

Computable longitudinal patient journeys from structured and unstructured EHR data

Received: 13 January 2026

Accepted: 2 September 2026

Published online: 10 September 2026

 Check for updates

Edward Kim ^{1,2,9}, Richard Foty ^{1,9}, Avnesh S. Thakor³, Jay S. Skyler ⁴, Lucy F. Robinson⁵, James D. Park ⁶, Monica Kraft⁷, Charles B. Cairns ^{8,10} & Vicki Seyfert-Margolis ^{1,10} 

Most electronic health record (EHR) data reside in unstructured text, recorded in encounters between healthcare providers and patients. These encounters capture symptom narratives, adverse effects, treatment rationales, adherence challenges, family history and other nuances that are not coded or only partially reflected in structured fields such as diagnosis and procedures codes. Yet the majority of real-world evidence studies rely on structured data fields. Here we present a new approach that uses large, pretrained language models to accurately extract computable clinical data from unstructured EHR text. Physician adjudication confirmed high accuracy against a blinded expert reference standard, with high expert physician inter-reviewer agreement. The framework is disease-agnostic and scales to large patient datasets across all clinical conditions. The extracted clinical entities are integrated with structured EHR data, embedded with medical ontologies, and organized into a knowledge graph, allowing all relationships between variables to be interrogated at the level of individual patients or at scale. An agentic-friendly programmatic interface navigates large-scale knowledge graph data, making analyses rapid, easy and cost-effective, both for exploration and hypothesis testing. This human-guided query environment enables acceleration of labor-intensive tasks, while study design, analytic decisions and interpretation remain under investigator control. To demonstrate clinical utility, we conducted a large-scale longitudinal analysis of treatment responses among individuals initiating glucagon-like peptide-1 receptor agonists (GLP-1 RAs) therapy. We identify large numbers of patients initiating GLP-1 RAs with high confidence, reconstruct patient-level trajectories, and model longitudinal changes in weight and hemoglobin A1c following GLP-1 RA initiation, revealing association between these parameters and response to therapy. Further, we characterize outcomes captured only in unstructured clinical notes and show time-to-event outcomes, moving beyond fixed-timepoint end points to provide a comprehensive understanding of real-world treatment responses.

Widespread adoption of electronic health records (EHRs) has created unprecedented opportunities for learning from real-world clinical care. Historically, EHR-based research and risk prediction focused predominantly on structured data fields, such as diagnosis and procedure codes. Yet most EHR data are unstructured, free-text documentation that capture symptom narratives, adverse effects, treatment rationales, adherence challenges, family history and nuances of disease and health status that are not coded or only partially reflected in structured fields. The structured EHR data represent the visible 20%, leaving the rich clinical narrative hidden¹.

A systematic review² found that less than 1% of real-world clinical studies utilized unstructured text; adding unstructured text raised median case-detection sensitivity from 62% to 78%. A recent survey reported that only 9 out of 984 pediatric research studies met their inclusion criteria, where the majority were excluded for not utilizing unstructured EHR data to identify adverse events³. Furthermore, only 2–4% of social determinants from unstructured clinical narratives are captured. Failure to extract unstructured insights blinds researchers to many drivers of healthcare outcomes, rendering much of a patient's journey invisible, causing many analyses and models to be flawed⁴.

Classical natural language processing (NLP) pipelines, using rule-based systems and concept matchers, such as cTAKES⁵ and Meta-Map⁶, have demonstrated that unstructured notes could be mined, but struggle with long context and have limitations analogous to key word searching. Biomedical AI remains largely 'unimodal' and fails to integrate narrative context essential for precision medicine⁷. Recently developed transformer-based models have improved extraction performance^{8–15} and minimized AI hallucinations¹⁶. Yet challenges remain to making unstructured EHR data extractions accurate and outputs valid for clinical use. Recent US Food and Drug Administration (FDA) guidance states 'large amounts of key clinical data are unstructured data within EHRs¹⁷, and extra care must be taken when converting this data into computable formats.'

We present a new approach using large, pretrained language models to accurately extract computable clinical data from unstructured EHR text. The extracted clinical entities are organized into a knowledge graph (KG), a data representation in multidimensional space, and paired with an agentic AI programmatic interface that allows for queries across multiple variables at the level of individual patients or at scale. The framework is disease-agnostic and scales to large patient datasets across all clinical conditions. The first section of our paper focuses on our extraction methods and their analytical validity, verifiability and key principles that ensure accuracy of extractions and analytic outputs. The second section focuses on the clinical utility of the extracted data demonstrated with the use case of a large-scale longitudinal analysis of treatment responses among individuals initiating glucagon-like peptide-1 receptor agonists (GLP-1 RAs). We identify large numbers of patients initiating GLP-1 RAs with high confidence, reconstruct patient-level trajectories and model longitudinal changes in weight and hemoglobin A1c (HbA1c) following GLP-1 RA initiation. Finally, we discuss the implications of this approach for large-scale analyses supporting clinical research and care.

