Science fair 2026 logbook

https://www.cysf.org/wp-content/uploads/logbooks.pdf

results

conclusions

future implications/ next steps

abstract

A preliminary evaluation of GLP-1RA and SGLT2i use among cardiac rehabilitation patients with atherosclerotic cardiovascular disease and type 2 diabetes.

Kassam K 1 , Rouleau CR 1,4 , Arena R 1,2 , Threlfall T 1 , Aggarwal S 1,4

1 TotalCardiology Research Network, Calgary AB; 2 Department of Physical Therapy, University of Illinois at Chicago, Chicago, IL; 3 Libin Cardiovascular Institute, University of Calgary, Calgary AB

Background: Glucagon like peptide 1 receptor antagonists (GLP-1RAs) and sodium glucose cotransporter 2 inhibitors (SGLT2i) reduce morbidity and mortality among adults with
atherosclerotic cardiovascular disease (ASCVD) and type 2 diabetes but remain underutilized. The purpose of this quality improvement project was to characterize GLP-1RA
and SGLT2i use in a cardiac rehabilitation (CR) setting.
Methods: A retrospective analysis was conducted among patients with ASCVD and type 2 diabetes referred to the TotalCardiology Rehabilitation program (Calgary, AB) between January- December 2024. Medication usage during the 12-week CR was ascertained as standard care by a registered nurse using patient self-report and reconciliation with the Pharmaceutical Information Network. SGLT2i and GLP1-RA use at baseline and program completion were derived by chart review.
Results: Among 551 patients identified as having ASCVD and Type 2 diabetes (79.7X% male, M= 65.1 years old), 77 reported taking a GLP- 1RA, 322 reported taking an SGLT2i, and 60 reported
taking both agents upon CR entry. Among [NUMBER251] patients who attended a

12-week
discharge appointment, 46 reported taking a GLP-1RA, 183 reported taking an SGLT2i, and 40 reported taking both. Among 258 patients who dropped out of CR, 42 reported taking a GLP-1RA, 169 reported taking an SGLT2i, and 32 reported taking both agents upon discharge.
Implications: A significant subset of patients attending CR would further benefit from therapy optimization for ASCVD and diabetes as part of their participation in a multidisciplinary risk reduction program. Work is underway at TotalCardiology Rehabilitation to evaluate a clinical decision support system to enhance medication adherence.

Logs

December 23, 2025

- Met with supervisor to discuss the project goals
- Reviewed the Excel spreadsheet for the first time over 550 patients, 68+ columns
- Immediately noticed the dataset is complex: there are summary flags (isGlpq1Intake, isGLP1Est, isGLP1) AND detailed Med1–Med14 records
- Question arose: why do we need new variables if summary flags already exist?
- Supervisor explained: the summary flags were entered manually and may not match the actual Med records — we need to independently verify using Med1–Med14
- Started thinking about what "medication status at a specific date" actually means — it's not just a simple lookup, you have to find the RIGHT Med record based on timing
- Noted the 3 variables we need:
    - Variable 1: Meds at Initial EST (baseline)

- Variable 2: Meds at FIRST 12-week EST (outcome, not repeat tests)
- Variable 3: Meds at time of separation for dropouts (most recent before leaving)

January 14, 2026

- Opened VS Code, set up Jupyter Notebook
- Copied the file path from the shared server to the notebook
  - Learned: on Windows, right-click the file → "Copy as path" → paste into notebook
  - Had to wrap the path in r:___ to handle the backslashes
- Wrote the data loading code (Part 1)
- Successfully loaded the dataset
- Converted all date columns to date time format → important step, some dates were being read as text
- Ran the initial data completeness analysis
  - Identified which columns have missing data
  - Generated the completeness bar chart
- Thought: some columns have quite a bit of missing data — this is going to matter when we try to match Med records to target dates
- Saved progress and ended for the day

January 15, 2026

- Ran Part 2: Medication Record Structure Analysis
- Key finding: not every patient has Med records in all 14 slots
  - Some patients have only 1–2 Med records
  - A small number have zero Med records at all — these patients will have missing variables

- Generated the Med slot utilization chart — Med1 through Med4 are most commonly filled, usage drops off after Med6
- Found patients who have medication flags set (like isGlpq1Intake = TRUE) but NO Med1–Med14 records at all
    - This is a red flag
        - how can we trust a flag if there's no underlying record?
    - This is exactly why we need to build variables from the Med records directly
- Started thinking about the algorithm: for any given target date, I need to scan Med1–Med14, find the record closest to that date (but on or before it), and pull the GLP1/SGLT2 flags
- Thought about edge cases: what if no Med record exists before the target date? What do we do?

January 16, 2026

- Ran Part 3: Temporal Logic Validation
- Checked whether dates follow the expected sequence:
    - DateReferralReceive → DateInitialMed → DateInitialEST → DateMedAfterInitialEst → Date12WeekEst
- Found some patients where dates are out of order which are likely data entry errors
- Calculated key time intervals:
    - Referral to Initial EST: median around 38 days
    - Initial EST to 12-week EST: median around 86 days (close to the expected 84 days / 12 weeks)
    - But some patients have 12-week EST dates that are >150 days after Initial EST
- This raised a concern: if Date12WeekEst can be >150 days out, it might be a REPEAT test, not the first 12-week assessment

    - Our Variable 2 specifically needs the FIRST 12-week EST
    - Will need to add a filter for this later
- Generated the temporal analysis charts — the histogram of Initial to 12-week interval clearly shows most patients cluster around 84 days, but there's a tail extending past 150 days

---

January 19, 2026

- Came back on Monday, ran Part 4: Medication Flag Consistency Validation
- This was the most important diagnostic step
- Compared the existing summary flags against what the Med1–Med14 records actually show at each timepoint
- Found discrepancies — some patients have flags that don't match their Med records
    - Example: isGlpq1Intake says TRUE but no Med record near DateInitialMed shows GLP1 = TRUE
    - This confirms the data entry inconsistency problem the supervisor warned about
- Calculated consistency rates for intake, post-EST, and current flags
- Generated the consistency visualization — clearly shows where the gaps are
- Thought: this is exactly the problem our new variables will solve — by going directly to the Med records, we bypass any errors in the summary flags
- Feeling more confident about the approach now

