

Identical results, hidden data problems: running CMS's digital quality measures on real hospital data

Sai Sharan Kasa

sk.sharankasa@gmail.com

Independent Researcher <https://orcid.org/0009-0002-7425-4516>

Research Article

Keywords: Quality Indicators, Health Care, Electronic Health Records, Health Level Seven, Data Accuracy, Hypoglycemia

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11394954/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author is employed by Amazon. This work was conducted independently, on personal time and with personal resources, using publicly available credentialed data; the employer had no role in the study and provided no funding, data, or resources.

Abstract

Objective

The Centers for Medicare & Medicaid Services (CMS) is moving hospital quality reporting to FHIR-based digital measures, with evidence so far mostly from synthetic test cases. This study asked whether two hospital-harm measures, severe hypoglycemia (CMS816) and severe hyperglycemia (CMS871), run over a complete hospital database without changing their results, and which data problems change them.

Materials and Methods

Under a preregistered plan, CMS's draft FHIR specifications were run with fqm-execution on MIMIC-IV-on-FHIR (431,231 stays, 2008–2019). A pipeline that cuts each patient's record into per-stay slices was compared stay by stay with the unmodified engine. In 88,234 stays from 2017–2019, measure-derived checks were applied and both measures rerun under alternative mappings and scenarios.

Results

The pipeline matched the unmodified engine on all 1,994 stays compared (discordance upper bound 0.19%); whole-record evaluation needed at least 98 times more CMS816 engine time. All 431,231 stays ran in 5.6 hours for US\$4.52. Medication-mapping choices moved up to 4.3% of CMS816's denominator while its rate changed under 0.1 points. In exploratory analyses, bedside glucose results without a unit hid 57 of 130 severe hypoglycemic stays (CMS816 0.70% vs 1.24% with mg/dL assigned; difference 0.54 points, 95% CI 0.41–0.68), and counting observation stays as inpatient doubled CMS816's denominator and events alike.

Discussion

With identical execution, every change came from the data. Checks built from the measure logic found these problems; profile validation did not.

Conclusion

What these measures report depends on data readiness, which should be tested before FHIR reporting becomes mandatory.

BACKGROUND AND SIGNIFICANCE

CMS plans to move every quality measure in its reporting programs to digital quality measures (dQMs), written in Clinical Quality Language (CQL) and run on HL7 FHIR data.[1, 2] Two hospital-harm measures show what is at stake. CMS816 counts inpatient stays in which a patient who received a glucose-lowering drug then had a glucose result below 40 mg/dL. CMS871 counts hospital days with severe hyperglycemia. Both became mandatory electronic clinical quality measures (eQMs) in the Hospital Inpatient Quality Reporting program from calendar

year 2026,[3] and draft FHIR versions were released for public comment in January 2026.[4] From the fiscal year 2028 payment determination, CMS also scores how closely hospitals' eCQM data match the medical record.[5]

Early electronically reported measures agreed with manual review with sensitivities between 46% and 98%.[6] Switching the medication data source of an adverse drug event measure changed a hospital's national ranking,[7] adding health information exchange data changed 15% of eCQM calculations,[8] and value-set and mapping choices changed who was counted even when rates hardly moved.[9, 10] For CMS816 and CMS871, the endorsement records report reliability and correlation-based validity but no testing of individual data elements.[11, 12] The closest accuracy study examined an earlier hypoglycemia measure,[13] and the public evidence for the FHIR versions consists of synthetic test cases.[14] A national hospital association asked CMS for "broad, real-world testing ... to confirm feasibility before advancing to mandatory implementation",[15] and NCQA expects discrepancies to arise from how data are "mapped into standard data models (e.g., FHIR)".[16]

Two things have kept such testing off real data. The first is computation. Running CQL over a patient's whole record is slow, and published large-scale CQL and FHIR analytics either use synthetic data or do not run quality measures.[17–19] On the open MIMIC-IV-on-FHIR demonstration data, CMS's measure logic took 51 minutes for five long-stay patients. The second is data readiness. Established data-quality frameworks ask whether records conform to their model, are complete and are plausible.[20, 21] A record can pass all of that and still be useless to a particular measure, for example a glucose result whose unit the measure cannot compare, or a dose recorded as completed although it was never given.

The two measures react to missing data in opposite ways, which makes them useful probes. Missing data almost always remove CMS816 events. CMS871 counts a day without any glucose result as hyperglycemic if the two days before it had results of at least 200 mg/dL,[22] so lost glucose records can add CMS871 event days.

The question behind this study is whether hospitals can trust CMS's FHIR measures computed on their own data. It has two testable parts. Can the measures run over a complete hospital database without changing the engine's results? And which data problems, found from the measures' own logic, change who is counted and the rates, and by how much?

MATERIALS AND METHODS

Design and preregistration

The plan, outcomes, sample sizes and analyses were registered on the Open Science Framework before any measure was run on credentialed data.[23] The two primary outcomes were whether the scaled pipeline gives identical results to the unmodified engine, and how data problems and mapping choices change the measures. Secondary outcomes were reproduction of CMS's test cases, runtime and cost, how well the readiness checks detect problems, agreement with an independent implementation, and completeness across hospitals. Seven deviations were recorded, each before the results it affects, and four of them are analyses dropped for time (Supplementary Table S1). Analyses added after registration are labeled exploratory.

Data

MIMIC-IV-on-FHIR version 2.1 holds the de-identified records of patients at Beth Israel Deaconess Medical Center in Boston from 2008 to 2019, converted to FHIR R4 from MIMIC-IV 2.2 and MIMIC-IV-ED 2.2.[24–28] It contains 299,712 patients and 431,231 hospital stays and, by design, keeps local codes for laboratory items, medications and diagnoses.[24] Dates are shifted for each patient, so each stay’s real years were recovered by linking its admission identifier to MIMIC-IV 3.1;[29] 431,044 stays (99.96%) linked. Rates and the readiness analyses used the 88,234 stays (52,022 patients) whose possible years all fall in 2017–2019, when the electronic medication administration record covered 98% of adult admissions. Each stay was evaluated in the one-year measurement period containing its discharge. A registered check counted the same source patterns in 2020–2022 admissions (Supplementary Methods 1).

Measures and execution engine

The FHIR specifications CMS816FHIRHHHypo and CMS871FHIRHHHyper (QI-Core 6.0.0) came from CMS’s public measure-content repository at a fixed commit,[14] with its bundled value-set expansions (Value Set Authority Center manifest of 17 November 2025). They were run with fqm-execution 1.8.5 (cql-execution 3.3.2, cql-exec-fhir 2.2.0),[30] the open-source engine that CMS’s measure authoring tool uses to run test cases for QI-Core measures, [31] with default options. Each package lacks one value set it declares, without which the engine stops; both came from the Value Set Authority Center (Supplementary Methods). CMS871 also declares its aggregation as “Sum” where the engine expects “sum”, and its results were read from the engine’s detailed output, not its summary report. With these in place, the engine reproduced CMS’s expected results for all 28 CMS816 and 26 CMS871 test cases.

Mapping the data to the measures

The baseline mapping followed one rule: keep values as exported and add only the structure the measures need to read them (Supplementary Methods). Hospital encounters received a SNOMED CT type from their MIMIC class, with emergency, ambulatory and same-day classes typed as inpatient admissions and the observation class as observation services. Laboratory and bedside glucose results received LOINC codes, and results recorded in mg/dL were given the UCUM unit system. Encounter diagnoses received dotted ICD-10-CM codes, a link from their encounter, an onset at admission and an active status. Without that status CMS’s shared library ignores them, as the independent implementation revealed. Every resource the measures retrieve was labelled with its QI-Core profile.

Medication administrations were linked to the National Drug Code (NDC) of their prescription, mapped to RxNorm (release of 5 October 2026) and moved from brand to generic concepts, because both medication value sets list only generic drugs and packs. Administrations without a usable NDC, and ICU infusion items, were mapped by review. I also ran the registered recall audit, searching the excluded codes for glucose-lowering drugs in three widening rounds (Supplementary Methods).

Four registered medication mappings were compared. The baseline (a) is NDC mapping plus the reviewed items. A conservative version (b) uses only single-product NDC matches, and a liberal version (c) adds mappings marked as possible during review. In a model-proposed version (d), an open-weight language model (gpt-oss-120b, run in the author’s own cloud account with invocation logging off) chose among RxNorm candidates for each local code.