Results

Extraction validation and knowledge graph assembly

We developed a framework that integrates structured and unstructured EHR data into a computable, reviewable, longitudinal representation of care (Extended Data Fig. 1). The framework combined structured clinical data with information extracted from narrative clinical notes, mapped these elements to existing biomedical ontologies, derived data-driven taxonomies for patient-centered concepts that established ontologies do not cover, and organized the result within a patient-centered KG. Across this workflow, self-improving AI research agents (AI agents iterating and improving their own agent skills) were used for EHR note extraction, ontology grounding, graph assembly,

retrieval, cohort construction and generation of reviewable analysis code under investigator oversight (Extended Data Fig. 2).

Patient data pipeline for extractions and analyses

A large de-identified ambulatory EHR data resource was used as our source data. This resource is derived from multiple healthcare provider systems and spans four distinct EHR systems, 1,337 physician practices and 8,871 unique providers. These source data included structured clinical records and longitudinal free-text documentation. Starting with a data lake of approximately 35 million patient EHR's, we selected 213,000 individuals with GLP-1 RA (semaglutide and tirzepatide) exposure in this longitudinal ambulatory EHR resource. Included patients had up to 10 years of available clinical history (Fig. 1). We analyzed unstructured EHR text from a random sample of approximately 1,000,000 clinical visits to ensure computational feasibility at scale, capturing the complete medical histories 31,235 patients in our data source (Fig. 1).

Extraction process quantification and validation

The patient narrative carries important clinical detail that coded fields omit and is not analyzable in its recorded form. We therefore decompose each note into a set of entities along with their associated atomic statements (single standalone propositions evidenced from the source note) using a large language model-based extraction pipeline. Large language models (LLMs) were used with structured generation and domain-specific prompting to derive computable clinical information from unstructured text. Because notes are longitudinal and heavily copy-forwarded, extraction was defined as a change-tracking operation where the current note is compared to the previously stored extraction. Findings are labeled as added, modified or removed, with stable terms carried forward without duplication (Extended Data Fig. 3). This allows extracted events to be placed on an encounter timeline and assembled into patient trajectories. Entities were prompted for extraction in the following categories: medication, self-reported symptoms, adherence, family history, hospitalizations, quantitative assessments and disease-status (Methods provides descriptions of each category). To establish that note-derived elements are accurate to support downstream analysis, we evaluated the extraction pipeline against a blinded physician reference standard. The reference set comprised 100 clinical visits spanning 96 providers, 86 practices and four EHR systems. A primary physician reviewer scored every model-extracted statement across three domains in all 100 visits. A second physician scored 30 of those visits, 10 alongside the primary reviewer for calibration and 20 independently, and a third physician, blinded to reviewer identity, adjudicated disagreements on the 20 independently scored visits to produce an adjudicated reference (Methods and Table 1).

Overall, precision and recall of the model achieved an F1 (harmonic mean of precision and recall) score of 99.4% (precision 99.2%, recall 99.6%), with per-domain F1 of 98.9% for medications, 99.5% for phenotypes and findings and 100% for quantitative assessments (Table 1).

Across the remaining 70 visits scored by the primary reviewer alone, the model achieved an F1 score of 97.6% (precision 99.3%, recall 95.9%; 1,756 true positives, 13 false positives, 75 missed findings), with per-domain F1 of 96.0% for medications, 97.4% for phenotypes and 99.4% for quantitative assessments (Table 1). In the planted-error probe (32 withheld correct findings identified by the model; Methods), reviewers recovered 26 of errors (81.2% sensitivity).

A 20-visit subset was independently scored by a second physician and adjudicated subsequently and independently by a third physician, to quantify inter-reviewer reliability and provide an adjudicated reference for the model's precision and recall (Methods). On this subset the two reviewers agreed on 94.7% of 698 jointly scored statements (Gwet's AC10.943) and the model's precision as judged by the primary reviewer (99.3%) matched its precision against the adjudicated reference, indicating that primary reviewer scoring was reliable.

