions

- Balance between internal validity (strict exclusions) and external validity (broader population)

⏰ 3. Temporal Logic: Time-Based Cohort Constraints

> Temporal logic defines when events must occur relative to each other and the study observation period. Proper temporal design is critical for establishing causality, avoiding immortal time bias, and ensuring sufficient observation periods. Time windows must be carefully specified for baseline periods, exposures, washout periods, and outcomes.

A. Index Date Definition



Timeline with Index Date
t = -365d
t = 0
t = +365d
Baseline Period
365 days lookback
Baseline HbA1c
INDEX
First diagnosis of Type 2 Diabetes
Follow-up Period
Minimum 365 days
Treatment Start
Outcome Event
Index Date Characteristics
✓ Clearly defined clinical event ✓ Objectively measurable ✓ Consistently identifiable
✓ Clinically meaningful ✓ Separates baseline from follow-up

📅 Index Date Selection Examples

Good Index Date Definitions:

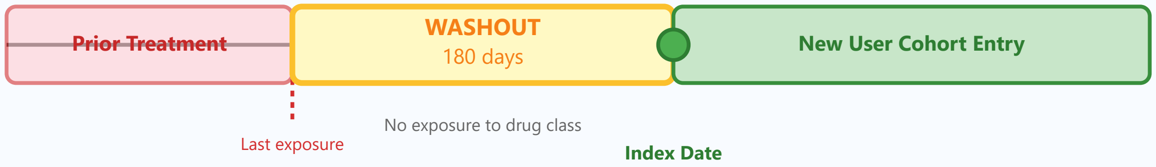
- **New Diagnosis:** First occurrence of Type 2 Diabetes diagnosis code (E11.x) with no prior occurrences in preceding 365 days
- **Treatment Initiation:** First prescription fill date for metformin among treatment-naive patients
- **Procedure Date:** Date of coronary artery bypass graft surgery
- **Laboratory Threshold:** First date when HbA1c ≥ 6.5% with subsequent diabetes diagnosis within 30 days

Poor Index Date Definitions:

- ❌ "Sometime during 2020" - too vague, creates variable baseline periods
- ❌ Random date selection - violates temporal causality
- ❌ Outcome date as index - reverse causality bias

B. Washout Periods

Washout Period Design



Common durations: 90, 180, or 365 days

Depends on drug half-life and disease state

🧪 Washout Period Rationale

Purpose: Ensure patients are "new users" or "incident cases" to enable valid causal inference

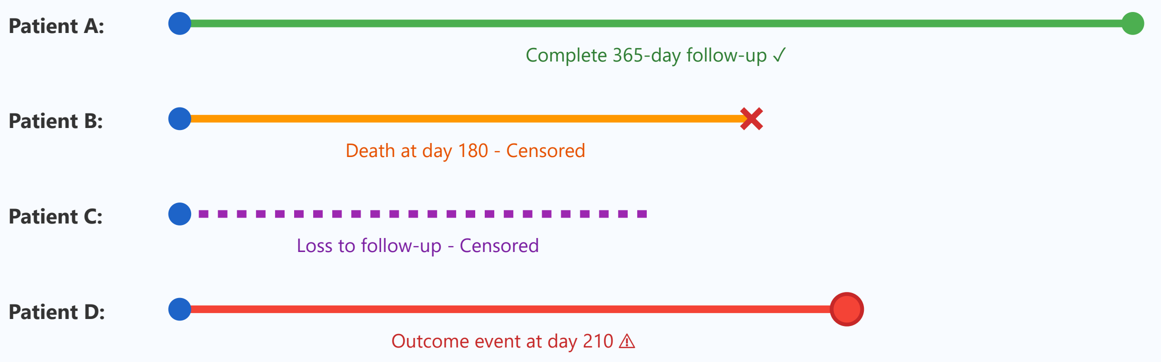
Duration Selection Guidelines:

- **Short washout (30-90 days):** Acute conditions, short-acting drugs (e.g., antibiotics)
- **Medium washout (180 days):** Chronic conditions, standard pharmacotherapy (e.g., antihypertensives)
- **Long washout (365+ days):** Long-acting drugs, conditions with remission/relapse patterns (e.g., biologics, psychiatric medications)

```
-- Implement 180-day washout for metformin SELECT de.person_id, de.drug_exposure_start_date
as index_date FROM drug_exposure de WHERE de.drug_concept_id = 1503297 -- Metformin -- Ensure
no prior metformin in washout period AND NOT EXISTS ( SELECT 1 FROM drug_exposure de2 WHERE
de2.person_id = de.person_id AND de2.drug_concept_id = 1503297 AND
de2.drug_exposure_start_date < de.drug_exposure_start_date AND de2.drug_exposure_start_date
>= de.drug_exposure_start_date - INTERVAL 180 DAY );
```

C. Event Sequence Ordering & Censoring

Censoring Events Timeline



Patient A:
Complete 365-day follow-up ✓
Patient B:
Death at day 180 - Censored
Patient C:
Loss to follow-up - Censored
Patient D:
Outcome event at day 210 ⚠

📊 Censoring Rules Implementation

Censoring Events (whichever occurs first):

- **Administrative censoring:** End of study period (Dec 31, 2023)
- **Death:** Date of death from any cause
- **Disenrollment:** End of continuous insurance/observation period
- **Outcome event:** Date of primary outcome occurrence
- **Competing risk:** Events that prevent outcome observation (e.g., transplant for ESRD outcome)
- **Treatment discontinuation:** If using "as-treated" analysis (optional)

```
-- Calculate person-time and censoring dates SELECT p.person_id, p.index_date, LEAST(
p.index_date + INTERVAL 365 DAY, -- Study end death.death_date, -- Death
obs.observation_period_end_date, -- Disenrollment outcome.outcome_date, -- Outcome event
'2023-12-31' -- Administrative end ) as censor_date, DATEDIFF(censor_date, p.index_date) as
follow_up_days, CASE WHEN outcome.outcome_date = censor_date THEN 1 ELSE 0 END as
event_occurred FROM cohort p LEFT JOIN death ON p.person_id = death.person_id LEFT JOIN
observation_period obs ON p.person_id = obs.person_id LEFT JOIN outcomes outcome ON
p.person_id = outcome.person_id;
```

🔑 Key Takeaways for Temporal Logic

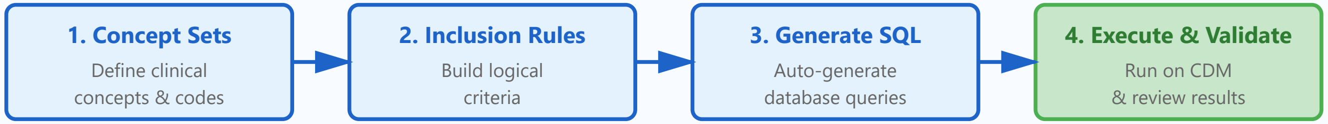
- Index date must be a specific, observable clinical event - not arbitrary
- Washout periods establish "new user" design for unbiased treatment comparisons
- Proper censoring prevents immortal time bias and selection bias
- Time-varying exposures require sophisticated survival analysis methods
- Document all temporal windows clearly for reproducibility

4. Implementation Tools: Building & Validating Cohorts

> Modern cohort identification requires specialized tools that combine clinical knowledge with computational efficiency. These tools help translate research questions into executable database queries, validate cohort definitions, and document the cohort creation process transparently. Proper implementation ensures reproducibility and regulatory compliance.