Every code it proposed was checked against RxNorm and the value sets. No model takes part in running the measures.

Running the measures at scale

The pipeline changes what the engine receives, never the engine itself. A filter built from the measures' data requirements keeps only the resource types and codes the measures can retrieve. Each stay then gets its own slice: the patient's filtered resources whose times overlap a window from 24 hours before the hospitalization starts (including an emergency or observation visit just before it) to one hour after discharge. The slice also holds the patient, the patient's conditions, records without times, and shared locations and organizations. Workers on a Kubernetes cluster (Amazon EKS, m7i.large nodes) each evaluate a batch of stays, write each stay's result exactly once, and pick up where they left off after an interruption. Each result also records the size of every named expression in the measure, so any change can be traced to the part of the measure that produced it.

Testing for identical results

The registered hypothesis was that the pipeline gives exactly the same initial population, denominator, exclusion and numerator for every stay, and the same CMS871 day counts, as the unmodified engine run on each patient's complete record for the year. The sample was 2,000 stays: 1,500 drawn at random and 125 from each of four situations where slicing could go wrong, namely the heaviest measure workload, emergency or observation time just before admission, readmission within 24 hours, and discharge at a year boundary. The run was repeated after deliberately killing two worker pods and one node. A registered stopping rule ended whole-record evaluations still running when the cluster reached its cost cap (Supplementary Table S1). The upper bound on the discordance rate is a 95% Wilson interval.[32]

An independent implementation

Both measures were also written in Python straight from CMS's human-readable specifications, without a CQL engine. In the four places where the text and the logic disagree (Supplementary Methods), the logic was followed. Its results were compared stay by stay with the engine's, on whole records for the 2,000-stay sample and on slices for the 2017–2019 cohort.

Runtime and cost

The pipeline was timed on the 2,000-stay sample five times on newly started nodes and five times on nodes already in use, then on 1, 2, 4 and 8 nodes, and once on all 431,231 stays.[33] Cost is node-hours at the on-demand price plus query charges, with the managed-cluster fee given separately. The full run and the readiness runs evaluated several stays per engine call; the timing runs used one.

Readiness checks

The checks come from what the measures require. Some come from the data requirements fqm-execution generates for each measure, such as resource type, profile, value set and status, and others from what the logic reads beyond them, such as units, timing and links between records. Each check is tied to the expression that

uses the data (Table 1). The checks were applied to every record the measures use in the cohort, as exported and after mapping. To test detection, seven kinds of problem were injected into 1%, 10% and 50% of eligible records with fixed, nested random seeds: units converted to mmol/L, profiles removed, standard codes removed, generic drug codes replaced by brand codes, diagnosis links removed, status set to unknown, and timing removed. Recall and precision of the matching check were computed for each.

Table 1

Readiness checks derived from the measures: what each requires, where the logic uses it, and what happens to the measure when it fails.

Sources: data requirements generated by fqm-execution from each measure, and the elements the CQL reads beyond them (Supplementary Methods 10).

Check	Requirement	Used by	If it fails
QI-Core profile declared	The record declares the QI-Core profile the measure asks for	Every record the measure asks for (the engine selects by declared profile)	Record invisible; no error unless the Patient lacks its profile
Code in the measure's value set	A standard code in the role's value set (LOINC, RxNorm, ICD-10-CM; encounters: a SNOMED CT type)	The measure's code filters	Record invisible
Drug code at the value set's term level	RxNorm code of a term type the value set lists (generic clinical drug or pack)	Medication value sets (exact code match)	Administration invisible: fewer CMS816 denominator and CMS871 population stays
Glucose value comparable in mg/dL	A number with a standard (UCUM) mass-per-volume unit	CMS816: result < 40 (and > 80 after it); CMS871: ≥ 200, > 300, > 600	Comparison is null: events and hyperglycemic days lost while the stay stays counted
Status the measure accepts	Results final, amended or corrected; administrations completed (CMS871 also in progress); diagnoses active or abated	Data-requirement status filters; prevalence period of diagnoses	Record dropped
Time element present	Result time; administration start (CMS816) or closed period (CMS871); encounter period; diagnosis onset	Hospitalization windows; 24-hour lookback from a dose	Record falls outside every window
Patient reference	The record points to its patient	Patient-level retrieval	Record invisible
Encounter links its diabetes diagnoses	The encounter points to the diagnosis	CMS871: Encounter with Existing Diabetes Diagnosis	Stay leaves CMS871 unless it qualifies otherwise (only event-free stays can leave)
Administration status confirmed by an eMAR event (MIMIC)	A recorded administration event behind a completed status	CMS816: Hypoglycemic Medication Administration; CMS871: Encounter with Hypoglycemic Medication	Status unverifiable
Status agrees with the eMAR event (MIMIC)	Completed only if the dose was given	As above	A not-given dose counts: stays added to CMS816's denominator
Stay typed as observation (information)	Encounter type in the Encounter Inpatient value set	Both measures: Qualifying Encounter	Stay outside both measures

Effect of data problems and mapping choices

Both measures were rerun under each mapping variant and each scenario and compared with the baseline. The registered scenarios corrected medication status from the recorded administration event, as lower and upper bounds because some administrations have no event. They also applied each problem at its full extent as exported (units, profiles, codes, diagnosis links), used laboratory glucose only, and injected the two problems that touch few stays: diagnosis links, and drug codes at the wrong term level. Two exploratory scenarios were added after the first readiness results: observation-class stays typed as inpatient admissions, and bedside glucose results without a unit given mg/dL, the unit of their item in MIMIC's dictionary. The engine is deterministic, so only stays whose slice changed were rerun, and every other stay kept its baseline result.

CMS816 is a proportion (stays with a severe hypoglycemic event per denominator stay) and CMS871 a ratio (hyperglycemic days per 100 eligible days). Confidence intervals come from a stay-level bootstrap with 2,000 resamples, paired by stay with Poisson weights,[34] except Wilson intervals for CMS816 rates.[32] Changes in who is counted are reported as stays entering and leaving each population, together with the parts of the measure that changed in those stays. Counts from 1 to 9 are shown as < 10.

Ethics, data use and AI use

All data are de-identified and used under PhysioNet data use agreements. Beth Israel Deaconess Medical Center's review board approved sharing MIMIC-IV and waived consent;[27] eICU meets the HIPAA Safe Harbor standard. [35] Only aggregate counts left the author's cloud account. AI assistance with writing and code is described in the Acknowledgments.

RESULTS

The data as exported

Of the 431,231 hospital stays, 88,234 formed the 2017–2019 cohort (Figure 2). As exported, no stay could enter either measure. No hospital encounter had a type from CMS's inpatient value set, laboratory, medication and diagnosis codes were all local, and no resource declared a QI-Core profile (Figure 4). After mapping, 166,151 stays overall (38.5%) and 53,762 cohort stays (60.9%) were observation encounters, which neither measure counts.

Identical results at scale

For all 1,994 stays that could be compared, the scaled pipeline and the unmodified engine agreed on every population and every CMS871 day count, in each of the five sample groups (0 discordant; upper bound 0.19%; Table 2). Fewer than 10 stays per measure were stopped at the cost cap before a whole-record result. Results were again identical for all 2,000 stays after two pods and a node were killed mid-run, with each stay written exactly once, and when several stays were evaluated per engine call. The independent implementation agreed with the engine on all 1,994 stays and on all 88,234 stays of the cohort.

Table 2. Stay-level agreement. A stay is identical only if every population (initial population, denominator, exclusions, numerator) and, for CMS871, both per-stay observations (eligible days and hyperglycemic days) match. Upper bound: upper limit of the 95% Wilson interval for the discordance rate (0 discordant). Fewer than 10 stays per measure had no whole-record result (evaluations stopped at the registered cost cap).

Comparison	Stays	Compared	Identical, CMS816	Identical, CMS871	Discordant	95% upper bound, %
Scaled pipeline vs unmodified engine on whole records	Random sample	1,497	1,497	1,497	0	0.26
	Highest measure workload	124	124	124	0	3.00
	Emergency or observation time just before admission	123	123	123	0	3.03
	Readmission within 24 hours	125	125	125	0	2.98
	Discharge at a year boundary	125	125	125	0	2.98
	All	1,994	1,994	1,994	0	0.19
Fault test (2 pods and 1 node killed) vs undisturbed run	All	2,000	2,000	2,000	0	0.19
Several stays per engine call vs one	All	2,000	2,000	2,000	0	0.19
Independent implementation vs engine, whole records	2,000-stay sample	1,994	1,994	1,994	0	0.19
Independent implementation vs engine, slices	2017–2019 cohort	88,234	88,234	88,234	0	0.0044

Across the compared stays, CMS816 needed 98.7 times more engine time on whole records than on slices (1,432,304 against 14,518 seconds), a lower bound because the stopped evaluations are left out, and CMS871 needed 7.5 times more. The slowest stay took 35.3 hours on its whole record and 12 minutes as a slice (Figure 6).