January 20, 2026

- Did not write new code today — spent the day thinking through the algorithm design
- Drafted the logic on paper:
    - Input: a patient row, a target date, a medication type (GLP1 or SGLT2)
    - Step 1: Scan Med1–Med14 for all records with a date

    - Step 2: Separate into "before or on target date" and "after target date"
    - Step 3: If records exist before target → use the closest one (highest priority)
    - Step 4: If no records before target → check for records within 7 days after (grace window)
    - Step 5: If still nothing → return None with explanation
    - Output: medication status, source record, days difference, confidence flag
- Realized that Variable 3 needs DIFFERENT logic: instead of "closest to target date," it needs "most recent before separation date"
    - Example: if a patient separates June 15, and has Med records on May 1 and June 10, we want June 10 (most recent), not necessarily the closest
    - This is an important distinction — will need a separate function for Variable 3
- Wrote down the confidence flag definitions:
    - high: record within 7 days before target
    - medium: record 8–30 days before target
    - low_post_date: used a record after target (grace window)
    - no_med_records: patient has no Med records at all
    - no_suitable_record: has Med records but none near the target date

---

January 21, 2026

- Coded the core temporal matching function (find_med_status_at_date)
- Built Variable 1: Medication Status at Initial EST
- Ran the code — it worked on the first attempt (surprisingly!)
- Checked the output:
    - Got medication status (TRUE/FALSE) for most patients
    - Confidence flags distributed as expected

        - majority high confidence
    - Some patients returned None
        - these are the ones with no Med records near Initial EST
- Tested a few patients manually using the lookup function to verify:
    - Picked a patient, found their DateInitialEST, manually looked at Med1–Med14 dates
    - Confirmed the algorithm selected the correct Med record
- Felt good about Variable 1 — logic is clean and results make sense
- Thought: the confidence flags are really useful — they let us see exactly how reliable each data point is

---

January 22, 2026

- Built Variable 2: Medication Status at 12-Week EST
- Initially used the same logic as Variable 1 — just pointed at Date12WeekEst instead of DateInitialEST
- Code produced results for all patients with a 12-week EST date
- But then remembered the concern from Day 4: some patients have Date12WeekEst >150 days from Initial EST
    - These are likely REPEAT tests, not the first 12-week assessment
    - Our requirement is specifically the FIRST 12-week EST
- Wrote a diagnostic check: calculated days from Initial EST to Date12WeekEst for all patients
    - Found 39 patients (13.4%) with Date12WeekEst >150 days
    - These 39 need to be excluded from Variable 2
- Realized the current code doesn't filter these out — needs to be fixed
- Made a note to fix this tomorrow

January 23, 2026

- Added the 60–150 day filter to Variable 2
  - Only accepts Date12WeekEst if it falls 60–150 days after DateInitialEST
  - 39 patients excluded and flagged as "rejected_repeat_est"
  - 251 patients retained as valid first 12-week assessments
- Ran the corrected code
  - output now clearly shows "Used: 251, Rejected: 39"
- Moved on to Variable 3: Medication Status at Separation
- Initially built it using the same "closest record" logic as Variables 1 and 2
- Then caught the error: Variable 3 needs MOST RECENT logic, not closest
  - "Most recent" = the latest Med record on or before separation date
  - "Closest" = could pick a record after separation if it's closer in time
  - For dropouts, we want their LAST KNOWN medication status before they left
- Rewrote Variable 3 with a separate function (find_most_recent_med_at_date)
  - This function finds all Med records on or before SeparationDate
  - Selects the one with the latest date (not smallest difference)
- Tested with the lookup function on a dropout patient — confirmed it's using the most recent record

---

January 26, 2026

- Built the patient lookup function
  - allows checking any individual patient by name or ULI
- Spent the morning testing 5–6 patients across different categories:
  - A patient who completed the program (has Var1 and Var2)
  - A patient who dropped out (has Var1 and Var3)
  - A patient with low confidence flags

    - A patient with no Med records
- All results checked out
    - the algorithm is selecting the correct Med records
- One interesting case: a patient where the closest Med record was 12 days before the Initial EST
    - Confidence flagged as "medium" — correct behavior
    - The lookup function clearly showed why this record was chosen
- Thought: the audit trail (showing source, days_diff, confidence) is really powerful — anyone can verify the logic without understanding the code

---

January 27, 2026

- Ran the statistics code to calculate all numbers needed for the results paragraph
- Key findings:
    - Baseline GLP-1RA use: 28/538 (5.2%) — quite low given guideline recommendations
    - 12-week GLP-1RA use: 54/295 (18.3%) — significant increase
    - SGLT2i increase was even more dramatic: 21.0% → 72.9%
    - Both agents at baseline: only 4.5% — very few patients on dual therapy at entry
- The chi-square tests confirmed statistical significance for both GLP-1RA and SGLT2i increases
- Thought: the SGLT2i result (21% → 73%) is remarkable
    - this program is clearly having a major impact on medication optimization
- Noted the question about denominators: should percentages be based on patients with data, or total cohort? Decided on patients with data (more accurate, standard practice)

January 28, 2026

- Received results from another analyst to compare against
- Ran the discrepancy check code
  - compared old summary flags vs. new variables side by side
- Confirmed that the existing summary flags and our new variables are measuring slightly different things:
  - Old flags: based on manually entered summary columns
  - New variables: independently derived from Med1–Med14 records with confidence scoring
- Both are valid but serve different purposes:
  - Old flags: quick reference for clinical staff
  - New variables: research-grade, independently verified, auditable
- Discussed the "discharge" definition with supervisor:
  - Their discharge metric (360 patients) includes everyone with a most recent med list
  - Our Variable 3 is specifically for dropouts only
  - Both are useful — just measuring different things