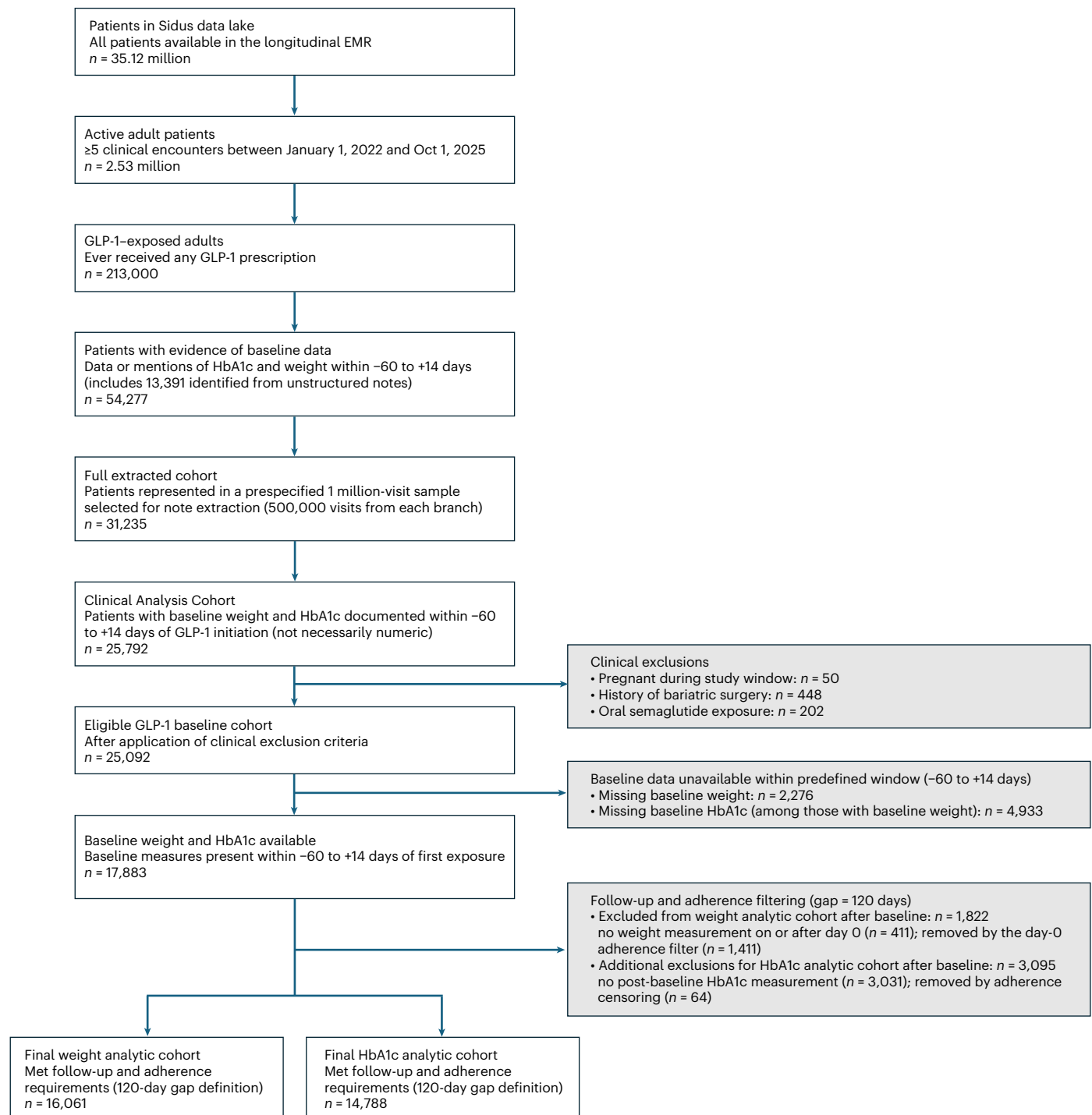


Fig. 1 | Flow diagram. Summarizing cohort selection, baseline assessment pathways and definition of the final analytic cohorts for longitudinal weight and HbA1c analyses.

Second, we compared our models with existing machine-learning named entity extractors. A primary expert and two additional expert reviewers from multiple institutions independently reviewed the extracted entities produced by each method across 20 visits (Table 1 and Extended Data Table 1). Reviewers adjudicated machine-extracted output rather than annotating notes *de novo*; accordingly, we report precision and characterize error types rather than recall. For phenotype extraction, inter-rater agreement was near-perfect for our GPT-based pipeline ($\kappa = 0.955$ – 1.000 across pairs) and substantially lower for the non-LLM baseline PhenoTagger ($\kappa = 0.600$ – 0.800), reflecting

the latter's inconsistent handling of clinical language. Our pipeline achieved 99.5% precision on 1,889 adjudicated phenotype extractions; PhenoTagger reached only 15.3%, with 83.4% of its extractions representing negation errors (phenotypes asserted as present when they were in fact absent). For medication extraction, inter-rater agreement was again highest for our pipeline ($\kappa = 0.917$ – 1.000) compared to BERT Large Med named entity recognition (NER)^{18,19} ($\kappa = 0.781$ – 0.964) and BioNER²⁰ ($\kappa = 0.788$ – 1.000). Our pipeline achieved 95.5% precision on 221 adjudicated medication extractions, versus 81.6% for BioNER and 65.4% for BERT Large Med NER. Together, these findings indicate that

Table 1 | Extraction validation

GPT-5.2 Low model-extracted statements versus physician reference						
Domain	Adjudicated reference (20 visits)			Single reviewer (visits 31–100)		
	Precision	Recall	F1	Precision	Recall	F1
Medications	98.9%	98.9%	98.9%	98.0%	94.1%	96.0%
Phenotypes & findings	99.1%	100%	99.5%	100%	95.0%	97.4%
Assessments	100%	100%	100%	98.8%	100%	99.4%
Overall	99.2%	99.6%	99.4%	99.3%	95.9%	97.6%
Comparison with other extraction models (20 visits)						
Domain	Model	N	Precision	Negation error		
Phenotypes	GPT-5.2 Low	1,889	99.5%	0.0%		
	PhenoTagger	1,847	15.3%	83.4%		
Medications	GPT-5.2 Low	221	95.5%	2.3%		
	BioNER	179	81.6%	3.8%		
	BERT Large Med NER	344	65.4%	3.4%		

Metrics are entity-level. Top (validation against physician review): adjudicated results are from a 20-visit subset independently scored by two physicians and reconciled by a blinded third physician adjudicator; single-reviewer results are from one physician across visits 31–100. Bottom (comparison with other models): N is the number of physician review events per method over the same 20-visit set (repeated review of a shared sample rather than independent patient samples); negation error is the proportion of extractions asserted as present when the note recorded them as absent.

our extraction framework recovered clinically relevant information from unstructured notes with high validity and was superior to the non-LLM comparator methods.