Runtime and cost

On the 2,000-stay sample with 8 nodes, runs on fresh nodes took 24.7 minutes on average (95% CI 23.7–25.6; coefficient of variation 3.2%) and runs on warm nodes 25.7 minutes (3.8%). Wall time fell in step with the number of nodes, from 201 minutes on one node to 25 on eight, 8.2 times faster, while the node cost stayed near US\$0.33 per run (Figure 6). All 431,231 stays ran in 5.6 hours on 8 nodes, using 44.8 node-hours that cost US\$4.52 at on-demand prices, about 1 cent per 1,000 stays. The cluster fee added US\$0.56 and the 8-minute preparation in Athena US\$0.34. The median stay took 0.11 seconds of engine time for CMS816 and 0.14 for CMS871, but the slowest 1% of stays used 59% of CMS816’s engine time.

Problems the checks found

Three problems remained after mapping (Figure 4; Supplementary Table S2). The first was glucose units. Of 824,809 glucose results, 211,894 (25.7%) could not be compared with the measures' mg/dL thresholds, and nearly all of them (211,672) were bedside readings with a number but no unit. The gap is in MIMIC-IV's own records, where 85% of these bedside values carry their unit only in the item dictionary, and 82% in 2020–2022 (Supplementary Table S5). The second was medication status. Of 450,394 electronic administration records for CMS816 drugs, 32,337 (7.2%) had no administration event to back their “completed” status, and 2,032 were marked completed although their recorded event said otherwise. The third was timing. In 60,890 of about 60,910 cohort stays with a linked emergency visit, the visit ended after the recorded admission time, so results and doses from the emergency department fell outside both measures' windows. The same holds for 96% of 2020–2022 admissions with an emergency visit.

The checks caught every injected problem. Recall was 1.0 for all seven kinds at every rate, and precision was 1.0 except for units, where the 211,894 results that already failed counted against it (Supplementary Table S3).

Effect on the measures

At baseline, 73 of 10,487 denominator stays had a severe hypoglycemic event (CMS816 0.70%; 95% CI 0.55–0.87), and there were 2,130 hyperglycemic days among 63,940 eligible days in 13,302 stays (CMS871 3.33 per 100 days; 95% CI 3.13–3.53) (Table 3). For context only, hospitals reporting voluntarily in 2023 averaged 1.28% across 522 hospitals and 0.077 hyperglycemic days per eligible day across 310 hospitals.[11, 12]

Table 3. Effect of each medication mapping and data scenario on CMS816 and CMS871 in the 2017–2019 cohort (88,234 stays). Only stays whose input changed were re-evaluated; all others keep their baseline result. Entering / leaving: stays added to or removed from the population relative to baseline. CMS816 rates with Wilson 95% CIs; CMS871 rates and all differences with 95% CIs from a paired stay-level bootstrap (2,000 resamples). Counts from 1 to 9 are shown as <10.

A. CMS816, severe hypoglycemia (stays with an event per 100 denominator stays)

Variant or scenario	Stays re-run	Denominator	Denominator entering / leaving	Events	Events entering / leaving	Rate, % (95% CI)	Difference, points (95% CI)
Baseline (mapping a)	88,234 (all)	10,487		73		0.70 (0.55–0.87)	Reference
Mapping b, conservative	8,616	10,034	0 / 453	66	0 / <10	0.66 (0.52–0.84)	–0.04 (–0.09 to +0.01)
Mapping c, liberal	1,722	10,594	107 / 0	74	<10 / 0	0.70 (0.56–0.88)	0.00 (–0.01 to +0.02)
Mapping d, model-proposed	8,785	10,468	202 / 221	73	<10 / <10	0.70 (0.56–0.88)	0.00 (–0.03 to +0.03)
Medication status corrected, lower bound	6,426	10,440	0 / 47	72	0 / <10	0.69 (0.55–0.87)	–0.01 (–0.03 to 0.00)
Medication status corrected, upper bound	1,571	10,465	0 / 22	73	0 / 0	0.70 (0.56–0.88)	0.00 (0.00 to 0.00)
Brand drug codes, injected 1%	3,774	10,480	0 / <10	73	0 / 0	0.70 (0.55–0.87)	0.00 (0.00 to 0.00)
Brand drug codes, injected 10%	13,520	10,419	0 / 68	72	0 / <10	0.69 (0.55–0.87)	–0.01 (–0.03 to +0.01)
Brand drug codes, injected 50%	19,736	9,898	0 / 589	61	0 / 12	0.62 (0.48–0.79)	–0.08 (–0.15 to –0.02)
Diagnosis links removed, injected 1%	218	10,487	0 / 0	73	0 / 0	0.70 (0.55–0.87)	No stay changed
Diagnosis links removed, injected 10%	2,263	10,487	0 / 0	73	0 / 0	0.70 (0.55–0.87)	No stay changed
Diagnosis links removed, injected 50%	11,238	10,487	0 / 0	73	0 / 0	0.70 (0.55–0.87)	No stay changed
Diagnosis links as exported	22,243	10,487	0 / 0	73	0 / 0	0.70 (0.55–0.87)	No stay changed
Glucose units as exported	80,197	10,487	0 / 0	0	0 / 73	0.00 (0.00–0.04)	–0.70 (–0.87 to –0.54)
Profiles as exported,	80,735	0	0 / 10,487	0	0 / 73	No denominator stays	Not defined

results and administrations							
Codes as exported (local), results and administrations	80,735	0	0 / 10,487	0	0 / 73	No denominator stays	Not defined
Laboratory glucose only (registered sensitivity)	12,485	10,487	0 / 0	73	0 / 0	0.70 (0.55–0.87)	No stay changed
Bedside glucose unit assigned (exploratory)	12,147	10,487	0 / 0	130	57 / 0	1.24 (1.05–1.47)	+0.54 (+0.41 to +0.68)
Observation stays typed as inpatient (exploratory)	54,952	21,030	10,574 / 31	151	78 / 0	0.72 (0.61–0.84)	+0.02 (–0.09 to +0.13)

B. CMS871, severe hyperglycemia (hyperglycemic days per 100 eligible days)

Variant or scenario	Stays re-run	Population	Population entering / leaving	Stays with an event day	Entering / leaving	Event days / eligible days	Rate per 100 days (95% CI)	Difference (95% CI)
Baseline (mapping a)	88,234 (all)	13,302		1,400		2,130 / 63,940	3.33 (3.13–3.53)	Reference
Mapping b, conservative	8,616	12,982	0 / 320	1,400	0 / 0	2,130 / 62,561	3.40 (3.20–3.61)	+0.07 (+0.06 to +0.08)
Mapping c, liberal	1,722	13,368	66 / 0	1,400	0 / 0	2,130 / 64,207	3.32 (3.12–3.51)	–0.01 (–0.02 to –0.01)
Mapping d, model-proposed	8,785	13,247	150 / 205	1,400	0 / 0	2,130 / 63,720	3.34 (3.14–3.54)	+0.01 (0.00 to +0.02)
Medication status corrected, lower bound	6,426	13,288	0 / 14	1,400	0 / 0	2,130 / 63,882	3.33 (3.13–3.53)	0.00 (0.00 to +0.01)
Medication status corrected, upper bound	1,571	13,294	0 / <10	1,400	0 / 0	2,130 / 63,912	3.33 (3.13–3.53)	0.00 (0.00 to 0.00)
Brand drug codes, injected 1%	3,774	13,298	0 / <10	1,400	0 / 0	2,130 / 63,930	3.33 (3.13–3.53)	0.00 (0.00 to 0.00)
Brand drug codes, injected 10%	13,520	13,268	0 / 34	1,400	0 / 0	2,130 / 63,819	3.34 (3.13–3.53)	+0.01 (0.00 to +0.01)
Brand drug codes, injected 50%	19,736	13,062	0 / 240	1,400	0 / 0	2,130 / 62,824	3.39 (3.19–3.59)	+0.06 (+0.05 to +0.07)
Diagnosis links removed, injected 1%	218	13,291	0 / 11	1,400	0 / 0	2,130 / 63,910	3.33 (3.13–3.53)	0.00 (0.00 to 0.00)
Diagnosis links removed, injected 10%	2,263	13,169	0 / 133	1,400	0 / 0	2,130 / 63,553	3.35 (3.15–3.55)	+0.02 (+0.02 to +0.03)
Diagnosis links removed, injected 50%	11,238	12,616	0 / 686	1,400	0 / 0	2,130 / 62,055	3.43 (3.22–3.63)	+0.10 (+0.09 to +0.11)
Diagnosis links as exported	22,243	12,014	0 / 1,288	1,400	0 / 0	2,130 / 60,496	3.52 (3.31–3.72)	+0.19 (+0.17 to +0.21)
Glucose units as exported	80,197	12,171	0 / 1,131	0	0 / 1,400	0 / 58,063	0.00 (0.00–0.00)	–3.33 (–3.53 to –3.13)