January 29, 2026

- Spent the day drafting the full write-up:
  - Brief description
  - Problem statement (detailed — covers clinical context, data challenge, research gaps)
  - Methods (algorithm design, confidence flags, quality validation)
  - Analysis and results
  - Conclusion and recommendations
- Reviewed the write-up for accuracy — all numbers match the code output
- Thought about future work:

- The 39 excluded patients (repeat ESTs) could be recovered with manual chart review
- Could explore predictors of medication initiation during CR
- The pipeline could be adapted for other medications or timepoints

---

January 30, 2026 — Day 14: Preparation for Doctor Meeting

- Reviewed all outputs one final time before the meeting
- Confirmed the 3 variables are complete and correct:
  - Variable 1: Baseline meds at Initial EST ✓
  - Variable 2: Outcome meds at FIRST 12-week EST (251 patients) ✓
  - Variable 3: Dropout meds at separation (most recent before leaving) ✓
- Prepared key talking points for the meeting:
  - All patients have ASCVD and T2D (inclusion criterion)
  - Significant increases in both GLP-1RA and SGLT2i use during the program
  - Data quality is strong — majority of matches are high confidence
  - Can verify any individual patient in seconds using the lookup function
- Practiced a quick demo using lookup_patient()

January 22nd

- showed progress to Dr. Rouleau and was thoroughly impressed with the work completed
- walked through the finished copy which made major progress.
  - significant innovation and part of study that could get manuscript
- presenting findings at Tine Haworth Cardiovascular Research Day 2026.
  - made an abstract/ making a poster for the event

timetable

Tasks

- [ ] research
- [ ] brainstorm of project
- [ ] project selection
- [ ] understand the dataset
- [ ] design variables
- [ ] validate variables
- [ ] finalize results

Timetable (Planned Schedule)

Objectives	Goal	Target date
1	Research projects and different innovations	November
2	Brainstorm potential projects of interest and identify their potential	December
3	Finalize selection of project	January
4	Learn to understand the 551 patient dataset	Jan 13-15
5	Design the variables for optimal flow	Jan 19-24
6	Validate the variables	Jan 27-30
7	Finalize results,	Feb 2
8	Add phase 2 - stress budgeting	Feb 15th
9	Research wearables	Feb
10	Develop modelling to utilize wearable data for prediction	March

11	Tie everything together under topic of Cardiac rehab	

project selection

Worked with perplexity to brainstorm potential projects, and then with a cardiac rehab centre, gained access to data from patients to work on a project.

brainstorm for projects (1)

brainstorm for projects

- Learning fair and value optimized treatment rules for Cardiovascular prevention

  Core idea: instead of just predicting risk, you **redesign who gets treated** (e.g., statins, intensive blood-pressure therapy) to maximize health benefit while minimizing unfairness between groups.

  What you'd do:

  - Use an open CVD cohort (risk factors, treatments, outcomes).
  - Train a solid risk model, then simulate treatment policies like "treat if 10-year risk > X%."
  - Add fairness constraints (e.g., equalize under-treatment rates between sexes or ethnic groups) and cost-effectiveness constraints using published thresholds.
  - Output a tool that lets a user slide thresholds and instantly see changes in events prevented, costs, and inequity.

  Why it's top-notch: current reviews stress that AI in cardiac care often ignores fairness and value despite big concerns about bias and cost; a project that **quantitatively fixes this** is very forward-looking.

- SGLT2/ GLP1

Core concept: “AI cardio-diabetes co-pilot” for GLP-1/SGLT2 + rehab

One project:

An AI co-pilot that helps clinicians and patients decide: “Which therapy, when, for whom, with what rehab support?” – for people with CVD + type 2 diabetes enrolled in your program.

It would have four tightly integrated “brains,” backed by your dataset:

1. **Benefit brain – who benefits from GLP-1/SGLT2 (and when):**
    - Use your med-timing variables (isGLP1Intake, isSGLT2Intake, isGLP1Est, isSGLT2Est, isGLP1, isSGLT2, dates) plus lipids, BP, BMI, PeakMET, PHQ-8/GAD-7 and program outcomes.
    - Train ML models to estimate:
        - Probability of good functional response (e.g., large PeakMET gain, BP improvement) with vs without GLP-1/SGLT2 exposure.
        - Effect of **early vs late** initiation relative to referral and initial EST on those outcomes.
    - This mirrors cutting-edge work on phenotype-based targeting of SGLT2i vs GLP-1RA, but you apply it to **cardiac rehab outcomes**, which current T2D AI papers do not.
2. **Drop-out brain – who will fail rehab without extra help:**
    - Use ProgramStatus, SeparationStatus, SeparationDate, PHQ-8, GAD-7, baseline fitness (PeakMET), and risk factors to predict non-completion and poor improvement.
    - Then link those predictions to GLP-1/SGLT2 review status and timing to ask:
        - “If we give this patient an earlier GLP-1/SGLT2 review plus extra support, how much does their completion probability improve (model-estimated)?”
    - Current AI-rehab work looks at personalization and predictors separately; you’d explicitly connect **pharmacotherapy + psychosocial**

risk + rehab adherence in one model.

3. **Phenotype brain – discovering hidden subgroups that guide decisions:**
    - Cluster patients using all your structured data: clinical risk factors, mental health, PeakMET, BMI/waist, smoking, med status and trajectories, and program outcomes.
    - For each phenotype, show:
        - Typical GLP-1/SGLT2 strategy.
        - Typical rehab outcome and drop-out risk.
    - Then the platform can say: "This new patient looks like Phenotype 3 (high mental-health burden, low fitness, no GLP-1/SGLT2 yet) – in our data, early GLP-1 + extra support produced the best result."
4. **Simulation brain – a mini target-trial / policy simulator:**
    - Inspired by target trial emulations and AI-driven pharmacotherapy papers, you let users test **strategies** on your cohort:
        - "What if we start GLP-1RA at referral in all eligible?"
        - "What if we use your benefit-brain to selectively start GLP-1/SGLT2 only in patients with model-predicted large gains?"
        - "What if we auto-flag high drop-out risk patients for tele-coaching?"
    - The system recomputes projected distributions of PeakMET, BP, and completion for your 300–400 patients under each strategy.