A. OHDSI ATLAS Platform

ATLAS Cohort Builder Workflow



1. Concept Sets
Define clinical concepts & codes

2. Inclusion Rules
Build logical criteria

3. Generate SQL
Auto-generate database queries

4. Execute & Validate
Run on CDM & review results

Key ATLAS Features

- ✓ Standardized vocabularies (SNOMED, RxNorm, LOINC)
- ✓ Visual cohort definition interface (no SQL required)
- ✓ Automatic attrition reporting
- ✓ Cohort characterization analytics
- ✓ JSON export for reproducibility
- ✓ Multi-database compatibility (OMOP CDM)

ATLAS in Practice

Workflow Example: Type 2 Diabetes Cohort

Step 1 - Create Concept Sets:

- "Type 2 Diabetes" → Include: 201826 (Type 2 DM), 443238 (DM Type 2 w/o complication)

- "Anti-diabetic Drugs" → Include descendants of: 1502809 (Antidiabetic agents)
- "Exclusion Diagnoses" → Include: 201254 (Type 1 DM), 4058243 (Gestational DM)

Step 2 - Define Entry Events: First condition occurrence of Type 2 Diabetes

Step 3 - Add Inclusion Criteria:

- Have at least 1 measurement of HbA1c in baseline period
- Age between 18 and 75 at index
- 365 days of continuous observation before index

Step 4 - Export & Execute: Generate SQL and run against your CDM database

B. SQL Query Optimization

⚡ Performance Best Practices

Query Optimization Strategies:

```
-- ❌ SLOW: Nested subqueries without indexes SELECT person_id FROM condition_occurrence
WHERE person_id IN ( SELECT person_id FROM drug_exposure WHERE drug_concept_id = 1503297 ); -
- ✓ FAST: JOIN with proper indexes SELECT DISTINCT c.person_id FROM condition_occurrence c
INNER JOIN drug_exposure d ON c.person_id = d.person_id WHERE d.drug_concept_id = 1503297; --
Even better: Use temporary tables for complex cohorts CREATE TEMPORARY TABLE
diabetes_patients AS SELECT person_id, MIN(condition_start_date) as index_date FROM
condition_occurrence WHERE condition_concept_id IN (201826, 443238) GROUP BY person_id;
CREATE INDEX idx_diabetes_person ON diabetes_patients(person_id); CREATE INDEX
idx_diabetes_date ON diabetes_patients(index_date);
```

Performance Tips:

- Always index person_id, date columns, and concept_id columns
- Use EXISTS instead of IN for large subqueries
- Partition large tables by year or person_id range
- Filter early: apply WHERE clauses before JOINs when possible

- Use EXPLAIN PLAN to identify bottlenecks

C. Cohort Validation & Quality Metrics

Cohort Quality Assessment Dashboard

Sample Size
5,432
Adequate power ✓

Baseline Balance
SMD<0.1
Well-matched ✓

Follow-up Time
412 days
Median (IQR: 365-730)