Profiles as exported, results and administrations	80,735	8,675	0 / 4,627	0	0 / 1,400	0 / 37,544	0.00 (0.00–0.00)	–3.33 (–3.53 to –3.13)
Codes as exported (local), results and administrations	80,735	8,675	0 / 4,627	0	0 / 1,400	0 / 37,544	0.00 (0.00–0.00)	–3.33 (–3.53 to –3.13)
Laboratory glucose only (registered sensitivity)	12,485	13,301	0 / <10	1,416	20 / <10	2,159 / 63,899	3.38 (3.18–3.57)	+0.05 (+0.02 to +0.07)
Bedside glucose unit assigned (exploratory)	12,147	13,444	142 / 0	1,761	361 / 0	2,931 / 64,524	4.54 (4.29–4.79)	+1.21 (+1.10 to +1.33)
Observation stays typed as inpatient (exploratory)	54,952	29,597	16,317 / 22	2,633	1,242 / <10	4,055 / 113,698	3.57 (3.40–3.72)	+0.24 (+0.10 to +0.38)

The medication mappings moved stays more than rates (Figure 5). The conservative mapping removed 453 stays (4.3%) from CMS816’s denominator and 320 (2.4%) from CMS871’s population, and the model-proposed mapping swapped 423 CMS816 denominator stays in and out. Yet no CMS816 difference reached 0.1 percentage points, every interval included zero, and CMS871 moved by –0.01 to +0.07 per 100 days. The model reviewed all 1,418 medication entries in 12.5 minutes for US\$0.25. It agreed with the reviewer on 961 of 971 excluded codes, and one disagreement was an insulin product the reviewer had missed (5,725 administrations), which reopened the audit. On the 447 entries needing judgment it agreed on only 259, and its high-confidence answers alone would have kept 101 wrong mappings (Supplementary Table S6). Correcting medication status removed at most 47 CMS816 denominator stays. Replacing generic codes with brand codes in half of all administrations removed 589 denominator stays and 12 of the 73 events (difference –0.08 points; –0.15 to –0.02).

Three data problems mattered more. Giving bedside glucose results their missing unit, an exploratory scenario, added 57 severe hypoglycemic stays, so 44% of the 130 events had been hidden. CMS816 rose to 1.24% (difference 0.54 points; 95% CI 0.41–0.68) and CMS871 to 4.54 per 100 days (difference 1.21; 1.10–1.33). Diagnosis links as exported removed 1,288 stays (9.7%) from CMS871. None of them had a hyperglycemic day, because a stay with one also qualifies through its glucose results, so the rate went up by 0.19 per 100 days (0.17–0.21). Typing observation stays as inpatient, also exploratory, roughly doubled both measures. CMS816’s denominator grew from 10,487 to 21,030 stays and its events from 73 to 151, and CMS871’s population from 13,302 to 29,597 stays. CMS816’s rate hardly moved (0.72%; difference 0.02 points, –0.09 to 0.13), while CMS871 rose by 0.24 per 100 days (0.10–0.38).

In every scenario, the parts of the measure that changed were the ones each problem acts on (Supplementary Table S4). The registered glucose-source analysis isolated the bedside effect. With laboratory glucose only, CMS816 did not change at all, because bedside results as recorded had contributed no events. CMS871 rose by 0.05 per 100 days (0.02–0.07) as 20 stays gained a hyperglycemic day through the rule for days without a test. At their full extent as exported, units left CMS816 with all 10,487 stays and no events. Local codes or missing

profiles on results and administrations emptied CMS816 entirely, while CMS871 kept 8,675 stays at a rate of zero. The engine raised no error in any of these cases.

Completeness across hospitals

In the eICU Collaborative Research Database 2.0,[35, 36] which covers 200,234 adult ICU stays at 208 US hospitals in 2014–2015, recording of the measures' data varied far more than chance would allow (Supplementary Figure S1; funnel limits [37]). The median hospital recorded a bedside glucose for 61.0% of stays (interquartile range 46.6–76.5%). Of 205 hospitals with at least 10 stays, 25 recorded no bedside glucose at all and 24 recorded no insulin. Diabetes, which describes the patients rather than the record, varied least. In this database, a hospital without a given data feed contributes none of that data.[36]

DISCUSSION

CMS's FHIR hospital-harm measures ran over every stay of a large hospital database without a single stay-level result changing, and what did change the results came from the data. The biggest effect came from a small detail of how bedside glucose was recorded: without a unit, those readings hid 44% of severe hypoglycemic events.

Why identical results matter

The identical results do two jobs. They make the pipeline a control: it reproduced a deterministic engine exactly, so every difference between scenarios in this study comes from the data or the mapping and none from how the measures were run. They also make cheap, hospital-scale testing trustworthy. Whole-record evaluation of the 2,000-stay sample alone took more than 400 CPU-hours, while each scenario here reran only the stays whose data changed. Identical results are not the same as valid results, though. A faithful execution carries the measure's own assumptions with it,[38] and chart review is still needed to validate the measures.

Each problem pushes in a predictable direction

Each problem worked through specific parts of the measures, in the direction the logic predicts. A glucose result without a usable unit can never be compared with 40 or 300 mg/dL, so it removes CMS816 events and CMS871 hyperglycemic days while the stay still counts. The FHIR specification itself notes that CMS816 "is currently confined to using mg/dL".[39] A dose recorded as completed but never given adds a stay to CMS816's denominator. A stay typed as an observation encounter drops out of both measures. Missing data can also raise a harm rate. A lost diagnosis link can only remove CMS871 stays without hyperglycemic days, and removing bedside tests created hyperglycemic days through the rule for days without a test. Because the direction follows from the logic, readiness checks can be generated from the measure itself and ranked by their likely effect, instead of being written by hand for every data source. Fixes can be as cheap: one rule giving each bedside item its dictionary unit made all 57 hidden events count.

Stable rates can hide different patients

Rates turned out to be a poor guide to readiness. The mapping variants and observation typing changed who was counted, by up to 4.3% and by a factor of two, while CMS816's rate moved by less than 0.1 points. Value-set choices did the same to a statin measure.[9] A hospital that compares its FHIR rate with its current eCQM rate

could see agreement while counting different patients, which is why readiness is better judged stay by stay. A language model was a useful second reader of medication codes, but its confident answers still needed review.

Profile validation does not catch these problems

None of the problems that changed the results breaks a FHIR profile. Records from the open demonstration data with a bedside glucose but no unit, an encounter without links to its diagnoses, and an observation-class encounter all passed the HL7 FHIR validator against QI-Core 6.0.0 without errors (Supplementary Methods). No profile can tell a dose that was given from one that was not. The one problem that profiles do govern, the declared profile itself, caused trouble of a different kind. When the patient record lacked its QI-Core profile the engine stopped with an error, but when other resources lacked it the measures quietly returned zeros; at full scale, CMS871 reported a rate of zero over 8,675 stays. This matters because CMS's authoring guidance is moving its dQMs from QI-Core to US Quality Core on US Core 6.1.0,[40] and the profiles declared in hospitals' data will change with it.