Together, this becomes an **AI platform that:**

- Selects and times GLP-1/SGLT2 more intelligently.
- Identifies at-risk patients early and suggests concrete interventions.
- Understands your population's phenotypes.
- Lets clinicians test "what-if" policies before changing real care.

Existing work does pieces of this—precision GLP-1/SGLT2 response prediction, AI in rehab, phenotype-based targeting, target trial emulation—

but **not a single, rehab-integrated, cardiorenal-therapy-aware co-pilot built around one real program dataset.**

How this can genuinely improve the world (not just impress judges)

You can make a credible case that, if scaled:

- It could help **cardiac rehab programs** decide where to focus limited GLP-1/SGLT2 access (which patients, when) to maximize functional recovery and completion, not just lower A1c.
- It addresses the real, current challenge of integrating new GLP-1/SGLT2 guidelines into complex cardiometabolic care pathways, which major societies now emphasize.
- It turns raw program data into an actionable **service**: a prototype interface your partnering clinic could actually use as a decision-support and quality-improvement tool.

That's exactly the kind of project that can lead to pilot deployments, industry or hospital collaborations, and truly **change practice** beyond the fair.

Ensuring novelty vs the literature

We obviously can't mathematically prove "never been done," but the literature strongly suggests your specific integration is new:

- GLP-1/SGLT2 AI work focuses on predicting drug responses (e.g., HbA1c benefit) and individual treatment effects, not rehab outcomes, timing in rehab, or program completion.
- Cardiac rehab AI work focuses on exercise prescriptions and adherence prediction, not AI-guided GLP-1/SGLT2 strategies.
- Pharmacotherapy AI reviews describe the *potential* of integrated, multimodal platforms for cardiovascular pharmacotherapy, but mostly as a future vision.

Your “co-pilot” sits at the intersection of all three—and uses **real program data** in a concrete prototype. That intersection is where originality is highest.

Your interface: how everything appears in one project

Think of a single web dashboard with three main tabs:

- **Patient view:**
  - Input a (de-identified) patient profile (or pick an example).
  - Shows phenotype label, predicted rehab success, GLP-1/SGLT2 benefit estimate, and suggested timing.
  - Simple explanation panel: “Key drivers: high BMI, low baseline MET, high PHQ-8.”
- **Strategy simulator:**
  - Sliders/buttons to choose policies (early vs late GLP-1/SGLT2, targeted vs universal, extra support for top-risk quartile).
  - Outputs projected distributions of MET improvement, BP control, and completion for the whole cohort.
- **Insight explorer:**
  - Visualizations of discovered phenotypes, their characteristics, and their observed outcomes under different therapies.

- SGLT2/ GLP1 part 2: rehab trajectory

New idea A: "Rehab trajectory forecaster" with adaptive exercise prescriptions

Concept: Predict each patient's 12-week PeakMET trajectory based on baseline + early signals (initial EST, PHQ-8/GAD-7, med status), then generate **personalized weekly exercise targets** that maximize MET gain while minimizing drop-out risk.

Why nationals-caliber:

- Recent AI cardiac rehab work shows ML can personalize VO2max/exercise prescription and boost adherence, but real-world implementations are scarce.
- Your data has the perfect longitudinal structure (initial → 12-week EST, program status, mental health).

What you'd build:

- Time-series model (LSTM or gradient boosting with time features) predicting MET trajectory.
- Optimization layer: "Given this patient's predicted trajectory and drop-out risk, what weekly MET targets maximize 12-week gain?"
- Interface: input patient → see predicted trajectory ± personalized weekly plan.

Novelty: Few student projects do **prescriptive analytics** for rehab (not just prediction). Your mental health + GLP-1/SGLT2 integration makes it unique.

Feasibility: Perfect fit for your data. Complements the SGLT2/GLP-1 project.

New idea B: "Cardio-diabetes phenotype atlas" with treatment recommender

Concept: Discover 4-6 data-driven phenotypes from your full dataset, validate them against outcomes (MET, BP, completion), then build a recommender: "Patients like you typically do best with [GLP-1 early + extra psych support / SGLT2 only / standard rehab]."

Why nationals-caliber:

- Phenotype discovery + precision medicine is exploding in diabetes/cardiology, but mostly in large cohorts. Doing it convincingly in a smaller, real program dataset with actionable recommendations is impressive.

What you'd build:

- Autoencoder or advanced clustering → phenotype discovery.

- For each phenotype → characteristic treatment patterns, outcomes, drop-out rates.
- Simple recommender system (rule-based or ML) suggesting optimal strategy.

Novelty: Ties phenotypes explicitly to **program-level decisions** (meds + rehab support), which literature papers don't operationalize into tools.

Feasibility: Excellent. Pure data science, stunning visualizations.

background info

- Reviewed all column definitions and meanings in the dataset with the clinical team
  - reviewed the variable bank available
- Completed research on :

  GLP1

  These mimic specific hormones, ramping up glucose-dependent insulin from beta cells while reducing glucagon from alphas, plus hitting brain relays in NTS to curb appetite and inflammation via mesolimbic tweaks. Various studies have displayed 13-20% drops in MACE (MI, stroke, CV death), 14% in CV mortality, and 13% all-cause mortality versus placebo in T2D cohorts, even spilling benefits to non-diabetics with obesity.

  SGLT2

  SGLT2 inhibitors stop the kidneys from taking back all the glucose, so you increase urination and blood sugar falls even without much insulin around. Bonus: they trim weight, lower BP, protect the heart from failure no matter the pump strength, and slow kidney decline – works great even non-diabetics sometimes. Downsides include increased risk of genital mycotic infections (yeast infections) and urinary tract infections due to glucose excretion in urine.