After validation, we ran the full extraction pipeline over 1 million clinical visits, recovering more than 49 million clinical entities, including 14.6 million self-reported symptoms, 12.8 million phenotype data points, 1.6 million medication switches and 1.3 million adherence events. The full pipeline was run using the more cost-efficient GPT-5 mini and had a 0.80 concept recall with GPT-5.2 (Table 2). Verification scores of key clinical entity extractions from the source text (symptoms, disease, phenotypes, assessments, history, current medication and medication adherence) were high, ranging from 94.1% to 98.4% (Table 2). Medication switching among agents and dose had slightly lower verification rates (74% to 81%), mainly due to contradictory statements in the source text.

Concept discovery

When standard medical terminology already exists, we map extracted terms to associated vocabulary, such as tenth revision of the International Classification of Diseases (ICD-10), Human Phenotype Ontology (HPO) and Unified Medical Language System (UMLS). Many patient-centered concepts have no standard terminology, we therefore derived groupings directly from the notes themselves. Based on questions, we organize note-derived events into data-driven taxonomies. In this case, we included medication symptoms, medication stops, medication switches, dose change reasons, adherence barriers, adherence recommendations and hospitalization reasons. We embedded short text spans from verified events, clustered them by semantic similarity and assigned labels to each cluster. A second grouping step organized these labels into higher-level parent categories, yielding a two-level taxonomy of recurrent clinical categories.

This approach identified concise, reusable groupings from unstructured documentation, while preserving linkage to the underlying note text. In medication switching, for example, the pipeline grouped related reasons for switching medications into clinically coherent categories that could be used as downstream analytic filters (Fig. 2a). A low-dimensional projection of the embeddings (Fig. 2a)

separated the discovered groupings, and their stability across projection parameter settings was assessed as described in Methods.

These findings indicate that the concept-discovery pipeline organized heterogeneous narrative statements into taxonomies linked to underlying note text for downstream analysis.

Computable elements for clinical analyses

We assembled a patient-centered KG for longitudinal analyses. A KG represents data as entities (nodes) connected by explicit semantic relationships (edges). In the patient-centered graph used here, nodes represent patients, visits, ontology concepts and clinical entities drawn from structured records and unstructured notes, and edges carry temporal as well as semantic relationships, with each element retaining a link to its source²¹. This architecture allows for graph traversal, revealing connections across any data, for example from a patient node to their medication history to specific adverse events extracted from notes. The combination of extracted structured and unstructured data with the agentic framework allows downstream users to query multiple clinical parameters at the level of individual patients or at scale (millions of patient visits).

The KG contained the clinical analysis cohort defined in Fig. 1 (25,092 patients with baseline weight and HbA1c documented from 60 days before to 14 days after GLP-1 RA initiation). This KG contains 88 million nodes, 241 million edges and is mapped to more than 58,000 UMLS concepts, 8,736 HPO concepts and 39,827 unique medications.

Each patient was represented as a temporally linked KG that integrates unstructured note-derived phenotypes with structured EHR data. As shown, addition of unstructured variables substantially increases size, density and complexity of the patient KG (Fig. 2b). Notably, each element remains linked to source evidence, visit metadata, assertion or negation status and verification level. Each element is grounded to reference terminologies (HPO, UMLS and RxNorm) or our data-driven taxonomy. Retrieval, grounding and graph-construction steps were performed using agents that generated reproducible code and preserved provenance links to underlying source data. Our unified representation supported both cohort-level analytics and auditable reconstruction of longitudinal patient journeys. Shared visit anchors enabled temporal queries linking glucagon-like peptide-1 receptor agonist (GLP-1 RA) exposure to emergent or resolving phenotypes and to contemporaneous changes in laboratory values, vital signs and coded diagnoses, while preserving event order and status changes across visits. This representation also generated patient-level journey reconstructions with linked source evidence (original EHR) and uncertainty quantification metadata (Fig. 2c).

Analyses over this graph were carried out by research agents working in an analytic environment. Given a research question, an agent retrieves the relevant subgraph, drafts an analysis plan, writes the code that implements it and builds intermediate data structures that the investigator inspects before results are produced. Every step is written to persistent, reviewable code, so the path from a research question to a reported number is reproducible.

Clinical application: Real-world outcomes after GLP-1 RA initiation in routine care

Analytic validity does not by itself demonstrate clinical utility. We therefore applied the framework to GLP-1 RAs, a drug class whose rapid uptake has reshaped the management of type 2 diabetes (T2DM) and obesity at population scale.

We used our AI agent workflows in our KG composed of patients who initiated GLP-1 RA therapies (semaglutide and tirzepatide), to identify eligible patients, longitudinally align note-derived and structured observations around treatment initiation, and generate reviewable analytic code to assemble longitudinal cohorts for weight and glycemic analyses. Cohort assembly and attrition are summarized in Fig. 1.