What this means for the FHIR transition

A national hospital association asked CMS for real-world testing before FHIR reporting becomes mandatory.[15] This study is one such test, on one center's data. CMS's accuracy validation compares a hospital's submitted eCQM data with its own medical record.[5] That comparison may not reveal a change caused by the data model or the mapping, which is exactly what a move from QDM to FHIR introduces. And measure packages should be tested with complete value sets: as published, the CMS871 package cannot be computed without an expansion from elsewhere, a defect also reported for CMS816's package.[41]

Limitations

The data come from one academic center through a research conversion.[24] Figures such as the share of observation encounters describe this center and this conversion rather than US hospitals, although they persisted into 2020–2022. Identical results do not depend on the data, and the checks are released so that other sites can run them on their own FHIR data. Dates are shifted, so measurement periods are built per stay. Without chart review, the baseline rates are not checked against clinical truth. The FHIR specifications are still drafts. The medication audit was mine, not a pharmacist's; its remaining miss rate is bounded in the Supplement. Whole-record evaluations of fewer than 10 stays per measure were stopped under the registered cost rule.

CONCLUSION

CMS's FHIR hospital-harm measures can be run over a complete hospital database with results identical to the unmodified engine, for about 1 cent of computing per 1,000 stays. What they report depends on the data. Glucose units, diagnosis links and encounter typing changed the results in directions the measure logic predicts, medication mapping changed who was counted far more than the rates, and profile checks caught none of it. Checks built from the measures themselves, released with this study, should be run before FHIR reporting becomes mandatory.

Declarations

DATA AVAILABILITY

MIMIC-IV-on-FHIR 2.1, MIMIC-IV 2.2, MIMIC-IV-ED 2.2 and MIMIC-IV 3.1 are available to credentialed users from PhysioNet under their data use agreements.[24-29, 42] The measure specifications and test cases are public.[14] Value-set expansions require a free UMLS license. Aggregate results (counts with 1–9 suppressed) are available in the repository below.

CODE AVAILABILITY

All code needed to reproduce the analyses, including the slicing SQL, the worker, the readiness checks and scenarios, the independent implementation and the analysis scripts, is available under the Apache-2.0 license at <https://github.com/sharankasa/cms-dqm-readiness>, archived at Zenodo (doi:10.5281/zenodo.23406902). No data are included.

FUNDING

None. The study was self-funded; cloud computing was paid personally, and the design does not depend on services specific to one cloud provider.

COMPETING INTERESTS

The author is employed by Amazon. This work was conducted independently, on personal time and with personal resources, using publicly available credentialed data; the employer had no role in the study and provided no funding, data, or resources.

AUTHOR CONTRIBUTIONS

Sai Sharan Kasa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration.

ACKNOWLEDGMENTS

The author used Claude (Anthropic; Claude Opus 5.5) as an assistant for writing and debugging analysis code and for editing the manuscript text. The author designed the study, reviewed and verified all code, numbers and text, and takes full responsibility for the content. The language model that proposed medication mappings for mapping variant (d) is described in Methods.

References

1. Centers for Medicare & Medicaid Services. About dQMs. eCQI Resource Center. <https://ecqi.healthit.gov/dqm/about-dqms> (accessed 15 October 2026).
2. Centers for Medicare & Medicaid Services. Digital Quality Measurement Strategic Roadmap. March 2022. https://ecqi.healthit.gov/sites/default/files/CMSdQMStrategicRoadmap_032822.pdf (accessed 15 October 2026).

3. Centers for Medicare & Medicaid Services. Medicare and Medicaid programs and the Children's Health Insurance Program; hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and policy changes and fiscal year 2025 rates; quality programs requirements; and other policy changes. Final rule. Fed Regist 2024;89:68986. <https://www.federalregister.gov/documents/2024/08/28/2024-17021> (accessed 15 October 2026).
4. Centers for Medicare & Medicaid Services. Draft dQMs. eCQI Resource Center. <https://ecqi.healthit.gov/dqm/draft-dqms> (accessed 15 October 2026).
5. CFR § 412.140(d)(2)(ii)(B). <https://www.law.cornell.edu/cfr/text/42/412.140> (accessed 15 October 2026).
6. Kern LM, Malhotra S, Barrón Y, et al. Accuracy of electronically reported "meaningful use" clinical quality measures: a cross-sectional study. *Ann Intern Med* 2013;158(2):77–83. doi:10.7326/0003-4819-158-2-201301150-00001
7. Kahn MG, Ranade D. The impact of electronic medical records data sources on an adverse drug event quality measure. *J Am Med Inform Assoc* 2010;17(2):185–91. doi:10.1136/jamia.2009.002451
8. D'Amore JD, McCrary LK, Denson J, et al. Clinical data sharing improves quality measurement and patient safety. *J Am Med Inform Assoc* 2021;28(7):1534–42. doi:10.1093/jamia/ocab039
9. Cholan RA, Weiskopf NG, Rhoton DL, et al. From concepts and codes to healthcare quality measurement: understanding variations in value set vocabularies for a statin therapy clinical quality measure. *EGEMS (Wash DC)* 2017;5(1):19. doi:10.5334/egems.212
10. Hripcsak G, Levine ME, Shang N, et al. Effect of vocabulary mapping for conditions on phenotype cohorts. *J Am Med Inform Assoc* 2018;25(12):1618–25. doi:10.1093/jamia/ocy124
11. Partnership for Quality Measurement. Rate of severe hypoglycemia among hospitalized patients (Hospital Harm – Severe Hypoglycemia, CBE 3503e). <https://p4qm.org/measures/3503e> (accessed 15 October 2026).
12. Partnership for Quality Measurement. Hospital Harm – Severe Hyperglycemia (CBE 3533e). <https://p4qm.org/measures/3533e> (accessed 15 October 2026).
13. Santos CAQ, Conover C, Shehab N, et al. Electronic measurement of a clinical quality measure for inpatient hypoglycemic events: a multicenter validation study. *Med Care* 2020;58(10):927–33. doi:10.1097/MLR.0000000000001398
14. cqframework. dqm-content-qicore-2025: dQM measure content using QI-Core 6.0.0, commit 5f840c8 (10 September 2026). GitHub. <https://github.com/cqframework/dqm-content-qicore-2025/tree/5f840c8792ab3d0c026fc17d07847d782d34e482> (accessed 15 October 2026).
15. Federation of American Hospitals. Letter to the CMS Administrator re: Fast Healthcare Interoperability Resources (FHIR) digital quality measures (dQMs). 25 February 2026. <https://fah.org/wp-content/uploads/2026/02/FAH-FHIR-Digital-Quality-Measure-Transition-Comments.pdf> (accessed 15 October 2026).
16. National Committee for Quality Assurance. NCQA shares new guidance for digital HEDIS reporting. 1 November 2024 (updated 2026). <https://www.ncqa.org/blog/ncqa-shares-new-guidance-for-digital-hedis-reporting/> (accessed 15 October 2026).
17. Grimes J, Brush R, Rhyzhikov N, et al. SQL on FHIR: tabular views of FHIR data using FHIRPath. *NPJ Digit Med* 2025;8:342. doi:10.1038/s41746-025-01708-w
18. Gulden C, Kampf M, Kraska D, et al. Benchmarking Fast Healthcare Interoperability Resources–based analytics: quantitative study of RESTful server queries and big data engines. *JMIR Med Inform*