- Learned that all patients in the registry have both ASCVD and Type 2 diabetes — this is an inclusion criterion

  ASCVD + T2D

  T2D elevates ASCVD risk, making it the top cause of morbidity and mortality in affected patients.

  Nearly one third of T2D patients have established ASCVD, manifesting as coronary heart disease, stroke, peripheral artery disease, or heart failure, with events occurring earlier and more severely than in non-diabetics. Globally, T2D triples ASCVD mortality risk.

  ASCVD is when gunk builds up in arteries, setting up heart attacks, strokes, or bad leg flow – LDL over years is the big driver, plus high BP, sugar issues, smokes, getting older. To head it off, calculators predict your 10-year odds and push statins if you're mid-to-high risk, diet tweaks, moving more. Diabetics over 40 pretty much get statins standard, ramp up intensity if kidneys are leaking protein; tossing in GLP-1s or SGLT2s really cuts events for the risky crowd.
- Reviewed the structure of Med1–Med14: each has a date, GLP1 flag (TRUE/FALSE), SGLT2 flag (TRUE/FALSE), and LinkedCv flag
- Confirmed that data entry is done manually by clinical staff — so errors and inconsistencies are expected

problem

Problem Statement

Atherosclerotic cardiovascular disease (ASCVD) and Type 2 diabetes (T2D) form a high-risk combination, carrying an elevated burden of cardiovascular events. Currently, there is specific emphasis on two drug classes that have demonstrated significant benefits: GLP-1 receptor agonists (GLP-1RAs) and SGLT2 inhibitors (SGLT2is).

GLP-1RA's have demonstrated reductions in major cardiovascular events, weight loss, and glycemic control improvements. SGLT2 inhibitors have been helpful in the reductions in heart failure hospitalization, cardiovascular death, and renal decline. The combination of both agents is increasingly recognized as a strong therapy in patients with concurrent ASCVD and T2D. Despite this evidence, real world uptake of these medications remains suboptimal.

Contains longitudinal medication record for each patient, captured across up to 14 discrete medication documentation events (Med1 through Med14). Each medication record contains a date and binary flags indicating whether the patient was taking a GLP-1RA, an SGLT2i, and whether the visit was linked to a cardiovascular event.

However, these summary flags present a significant data integrity challenge. Because the registry is maintained through manual data entry by clinical staff, discrepancies can arise between the summary flags and the underlying Med1–Med14 records. A summary flag may be set based on a recent appointment that does not correspond to the intended clinical timepoint. For example, the "current" medication flag may reflect a medication list from a date well after the 12-week EST, or a patient's intake flag may have been updated retroactively. Furthermore, patients who undergo multiple exercise stress tests introduce ambiguity into which assessment constitutes the "first" 12-week EST.

These issues create a fundamental tension: the summary flags are convenient for quick analysis, but they may not accurately reflect medication status at the specific clinical timepoints required for rigorous research. What was needed was a method to independently derive medication status at each timepoint directly from the longitudinal Med1–Med14 records, with transparent quality controls and a full audit trail.

Prior to this work, there was no standardized, automated method to determine a patient's GLP-1RA or SGLT2i status at a specific clinical date using the underlying medication records. People relied on the pre computed summary flags, which, as described above, are subject to human error and may not correspond to the intended timepoints. This created 3 sub problems:

1. no reliable variable capturing medication status at the time of the initial exercise stress test the true baseline for cardiac rehabilitation. The existing intake flag (isGlpq1Intake) is based on a medication list that may not align precisely with the EST date.
2. no mechanism to ensure that the 12-week medication outcome variable reflected the patient's *first* 12-week EST. Patients who underwent repeat assessments, with Date12WeekEst falling more than 150 days after the initial EST, could have their outcome contaminated by later clinical encounters.
3. for patients who did not complete the 12-week program, there was no variable capturing their *most recent* medication status at the time they left. The existing current medication flag reflects the most recent data query date, not the separation date, and therefore does not represent the patient's medication status when they actually exited the program.

Methods

Data was extracted from the cardiac rehab dataset, which is maintained as an Excel workbook on a institutional server. The dataset included all enrolled patients with confirmed ASCVD and Type 2 diabetes. Patient records included demographics (age, sex), clinical characteristics (BMI, lipid panel, blood pressure, peak MET, PHQ-8, GAD-7), program participation data (referral date, initial EST date, 12-week EST date, program status, separation status and date), and longitudinal medication records (Med1 through Med14), each containing a documentation date and binary flags for GLP-1RA use, SGLT2i use, and cardiovascular linkage.

Work was done with Python, VS code x Jupyter Notebook environment.

Variables

Variable 1 — Medication Status at Initial Exercise Stress Test (Baseline)

This variable captures whether a patient was taking a GLP-1RA or SGLT2i at the time of their initial exercise stress test. This represents the true baseline

medication status upon cardiac rehabilitation entry. The target date for matching was the value in DateInitialEST.

Variable 2 — Medication Status at First 12-Week Exercise Stress Test (Outcome)

This variable captures medication status at the patient's *first* 12-week exercise stress test and is intended as the primary outcome variable for program completers. To ensure that only the first 12-week EST was used, a filter was applied.

> Date12WeekEst was only accepted if it fell between 60 and 150 days after DateInitialEST.
>
> Patients whose Date12WeekEst was greater than 150 days from the initial EST were flagged as likely having undergone another assessment and were excluded from this variable to accomplish this variables goals.
>
> This filter identified and excluded 39 patients (13.4% of those with a recorded 12-week EST).

Variable 3 — Medication Status at Program Separation (Dropouts)

This variable uses the most recent medication status for patients who left the program before completing the 12-week assessment. It applies to patients who have a recorded SeparationDate without a valid 12-week EST. This variable identifies the *most recent*medication record on or before the separation date. All of this was done to capture the last known medication status before the patient left the program, regardless of how far back that record falls.