Missing Data
2.3%
Key covariates

Index Date Validity
98.7%
Proper sequence ✓

Reproducibility
100%
Re-run consistency ✓

✓ COHORT VALIDATED - READY FOR ANALYSIS

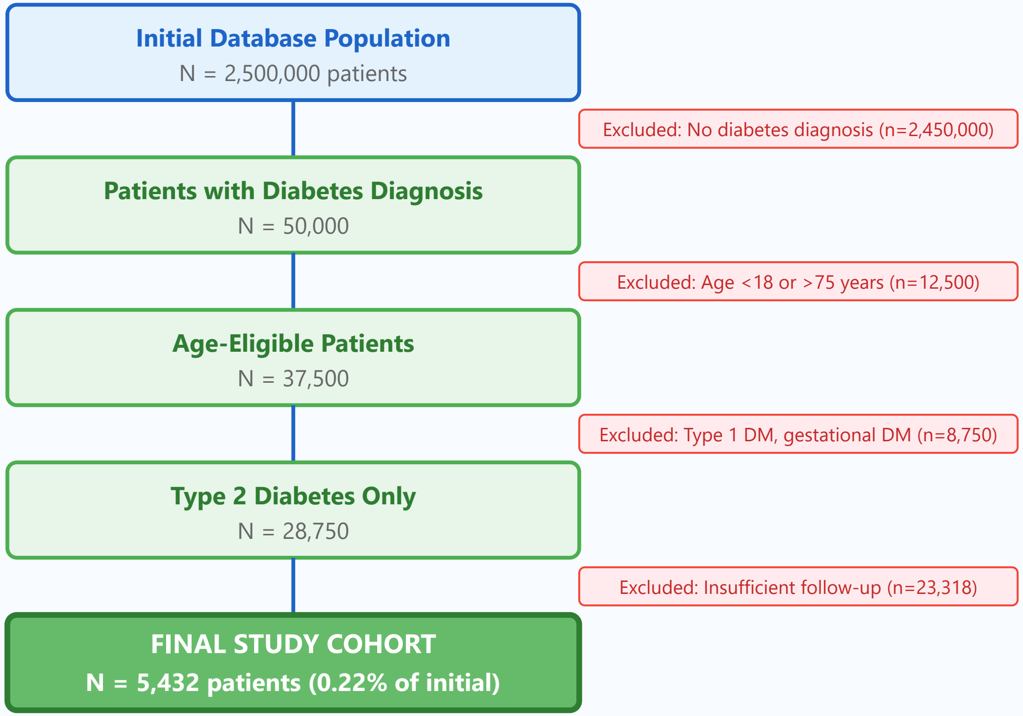
📋 Validation Checklist

Essential Quality Checks:

- ✓ **Face Validity:** Do patient characteristics match clinical expectations?
- ✓ **Temporal Validity:** Are all dates logical and in proper sequence?
- ✓ **Code Set Completeness:** Manual review of 50-100 random records
- ✓ **Attrition Documentation:** Every exclusion step justified and counted
- ✓ **Reproducibility Test:** Re-run produces identical cohort
- ✓ **External Validation:** Compare to published prevalence/incidence rates

D. Attrition Diagrams (CONSORT-style)

Study Population Attrition Diagram



Initial Database Population
N = 2,500,000 patients

Excluded: No diabetes diagnosis (n=2,450,000)

Patients with Diabetes Diagnosis
N = 50,000

Excluded: Age <18 or >75 years (n=12,500)

Age-Eligible Patients
N = 37,500

Excluded: Type 1 DM, gestational DM (n=8,750)

Type 2 Diabetes Only
N = 28,750

Excluded: Insufficient follow-up (n=23,318)

FINAL STUDY COHORT
N = 5,432 patients (0.22% of initial)

E. Sample Size & Power Calculations

📊 Power Analysis Example

Research Question: Does metformin reduce cardiovascular events compared to sulfonylureas?

Assumptions:

- Expected event rate in control group (sulfonylureas): 8% over 3 years
- Clinically meaningful hazard ratio to detect: HR = 0.75

- Alpha = 0.05 (two-tailed), Power = 0.80
- Allocation ratio: 1:1 (metformin:sulfonylureas)

Required Sample Size: 2,716 patients per group (5,432 total)

Expected Events: 217 cardiovascular events needed

```
## R code for power calculation library(powerSurvEpi) power <- powerCT.default( nE = 217, #
Number of events RR = 0.75, # Hazard ratio to detect alpha = 0.05, # Type I error power =
0.80 # Desired power ) # Result: Need 5,432 patients with 217 events for 80% power
```

Post-Hoc Check: After cohort identification (N=5,432), verify that expected event rate and follow-up time will yield sufficient events. If not, consider extending follow-up period or relaxing inclusion criteria.

🔑 Key Takeaways for Implementation Tools

- ATLAS provides standardized, reproducible cohort definitions across institutions
- SQL optimization is critical for large databases - use indexes, temp tables, and proper joins
- Validate every cohort with quality metrics before proceeding to analysis
- Document attrition transparently using CONSORT-style diagrams
- Power calculations should guide sample size requirements and feasibility
- Export cohort definitions as JSON for version control and sharing

Best Practices Summary

Successful cohort identification requires careful attention to inclusion/exclusion criteria, temporal logic, data quality, and validation. Always document your methodology transparently, use standardized tools when possible, and validate your cohort against clinical expectations and published literature. The goal is to create a cohort that is both scientifically valid and computationally reproducible.