2026;14:e82924. doi:10.2196/82924

19. Kastroulis A, Bonfini P, Litsas A. Introducing Hermes: executing Clinical Quality Language (CQL) at over 66 million resources per second (inexpensively). arXiv:2212.06043 [preprint]. 2022. <https://arxiv.org/abs/2212.06043> (accessed 15 October 2026).
20. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244. doi:10.13063/2327-9214.1244
21. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(1):144–51. doi:10.1136/amiajnl-2011-000681
22. Centers for Medicare & Medicaid Services. CMS871 v5.1.000, Hospital Harm – Severe Hyperglycemia. <https://ecqi.healthit.gov/sites/default/files/ecqm/measures/CMS871-v5.1.000-QDM.html> (accessed 15 October 2026).
23. Kasa SS. Running CMS’s digital quality measures at hospital scale: correctness-preserving execution and measure-derived data-readiness checks, evidence from MIMIC-IV and eICU [preregistration]. Open Science Framework; registered 8 October 2026. doi:10.17605/OSF.IO/JHKVT
24. Bennett AM, Ulrich H, van Damme P, et al. MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format. J Am Med Inform Assoc 2023;30(4):718–25. doi:10.1093/jamia/ocad002
25. [dataset] Bennett A, Wiedekopf J, Ulrich H, et al. 2024. MIMIC-IV on FHIR (version 2.1). PhysioNet. doi:10.13026/rrj1-ny66
26. [dataset] Johnson A, Bulgarelli L, Pollard T, et al. 2023. MIMIC-IV (version 2.2). PhysioNet. doi:10.13026/6mm1-ek67
27. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1. doi:10.1038/s41597-022-01899-x
28. [dataset] Johnson A, Bulgarelli L, Pollard T, et al. 2023. MIMIC-IV-ED (version 2.2). PhysioNet. doi:10.13026/5ntk-km72
29. [dataset] Johnson A, Bulgarelli L, Pollard T, et al. 2024. MIMIC-IV (version 3.1). PhysioNet. doi:10.13026/kpb9-mt58
30. The MITRE Corporation. fqm-execution (version 1.8.5) [software]. Apache-2.0. <https://github.com/projecttacoma/fqm-execution>
31. Centers for Medicare & Medicaid Services. MADiE publicly accessible release schedule, release 3.0.7 (9 July 2026). <https://www.emeasuretool.cms.gov/sites/default/files/2026-07/Publicly%20Accessible%20Release%20Schedule%203.0.7.pdf> (accessed 15 October 2026).
32. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22(158):209–12. doi:10.1080/01621459.1927.10502953
33. Georges A, Buytaert D, Eeckhout L. Statistically rigorous Java performance evaluation. In: Proceedings of the 22nd Annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages and Applications (OOPSLA ’07). ACM; 2007:57–76. doi:10.1145/1297027.1297033
34. Hanley JA, MacGibbon B. Creating non-parametric bootstrap samples using Poisson frequencies. Comput Methods Programs Biomed 2006;83(1):57–62. doi:10.1016/j.cmpb.2006.04.006

35. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178. doi:10.1038/sdata.2018.178
36. [dataset] Pollard T, Johnson A, Raffa J, et al. 2019. eICU Collaborative Research Database (version 2.0). PhysioNet. doi:10.13026/C2WM1R
37. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med 2005;24(8):1185–202. doi:10.1002/sim.1970
38. Wang SV, Sreedhara SK, Schneeweiss S; REPEAT Initiative. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. Nat Commun 2022;13(1):5126. doi:10.1038/s41467-022-32310-3
39. Centers for Medicare & Medicaid Services. CMS816FHIR v1.0.000, Hospital Harm – Severe Hypoglycemia (FHIR). <https://ecqi.healthit.gov/sites/default/files/ecqm/measures/CMS816FHIR-v1.0.000-FHIR.html> (accessed 15 October 2026).
40. Centers for Medicare & Medicaid Services. CMS FHIR Quality Measure Development IG, version 0.8.0-cibuild (continuous-integration build of 13 October 2026): authoring patterns. <https://build.fhir.org/ig/cqframework/cms-qmd/branches/main/patterns.html> (accessed 15 October 2026).
41. cqframework/dqm-content-qicore-2025, issue #57. <https://github.com/cqframework/dqm-content-qicore-2025/issues/57> (accessed 15 October 2026).
42. Pollard T, Moody BE, Lehman L, et al. PhysioNet as a global platform for biomedical research. Nat Health 2026;1(8):792–5. doi:10.1038/s44360-026-00096-z

Figures

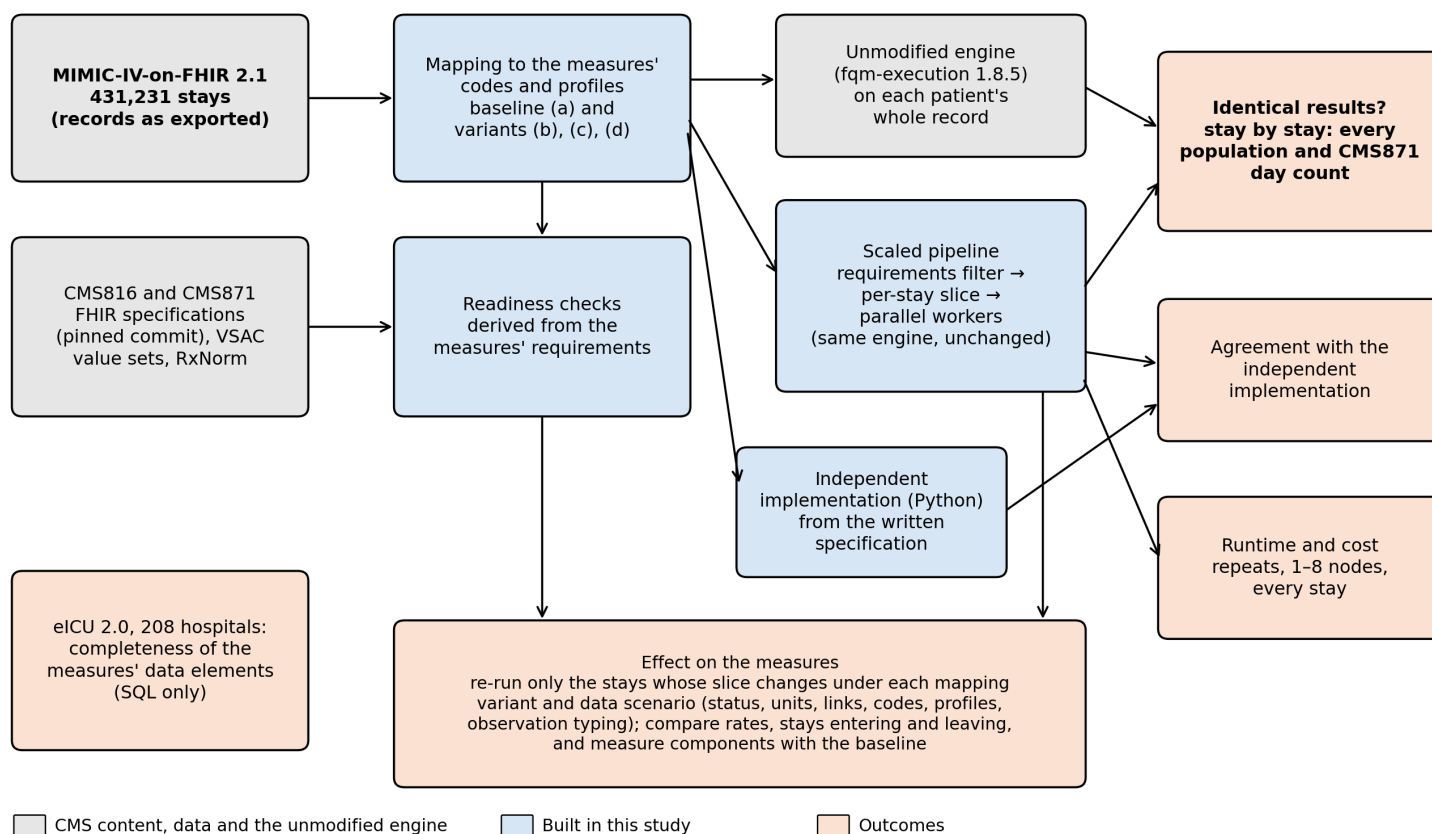


Figure 1

Study design. Grey: CMS content, data and the unmodified engine; blue: components built in this study; orange: outcomes.