Priority 1 — Record on or before target date (preferred): scans all Med1–Med14 records and identifies those dated on or before the target date. Among these, it selects the one closest to the target date. If this record falls within 7 days of the target, the match is assigned **high** confidence. If it falls between 8 and 30 days prior, the match is assigned **medium** confidence.

Priority 2 — Record within 7 days after target date (grace window): If no medication record exists on or before the target date, the algorithm checks for records falling within 7 days after the target. If found, the match is used but flagged with **low_post_date** confidence. This accounts for real-world

documentation delays while maintaining transparency about the temporal relationship.

No match: If no medication record exists within the acceptable window, the variable is set to missing (none) and the confidence flag indicates the reason, either no medication records exist for the patient at all (no_med_records), or records exist but none fall within a suitable timeframe (no_suitable_record).

For variable 3 specifically, a different method was used. Instead of finding the closest record to the target date, it identifies the *most recent* (latest-dated) record on or before the separation date. This ensures the variable reflects the patient's last documented medication status before exiting the program.

5. Confidence and Quality Flags

Each medication variable was accompanied by four metadata columns: the medication status (True/False/None), the source Med record used (e.g., Med3), the number of days between the source record and the target date, and a confidence flag. The confidence flag system provides a transparent audit trail, allowing downstream analysts to filter by data quality or conduct sensitivity analyses excluding low-confidence matches.

Confidence Level	Definition
high	Medication record within 7 days on or before target date
medium	Medication record 8–30 days before target date
medium_old_record	Medication record more than 30 days before target date (Variable 3 only)
low_post_date	Medication record used from after the target date (within 7-day grace window)
rejected_repeat_est	Date12WeekEst exceeded 150 days from Initial EST (Variable 2 only)
no_med_records	No Med1–Med14 records exist for this patient
no_suitable_record	Med records exist but none fall within the acceptable matching window

A patient lookup function was developed to enable individual-level auditing of each variable. Given a patient name or unique identifier, the function displays all three variables, the source Med record used for each, the temporal distance from

the target date, the confidence flag, and a complete chronological listing of all Med1–Med14 records. This tool was used to manually verify variable accuracy for a sample of patients across different program outcomes (completers, dropouts, and patients with low-confidence matches).

variables

Cardiac Rehab Medication Variables

Executive Summary

Three new variables have been created to track GLP-1 and SGLT2 inhibitor use at critical timepoints:

1. **Variable 1:** Baseline medication status (at Initial EST)
2. **Variable 2:** Outcome medication status (at first 12-week EST for completers)
3. **Variable 3:** Final medication status (at separation for dropouts)

Current Results:

- **550+ patients** analyzed
- **Variable 1:** Baseline data for all patients with Initial EST
- **Variable 2:** Outcome data for **251 program completers** (39 excluded as repeat ESTs)
- **Variable 3:** Dropout data for patients who separated early

Variable Tells You

Variable 1: Medications at Initial Exercise Stress Test

Purpose: What medications was the patient on when they started the cardiac rehab program? At baseline, was this patient on GLP-1 agonists or SGLT2 inhibitors?

Output Columns:

- `Var1_GLP1_AtInitialEST` → TRUE/FALSE/None (Is patient on GLP-1?)
- `Var1_SGLT2_AtInitialEST` → TRUE/FALSE/None (Is patient on SGLT2?)
- `Var1_GLP1_Source` → Which Med record was used (e.g., "Med3")
- `Var1_GLP1_DaysDiff` → Days between Med record and Initial EST
- `Var1_GLP1_Confidence` → Data quality flag

Example Interpretation:

```
Patient: John Smith
Var1_GLP1_AtInitialEST: TRUE
Var1_GLP1_Source: Med3
Var1_GLP1_DaysDiff: -2
Var1_GLP1_Confidence: high

Meaning: Patient was ON GLP-1 at program start. Data from Med
3
record dated 2 days before Initial EST. High confidence matc
h.
```

Variable 2: Medications at 12-Week EST (Completion Based)

Purpose: What medications was the patient on at their FIRST 12-week follow-up? For patients who completed the program, what was their medication status at the first 12-week assessment?

Important: This variable excludes patients whose Date12WeekEst appears to be a repeat EST (>150 days from Initial). Only the **first 12-week EST** is used to ensure true outcome measurement.

Output Columns:

- `Var2_GLP1_At12WeekEST` → TRUE/FALSE/None
- `Var2_SGLT2_At12WeekEST` → TRUE/FALSE/None
- `Var2_GLP1_Source` → Which Med record was used

- `Var2_GLP1_DaysDiff` → Days between Med record and 12-week EST
- `Var2_GLP1_Confidence` → Data quality flag

Current Results:

- **251 patients** included (first 12-week EST within 60-150 days of Initial)
- **39 patients** excluded (Date12WeekEst >150 days = likely repeat EST)

Example Interpretation:

```
Patient: John Smith
Var2_GLP1_At12WeekEST: TRUE
Var2_SGLT2_At12WeekEST: TRUE
Var2_GLP1_Confidence: high

Meaning: At 12-week follow-up, patient was on BOTH GLP-1 and
SGLT2.
If Var1 was different, this indicates medication change durin
g program.
```

Special Cases:

- If `Var2_GLP1_Confidence = "rejected_repeat_est"` → Patient has 12-week EST date but it's too far from Initial (likely a repeat test, not first assessment)
- If `Var2_GLP1_At12WeekEST = None` → Patient didn't complete 12-week assessment OR was rejected

Variable 3: Medications at Separation (Dropouts Only)

Purpose: For patients who left the program early, what were their medications at the time they separated? When dropouts left, what was their most recent medication status?

Important: This uses the **most recent** medication record before/at separation date, not the closest one. This gives us their last known medication status.