Table 2 | Cohort summary and extraction verification statistics

Total cohort summary of 1million visits from 31,235 patients.						
Category	Records	Entities	Score of 3	Score of 2	Score of 1	Score of 0
Self-R symptoms	884,246	14.6M	14.0M (96.2%)	474K (3.2%)	56K (0.4%)	22K (0.2%)
Phenotypes	890,681	12.8M	12.2M (95.5%)	479K (3.7%)	57K (0.4%)	41K (0.3%)
Diseases/conditions	902,145	9.2M	8.6M (94.1%)	477K (5.2%)	45K (0.5%)	21K (0.2%)
Current meds	746,818	4.4M	4.3M (97.8%)	82K (1.9%)	8K (0.2%)	5K (0.1%)
Assessments	611,576	2.6M	2.6M (98.4%)	34K (1.3%)	4K (0.2%)	3K (0.1%)
Family history	720,315	2.0M	1.9M (98.2%)	13K (0.7%)	10K (0.5%)	12K (0.6%)
Meds switched	516,708	1.6M	1.2M (75.5%)	287K (17.6%)	77K (4.7%)	32K (2.0%)
Adherence MDE	690,072	1.3M	1.3M (97.2%)	29K (2.2%)	6K (0.5%)	712 (0.1%)
Dosage change	245,953	429K	348K (81.1%)	52K (12.3%)	8K (2.1%)	19K (4.4%)
Medication stop	234,680	419K	411K (98.0%)	5K (1.4%)	1K (0.4%)	774 (0.2%)
Med adv. effect	213,187	330K	299K (90.4%)	18K (5.6%)	11K (3.5%)	1K (0.5%)
Hospitalizations	134,923	219K	199K (90.9%)	14K (6.4%)	4K (2.1%)	1K (0.5%)
Fine-tuning cohort: phenotypes of 100K visits from 3,271 patients.						
Phenotypes	94,805	1.56M	1.49M (90.9%)	59K (6.4%)	6K (2.1%)	4K (0.5%)
Analysis cohort: 15K patients; 141K visits; 1.05M phenotype events (365-day window).						
Baseline split	Visits	Events	Total visits	Visits %	Events %	Unique codes
After baseline	81,043	618,555	15K	57.40%	58.80%	5,010
Before baseline	60,228	432,928	15K	42.60%	41.20%	4,632

The cohort consists of 1,000,000 total visits across 31,235 unique patients using GPT-5 Mini Reasoning Low. Verification scores indicate the level of evidence support: 3 (fully supported), 2 (partially supported), 1 (inferred) and 0 (not supported). Records, the number of source records from which each domain was extracted; entities, the number of extracted data points that the four verification-score columns partition. M, million; K, thousand; Self-R, self-reported; Adherence MDE, adherence to medication, diet and exercise; Med adv. effect, medication adverse effect.

After applying plausibility, follow-up and persistence filters, the primary cohorts included 16,061 and 14,788 individuals for longitudinal weight and HbA1c analyses, respectively. The smaller HbA1c cohort reflected less frequent routine-care testing. For our persistence cohort (using a 120-day persistence definition), the weight cohort decreased from 16,061 at baseline to 1,247 at 12 months and 367 at 18 months, reflecting expected attrition over time (Supplementary Table 1).

Baseline characteristics varied across glycemic strata (Table 3). Individuals with normal glycemia were younger and had lowest baseline HbA1c, whereas those with T2DM and poorly controlled diabetes were older (more than 95% aged ≥ 40 years in poorly controlled group) and had progressively higher baseline HbA1c's. Baseline weight and body mass index (BMI) were highest in prediabetes and lowest in poorly controlled diabetes. Obesity was common across all groups, with more than two-thirds meeting criteria for class I–III obesity.

Weight-loss thresholds were broadly similar across strata, although $\geq 10\%$ and $\geq 15\%$ weight loss was most frequent in normal glycemia and least frequent in poorly controlled diabetes. By contrast, glycemic improvement increased with baseline HbA1c, whereas individuals with normal glycemia showed minimal absolute change.

Contribution of unstructured clinical documentation

Because these cohorts were assembled from both structured fields and note-derived values, we next quantified how much of the analyzable data each source contributed. Of the 14,788 patients in the HbA1c analytic cohort, 42.2% of baseline HbA1c values came from unstructured notes, 52.2% from structured EHR data and 0.6% from both. Documentation source was stable within patients over time. Among those whose baseline HbA1c came from notes, 84.0% of follow-up measurements also came from notes. These patients were followed more frequently than patients with structured baselines, averaging 7.31 versus 3.00 HbA1c observations, respectively, suggesting that for unstructured HbA1c documentation was the primary longitudinal data source. Across all patients, 70.0% of HbA1c observations were identified exclusively

from unstructured clinical notes, compared to 28.7% from structured laboratory data and 1.3% from both sources. In contrast, weight was captured predominantly in structured vital-sign data (98.8%). GLP-1 RA exposure relied on both sources, with 47.9% of events supported by structured prescription records, 33.4% exclusively by unstructured clinical notes and 18.6% by both.