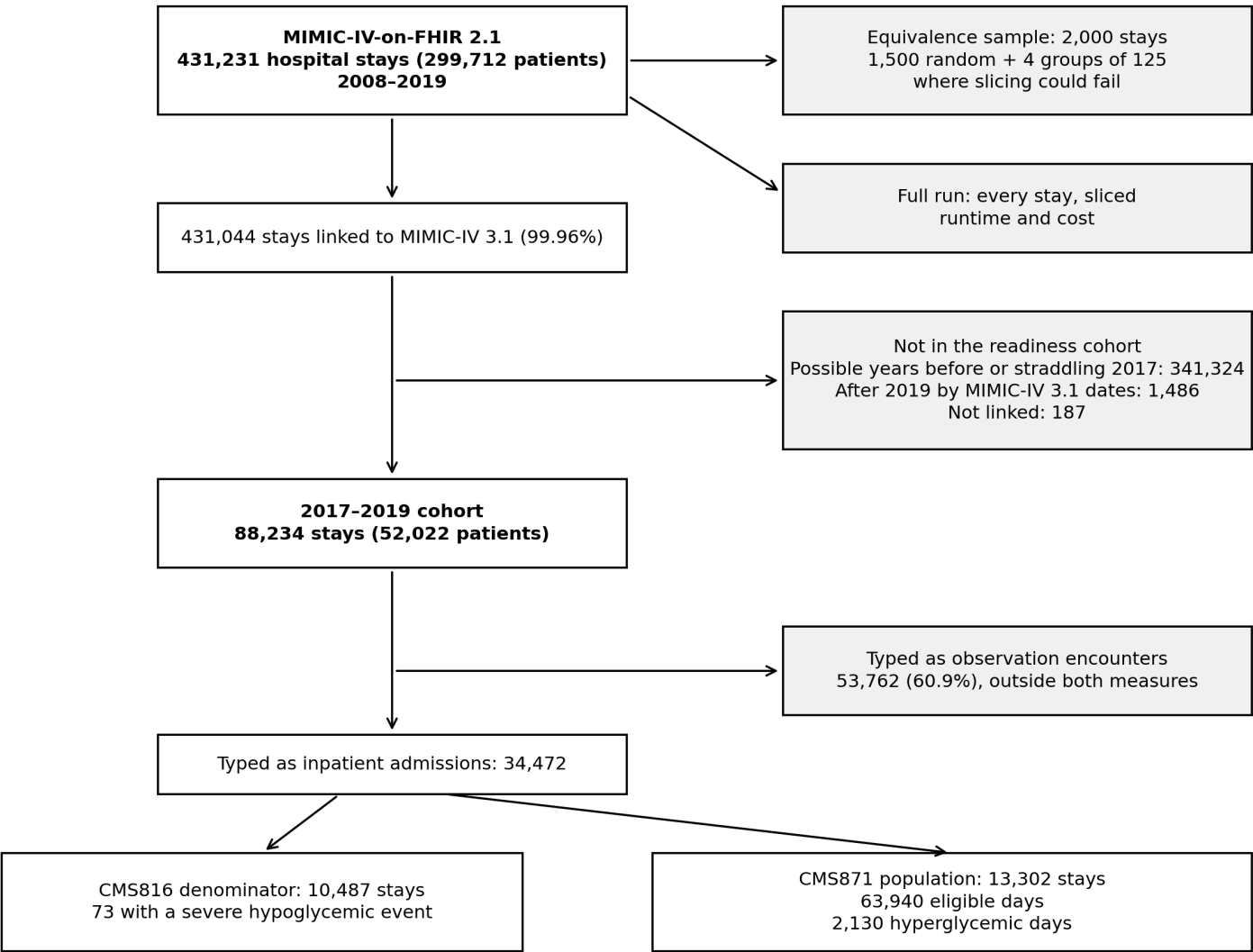
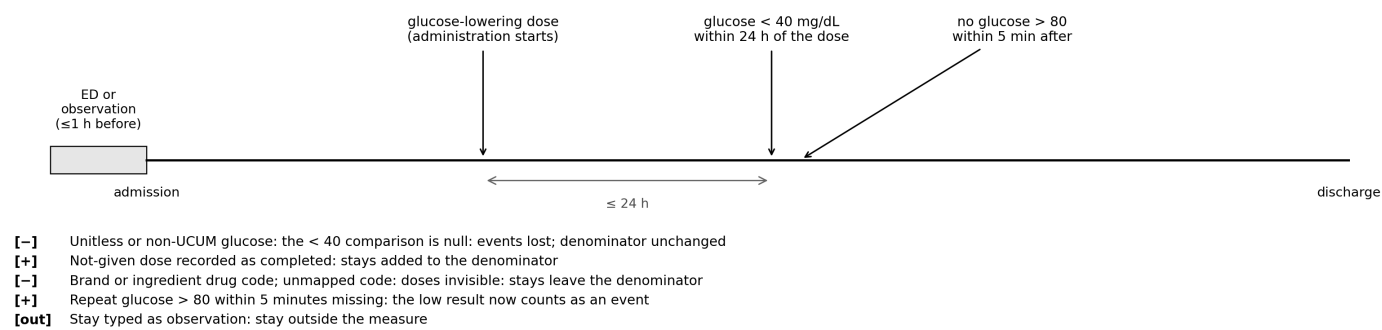


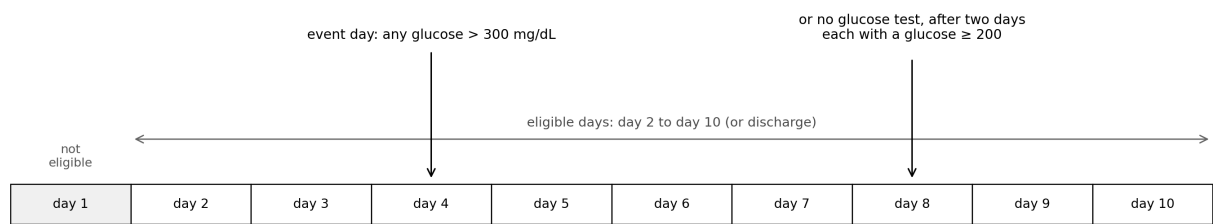
Figure 2

Stay flow from 431,231 stays to the 2017–2019 cohort and both measures' populations, with the equivalence sample and the full run.

A CMS816, severe hypoglycemia: one event per stay



B CMS871, severe hyperglycemia: event days among eligible days

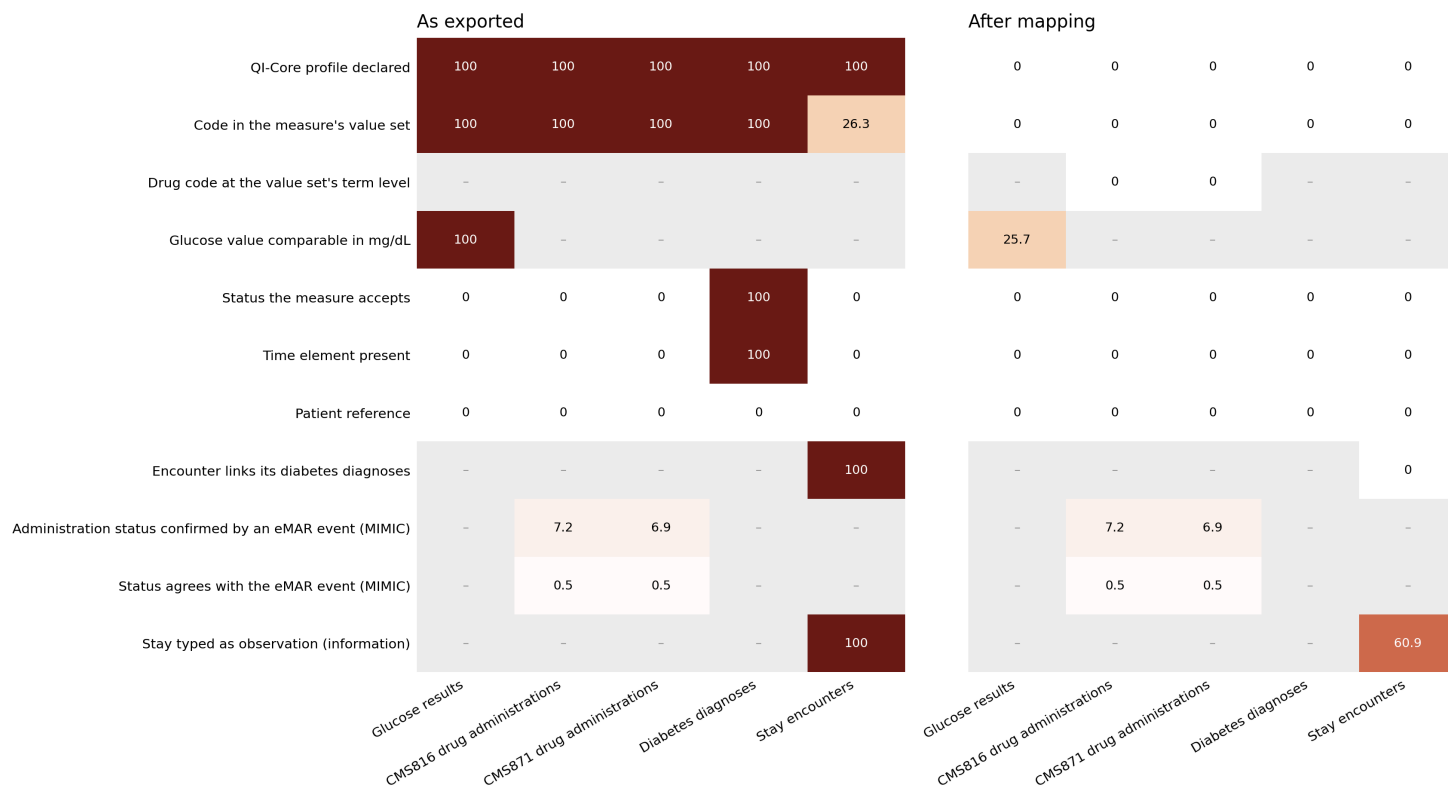


- [–]** Unitless or non-UCUM glucose: > 300 and ≥ 200 comparisons are null: event days lost
- [±]** Glucose tests missing from the record: a lost > 300 result removes an event day; the absence rule can create one
- [+]** Encounter not linked to its diabetes diagnosis: event-free stays leave the population: rate rises
- [–]** Glucose-lowering drug unmapped or at the wrong term level: stays leave the population
- [out]** Stay typed as observation: stay outside the measure

[–] lowers the count it acts on, [+] raises it, [±] either, [out] removes the stay from the measure. Logic as specified by CMS; days are whole 24-hour periods from the start of the hospitalization.