Output Columns:

- `Var3_GLP1_AtSeparation` → TRUE/FALSE/None

- `Var3_SGLT2_AtSeparation` → TRUE/FALSE/None
- `Var3_GLP1_Source` → Which Med record was used
- `Var3_GLP1_DaysDiff` → Days between Med record and separation
- `Var3_GLP1_Confidence` → Data quality flag

Applies To: Patients with SeparationDate BUT no valid 12-week EST

Example Interpretation:

```
Patient: Jane Doe
Var3_GLP1_AtSeparation: FALSE
Var3_SGLT2_AtSeparation: TRUE
Var3_GLP1_Source: Med5
Var3_GLP1_DaysDiff: -10
Var3_GLP1_Confidence: medium

Meaning: Patient dropped out. At time of separation, was on S
GLT2
but not GLP-1. Last medication record was 10 days before sepa
ration.
```

Confidence Flags

What They Mean:

Confidence Flag	Meaning	Action Needed
high	Med record within 7 days of target date	Trust this data fully
medium	Med record 8-30 days from target date	Generally reliable
medium_old_record	Med record >30 days before target	Review if critical for analysis
low_post_date	Used Med record AFTER target date (no earlier record available)	Flag for review
rejected_repeat_est	Date12WeekEst appears to be repeat test, not first	Excluded from Variable 2

Confidence Flag	Meaning	Action Needed
no_med_records	Patient has no Med1-14 documentation	Cannot determine status
no_suitable_record	Has Med records but none near target date	Cannot determine status

How to Use the Lookup Function

Check individual patients to verify data or investigate specific cases.

How to Use:

```
# In VS Code, after running all the variable creation code:

# Look up by patient name:
lookup_patient("Joe", df)

# OR Look up by ULI number:
lookup_patient(123456, df)
```

What It Shows:

The lookup function displays:

1. **Patient identification** (Name, ULI)
2. **Variable 1 details:**
    - Medication status at Initial EST
    - Which Med record was used
    - How many days from the EST
    - Confidence level
3. **Variable 2 details:**
    - Medication status at 12-week EST (if applicable)
    - Changes from baseline

- Whether patient completed program

4. **Variable 3 details:**
   - Medication status at separation (if dropout)
   - Time in program before separation
5. **All Med1-14 records:**
   - Complete medication history
   - Dates of each record
   - Shows which records were used for variables
6. **Data quality summary:**
   - Overall quality score
   - Any issues flagged

Example Output:

```
==========================================
PATIENT: Billy Bob (ULI: 12345678910)
==========================================

VARIABLE 1: Medications at Initial EST
──────────────────────────────────────

Initial EST Date: 2024-03-15

GLP-1 Status at Initial EST:
   • On GLP-1? True
   • Source: Med3
   • Days from EST: -2
   • Confidence: high

[... continues with Var2, Var3, all Med records, quality summ
ary ...]
```

1. Data Completeness Overview

Shows missing data by variable and patient-level completeness.

2. Medication Record Structure

Shows how many Med1-14 slots are being used per patient.

3. Temporal Analysis

Shows time intervals between key dates (referral, Initial EST, 12-week EST).

4. Medication Prevalence

Shows how many patients on GLP-1/SGLT2 at each timepoint.

Key Results:

- Baseline prevalence (Var1)
- Outcome prevalence (Var2 - among completers)
- Dropout prevalence (Var3)

5. Medication Changes

Shows patients who started or stopped medications during program.

Variable 2 Exclusions (39 patients):

Date12WeekEst was >150 days from Initial EST, suggesting it's a repeat test rather than the first 12-week assessment.

These 39 patients DO have a 12-week EST recorded, but it's likely their 2nd or 3rd test. Since you specifically want the FIRST 12-week EST, they were excluded to maintain data integrity.

They still have Variable 1 (baseline) data. If they later separated, they'd have Variable 3. They just don't have Variable 2 because we can't identify their TRUE first 12-week assessment.

Data Quality Philosophy:

This analysis prioritizes **accuracy over completeness**:

- Better to exclude questionable data than include inaccurate measurements
- Confidence flags allow sensitivity analyses (can test with/without low-confidence records)
- Transparent methodology supports publication and peer review

materials

Used python and Vs COde and data from TotalCardiology server.

results

Over 550 patients, all with confirmed ASCVD and Type 2 diabetes.

538 had medication data available at the time of initial exercise stress test (baseline).

290 patients had a recorded 12-week EST date; of these

251 were retained as valid first 12-week assessments (falling 60–150 days after the initial EST),

39 were excluded as likely repeat assessments.

The remaining patients who did not complete the 12-week program and had a recorded separation date contributed to the dropout cohort, with 360 patients having a most recent medication list available at time of discharge.

Among 538 patients with medication data at program entry, 28 (5.2%) were taking a GLP-1RA, 113 (21.0%) were taking an SGLT2i, and 24 (4.5%) were taking both agents. This indicates that the majority of patients entering cardiac rehabilitation with ASCVD and T2DM had not yet been initiated on GLP-1RA therapy, despite guideline recommendations, though SGLT2i use was notably more common at baseline.

Among 295 patients who attended a 12-week discharge appointment, 54 (18.3%) were taking a GLP-1RA, 215 (72.9%) were taking an SGLT2i, and 46 (15.6%) were

taking both agents. To determine whether these changes were statistically significant, chi-squared tests were performed comparing the observed medication proportions at intake versus at 12 weeks against the expected proportions under the null hypothesis of no change.

Among 360 patients with a most recent medication list at the time of discharge 64 (17.8%) were taking a GLP-1RA, 263 (73.1%) were taking an SGLT2i, and 55 (15.3%) were taking both agents. These figures are broadly consistent with the 12-week completer results, suggesting that medication optimization occurred across the program regardless of completion status.

cardiac_rehab_executive_summary.txt

<aside>

CARDIAC REHAB MEDICATION TRACKING ANALYSIS

STUDY OVERVIEW

This analysis examines medication usage patterns (specifically GLP 1 receptor agonists and SGLT2 inhibitors) among cardiac rehabilitation patients at three critical time points: baseline (Initial EST), 12 week follow up, and separation for early dropouts.