The addition of data from unstructured documentation contributed to the analyses in two distinct ways. Most consequentially, it made entire clinical domains analyzable that structured fields do not capture measures such as depression severity, pain intensity, waist circumference and adherence barriers. Second, it expanded ascertainment for domains that structured fields do capture: 6,238 patients entered the HbA1c cohort and 6,684 the weight cohort only because a baseline was recovered from clinical notes and would have been excluded using structured baselines alone (overlapping cohorts; counts are not additive). This tightened estimates for outcomes already supported by structured data. A sensitivity analysis using structured-only data preserved the direction of every association but widened the confidence intervals (Extended Data Table 2), confirming that note-derived information sharpened rather than distorted the results. Because weight and HbA1c are well captured in both structured and unstructured fields, this is a conservative test case on the value of unstructured data.

Stratified longitudinal weight and HbA1c trajectories

Having established that note-derived data broadened ascertainment without distorting estimates, we examined whether treatment response differed by baseline glycemic status. Model-based longitudinal trajectories showed greater and more rapid weight loss after GLP-1 RA initiation among individuals with lower baseline glycemic burden, with progressively attenuated weight loss across prediabetes, T2DM and poorly controlled diabetes (Fig. 3a). Separation between glycemic strata emerged early and persisted throughout follow-up. At 12 months, model-predicted weight change ranged from -7.7% (95% CI -8.7 to -6.7) in normal glycemia to -5.9% (-6.6 to -5.2) in prediabetes,

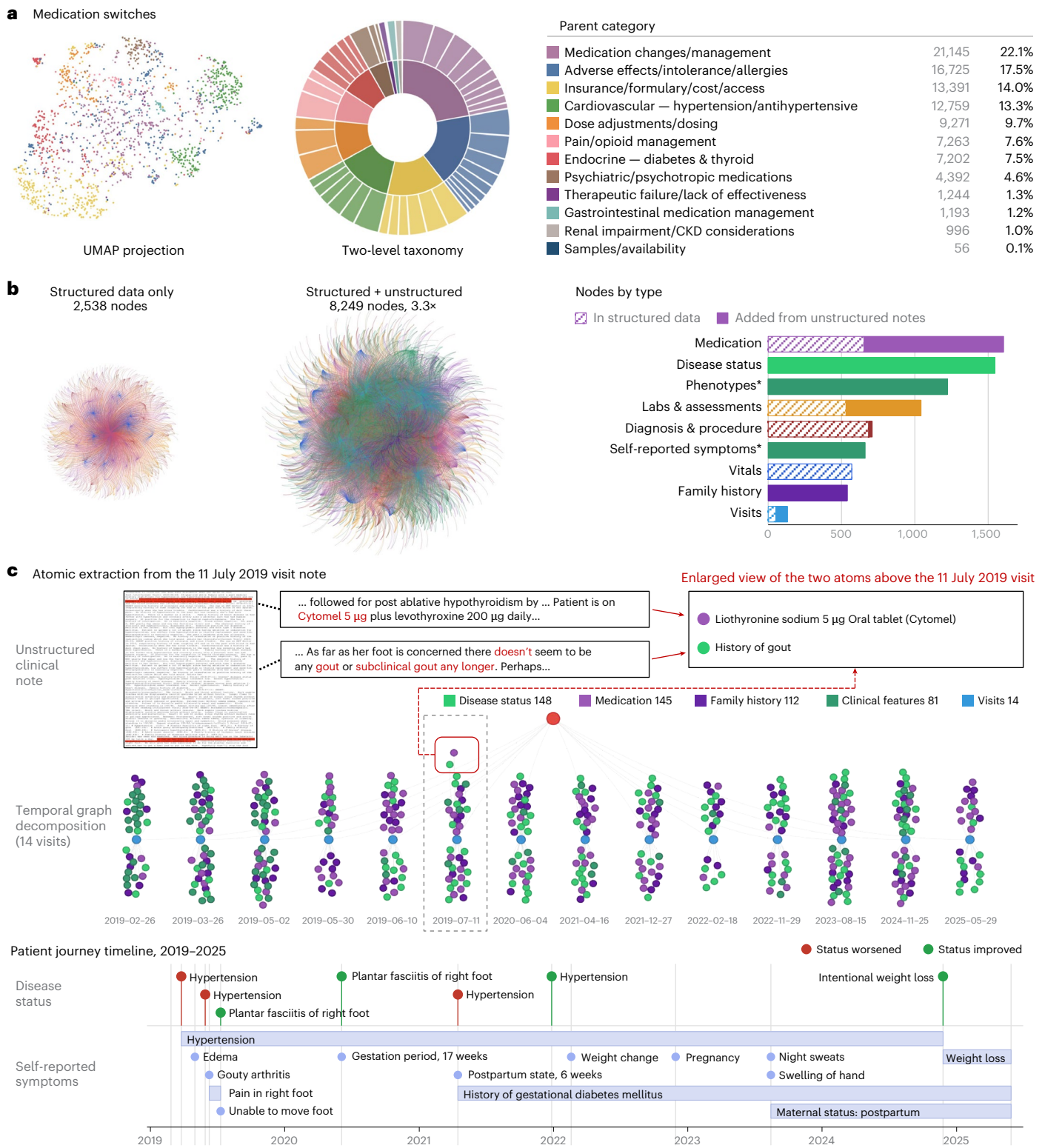


Fig. 2 | Data-driven taxonomies, knowledge graphs and reconstructed patient journeys derived from unstructured clinical notes. a, Right: two-level taxonomy of reasons for medication switching: 95,637 note-derived items in 50 subcategories and 12 parent categories. Inner ring, parent categories; outer ring, subcategories, unlabeled. Left: UMAP projection of embedded items, colored by parent category. **b,** Knowledge graphs for one patient (52 visits) from structured EHR data alone (2,538 nodes) and after adding note-derived entities