Figure 3

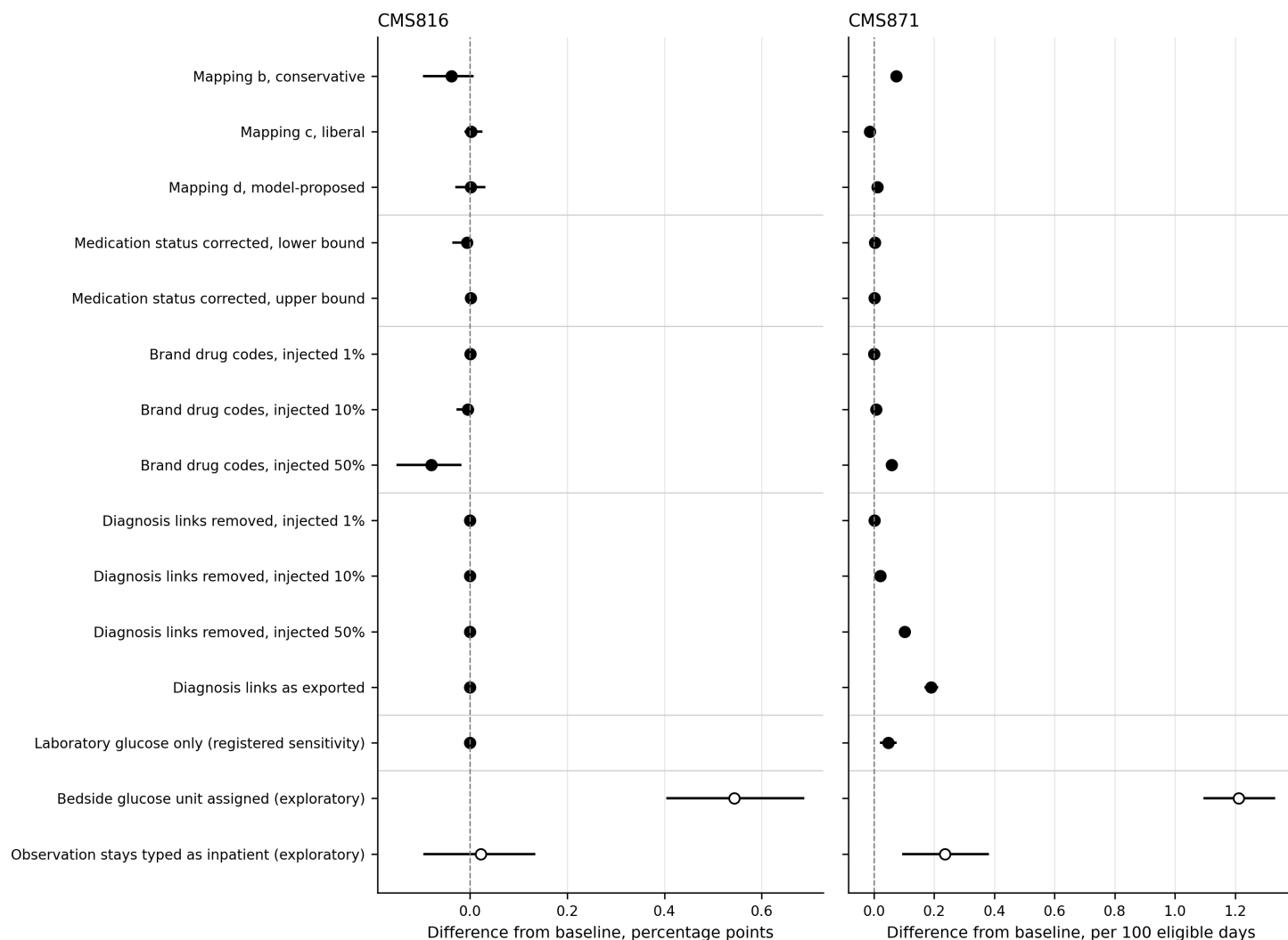
Where each readiness problem acts in the measure logic, and in which direction. Logic as specified in CMS's FHIR measure packages CMS816FHIRHHHypo and CMS871FHIRHHHyper (QI-Core 6.0.0, content commit 5f840c8).



Cells: % of records failing the check. Grey: check does not apply to that record type.

Figure 4

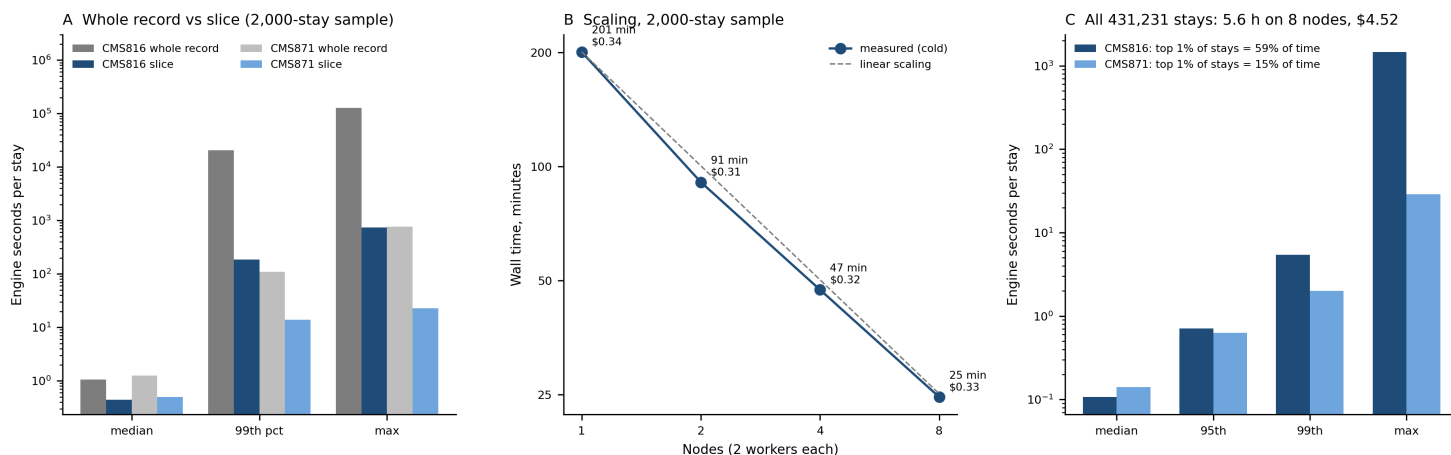
Readiness heatmap: % of records failing each check by record type, as exported and after mapping.



Filled: registered analyses. Open: exploratory. Lines: 95% CI (paired stay-level bootstrap). Units, profiles and codes as exported (rates fall to zero) are in Table 3.

Figure 5

Difference from baseline in each measure's rate under each medication mapping and data scenario, 2017–2019 cohort. Filled markers, registered analyses; open markers, exploratory; lines, 95% CIs from a paired stay-level bootstrap.



A: whole-record times exclude evaluations stopped at the cost cap (lower bounds). C: several stays per engine call; node cost at the on-demand price, cluster fee excluded.

Figure 6

Scale. A, per-stay engine time on whole records vs slices in the 2,000-stay sample; B, wall time by number of nodes; C, per-stay engine time over all 431,231 stays.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterial.docx](#)