DATASET SUMMARY

Total Patients: 551
Study Period: 2024 01 03 to 2024 12 28

Patient Distribution:
- Completed 12 week program: 290 (52.6%)
- Separated early: 259 (47.0%)
- Still active/other: 2

KEY FINDINGS

1. MEDICATION PREVALENCE AT BASELINE (Initial EST)
   GLP 1 Agonists: 77 patients (15.6%)
   SGLT2 Inhibitors: 322 patients (65.2%)
   Both medications: 60 patients
   Neither medication: 155 patients

2. MEDICATION CHANGES (Initial to 12 Week)
   Patients with both timepoints: 251

   Medication Initiation (among completers):
   • Started GLP 1: 9 patients
   • Started SGLT2: 26 patients

   Medication Discontinuation:
   • Stopped GLP 1: 4 patients
   • Stopped SGLT2: 8 patients

3. DATA QUALITY ASSESSMENT
   Overall Data Quality Score: 94.0%

   Temporal Consistency: 83.7% pass rate
   Medication Flag Consistency: 13.8% average pass rate

   Confidence Distribution for New Variables:
   • High confidence: 789 records
   • Medium confidence: 80 records
   • Low confidence: 2 records

VARIABLES CREATED

This analysis created 24 new variables (8 for each timepoint):

Variable Set 1: Medications at Initial EST (Baseline)
• Var1_GLP1_AtInitialEST, Var1_SGLT2_AtInitialEST
• Plus source, days_diff, and confidence metrics

Variable Set 2: Medications at 12 Week EST (Outcome)
• Var2_GLP1_At12WeekEST, Var2_SGLT2_At12WeekEST
• Plus source, days_diff, and confidence metrics

Variable Set 3: Medications at Separation (Dropouts)
• Var3_GLP1_AtSeparation, Var3_SGLT2_AtSeparation
• Plus source, days_diff, and confidence metrics

METHODOLOGY

Medication status was determined by finding the closest Med1 14 record to each target date, with the following priority:

1. Records on or before the target date (preferred)

2. Records up to 7 days after target date (if no prior record exists)
3. Confidence flags indicate data quality for each match

CLINICAL IMPLICATIONS

- Mean BMI: 28.2
  - Mean Peak MET: 6.5
  - Program completion rate: 52.6%

RECOMMENDATIONS

1. High quality medication tracking achieved for 88.6% of baseline records
2. Consider prospective medication tracking at standardized timepoints
3. Data quality checks identify 410 records needing review
4. Medication changes observed in 35 patients warrant further analysis

Report Generated: 2026 01 22 13:37:11

</aside>

Results Summary

Timepoint	N	GLP-1RA	SGLT2i	Both
CR Entry (Baseline)	538	28 (5.2%)	113 (21.0%)	24 (4.5%)
12-Week EST	295	54 (18.3%)	215 (72.9%)	46 (15.6%)
Discharge	360	64 (17.8%)	263 (73.1%)	55 (15.3%)

- GLP-1RA increase (intake → 12-week): $\chi^2 = 54.2$, $p < 0.001$
- SGLT2i increase (intake → 12-week): $\chi^2 = 27.8$, $p < 0.001$
- Variable construction completed: January 22, 2026

conclusions

- Cardiac rehab is an effective setting for GLP-1RA and SGLT2i optimization in patients with ASCVD and T2D
- GLP-1RA use increased 3.5-fold during the program (5.2% → 18.3%)
- SGLT2i use increased 3.5-fold during the program (21.0% → 72.9%)
- The new variables provide a reliable, independently verified method for tracking medication status at specific clinical timepoints
- Confidence flags allow quality control and sensitivity analysis
- All 3 variables (baseline, 12-week outcome, dropout separation) completed and validated

This project addressed a major challenge in cardiac rehabilitation research: the need to reliably determine medication status at specific clinical timepoints from longitudinal, human-entered medication records. Result is the 3 time specific medication variables at initial exercise stress test, first 12-week exercise stress test, and program separation.

This work was able to derive time-specific medication variables from longitudinal records in registry-based studies. The temporal matching algorithm, confidence flagging system, and individual patient verification tool provide a model that can be adapted to other medications, clinical timepoints, or patient populations. The built-in data quality validation also serves an ongoing operational function: by independently checking medication status against the underlying records, the project can identify subtle discrepancies in the existing summary flags, supporting continuous improvement in data entry accuracy.

future implications/ next steps

Future work will include a detailed examination of the 39 patients excluded from Variable 2 due to suspected repeat exercise stress tests, as manual chart review may recover valid first 12-week assessments for some of these individuals. This will explore the highest functioning potential of the variable. Additionally, further analysis could explore predictors of medication initiation during cardiac

rehabilitation including baseline clinical characteristics, and sociodemographic factors to inform targeted intervention strategies. Eve using the data available to develop a model that can identify the most appropriate time to put a patient on a GLP1 or SGLT2 presents a significant opportunity to continue to fill gaps in our world today.

This work was able to derive time specific medication variables from longitudinal records in data collected. The temporal matching algorithm, confidence flagging system, and individual patient verification tool provide a model that can be adapted to other medications, clinical timepoints, or patient populations. The built-in data quality validation also serves an ongoing operational function: by independently checking medication status against the underlying records, the project can identify subtle discrepancies in the existing summary flags, supporting continuous improvement in data entry accuracy. This project has significant potential to work in the scheme of cardiac rehab, and when coupled with my wearable introduction, the sky is the limit.

The connection between the variable work and the wearables?

- The actual data and variables help understand population-level patterns: How often are patients on GLP-1s or SGLT-2is at key timepoints in rehab?
- The wearable helps us understand individual-level patterns: When your heart rate, motion, and recovery patterns look abnormal, how do you feel, and might that relate to your risk?

As next steps, the wearable and model, could be used to guide safe exercise according to intensity for patients with ASCVD and T2D. Secondly, flag when someone is over-exerting themselves, which could be important for early decompensation detection. Finally, provide a personal baseline that doctors can use to compare against future episodes.