(8,249 nodes); rendered area is proportional to node count. Asterisks mark the two clinical-feature families that **c** shows as one. Assessments are note-derived quantitative values including HbA1c and PHQ-9. **c,** Reconstructed journey for an additional patient (14 visits): atomic extraction from one visit note, temporal decomposition of the graph by visit, and the extracted events on a time axis. Colors are as in **b**. Dashed outline, the source visit; red outline, the two atoms enlarged at right.

Table 3 | Baseline demographic and clinical characteristics of the primary study cohort, stratified by baseline glycemic status

Characteristic	Total (N=16,061)	Normal glycemia (n=3,468)	Prediabetes (n=4,316)	T2DM (n=5,864)	Poorly controlled diabetes (n=2,413)
Baseline HbA1c, %	7.3±1.9	5.4±0.3	6.1±0.2	7.6±0.7	10.8±1.5
Age, years	59.7±13.0	52.7±13.5	59.6±12.3	63.4±11.7	61.1±12.2
Weight, lbs	217.7±47.4	216.2±45.4	223.4±47.0	216.6±47.9	212.1±48.5
BMI, kg m ⁻²	35.1±6.6	35.4±6.5	36.3±6.5	34.6±6.6	34.0±6.6
Age group					
<40 years	1,103 (6.9%)	569 (16.4%)	268 (6.2%)	160 (2.7%)	106 (4.4%)
≥40 years	14,958 (93.1%)	2,899 (83.6%)	4,048 (93.8%)	5,704 (97.3%)	2,307 (95.6%)
Baseline BMI category					
Underweight	20 (0.1%)	3 (0.1%)	5 (0.1%)	6 (0.1%)	6 (0.2%)
Normal weight	674 (4.2%)	107 (3.1%)	107 (2.5%)	292 (5.0%)	168 (7.0%)
Overweight	2,938 (18.3%)	599 (17.3%)	599 (13.9%)	1,197 (20.4%)	543 (22.5%)
Obese class I	4,626 (28.8%)	1,033 (29.8%)	1,230 (28.5%)	1,669 (28.5%)	694 (28.8%)
Obese class II	3,777 (23.5%)	848 (24.5%)	1,106 (25.6%)	1,307 (22.3%)	516 (21.4%)
Obese class III	3,364 (20.9%)	728 (21.0%)	1,113 (25.8%)	1,105 (18.8%)	418 (17.3%)
Unknown	662 (4.1%)	150 (4.3%)	156 (3.6%)	288 (4.9%)	68 (2.8%)
Sex					
Female	9,787 (60.9%)	2,671 (77.0%)	2,861 (66.3%)	3,040 (51.8%)	1,215 (50.4%)
Male	6,274 (39.1%)	797 (23.0%)	1,455 (33.7%)	2,824 (48.2%)	1,198 (49.6%)
Race					
White	6,543 (40.7%)	1,418 (40.9%)	1,731 (40.1%)	2,485 (42.4%)	909 (37.7%)
Other	2,432 (15.1%)	403 (11.6%)	695 (16.1%)	881 (15.0%)	453 (18.8%)
Unknown/refused	7,086 (44.1%)	1,647 (47.5%)	1,890 (43.8%)	2,498 (42.6%)	1,051 (43.6%)

Values are mean±s.d. for continuous variables and n (%) for categorical variables. Percentages are column percentages calculated within each glycemic stratum; categories are mutually exclusive and exhaustive, so counts sum to the column total. Baseline glycemic status was assigned from the baseline HbA1c value: normal glycemia <5.7%, prediabetes 5.7 to <6.5%, T2DM 6.5 to <9.0%, and poorly controlled diabetes ≥9.0%. Baseline values are those recorded from 60 days before to 14 days after GLP-1 RA initiation. Unknown or refused responses are retained as a separate category rather than excluded. The table is descriptive; no statistical comparisons were made between strata.

–4.9% (–5.5 to –4.2) in T2DM and –2.7% (–3.6 to –1.8) in poorly controlled diabetes (Extended Data Table 3).

HbA1c trajectories showed the opposite gradient. Absolute HbA1c reduction increased with baseline glycemic severity, with largest declines in poorly controlled diabetes and smaller reductions in T2DM and prediabetes (Fig. 3b). Most HbA1c decline occurred early after initiation and then plateaued. At 12 months, model-predicted HbA1c change was minimal in normal glycemia (–0.01 percentage points; 95% CI –0.08 to 0.05), modest in prediabetes (–0.19; –0.29 to –0.08), larger in T2DM (–0.63; –0.77 to –0.49), and greatest in poorly controlled diabetes (–2.36; –2.63 to –2.08) (Extended Data Table 3). Despite these reductions, mean HbA1c in the T2DM and poorly controlled diabetes groups remained above diagnostic thresholds for diabetes during follow-up.

Weight trajectories differed by sex and age (Fig. 3c,d). Females and younger individuals (20–39 years) showed greater weight loss over follow-up than males and those aged ≥40 years, respectively, although trajectory shapes were similar across strata. At 12 months, model-predicted weight change was –6.1% (95% CI –6.5 to –5.7) in females versus –4.0% (–4.4 to –3.5) in males, and –8.1% (–9.5 to –6.7) in individuals aged 20–39 years versus –5.1% (–5.4 to –4.8) in those aged ≥40 years (Extended Data Table 3).

HbA1c trajectories showed minimal variation by sex, age or baseline BMI and were largely overlapping across these subgroups. Correspondingly, model-predicted 12-month HbA1c change differed little by sex, with reductions of –0.74 percentage points (95% CI –0.81 to –0.68) in females and –0.78 (–0.87 to –0.70) in males, and varied

only modestly by age, from –0.59 (–0.80 to –0.39) in individuals aged 20–39 years to –0.77 (–0.82 to –0.71) in those aged ≥40 years (Extended Data Table 3).

Time to weight loss and HbA1c reduction

Mean trajectories describe the average magnitude of change but not when patients reach clinically meaningful thresholds, we therefore also modeled time to clinically meaningful events. Time-to-event analyses showed slower and less frequent weight loss with increasing baseline glycemic burden (Fig. 3e(i–iii)). Median time to ≥5% weight loss was 210 days (95% CI 196–219) in normal glycemia, 238 days (219–257) in prediabetes, 285 days (266–302) in T2DM and 413 days (357–449) in poorly controlled diabetes. Median time to ≥10% weight loss was estimable only in normal glycemia (450 days; 95% CI 412–488), and median time to ≥15% weight loss was not estimable in any group.

Cox models showed the same pattern: relative to normal glycemia, hazards of achieving weight-loss thresholds declined progressively across worsening baseline glycemic categories, from 0.92 (95% CI 0.85–0.99) in prediabetes, 0.83 (0.76–0.89) in T2DM and 0.65 (0.59–0.72) in poorly controlled diabetes for ≥5% weight loss, to 0.80 (0.68–0.94), 0.63 (0.53–0.74) and 0.46 (0.37–0.58), respectively, for ≥15% weight loss (Extended Data Table 2).

Time-to-event analyses showed clear separation across baseline glycemic categories, with faster and more frequent HbA1c reduction at higher baseline HbA1c levels (Fig. 3e(iv–vi)). Median time to a ≥0.5-point HbA1c reduction was 99 days (95% CI 97–102) in poorly controlled diabetes and 151 days (146–160) in T2DM, compared to

468 days (412–528) in prediabetes; median time was not estimable in normal glycemia because events were infrequent. Median time to a ≥ 1.0 -point reduction was estimable in T2DM (294 days; 95% CI 272–314) and poorly controlled diabetes (106 days; 104–110); the median time to a ≥ 1.5 -point reduction was estimable only in poorly controlled diabetes (116 days; 95% CI 112–121).

Cox models showed the same gradient. Relative to normal glycemia, the hazard of achieving a ≥ 0.5 -point HbA1c reduction was 3.77-fold higher in prediabetes (95% CI 3.17–4.48), 10.98-fold higher in T2DM (9.32–12.93) and 20.08-fold higher in poorly controlled diabetes (17.00–23.72); this gradient strengthened at the ≥ 1.0 -point threshold, with hazard ratios of 3.00 (2.06–4.38), 29.16 (20.64–41.21) and 82.95 (58.66–117.29), respectively (Extended Data Table 2). Additional sensitivity analyses of patient trajectories including those for longitudinal weight and HbA1c, were across alternative GLP-1 RA persistence definitions, spanning gap thresholds from 30 to 730 days, all evaluated within the 18-month follow-up window; the widest threshold retains the full cohort with no persistence-based censoring.

Unstructured note-derived longitudinal trends

Many clinically important elements in general and in the case of GLP-1 RAs are only captured in unstructured clinical notes and have been discussed as potentially impacted by GLP-1 RAs in addition to the indicated end points of weight loss and HbA1c. Having recovered these measures as computable, date-stamped values, we analyzed their trajectories in the same cohort as an exploratory analysis.

Longitudinal change analyses required a baseline assessment within ± 30 days of GLP-1 RA initiation and at least one repeat assessment between 30 and 365 days, yielding analysis populations for depression (patient health questionnaire (PHQ)-9) (625 patients), general pain intensity (436 patients), waist circumference (99 patients), alcohol use (198 patients) and muscle strength (219 patients) (Extended Data Table 4). Characteristics of the wider populations from which these analysis populations were drawn, comprising all patients with at least one documented assessment in each domain, are given in Supplementary Table 2; across domains, most patients were aged ≥ 40 years, obesity was common and mean baseline HbA1c ranged from 6.4% to 7.6%.

Model-predicted changes at 3, 6, 9 and 12 months, with 95% CIs for every domain and baseline subgroup, are given in Extended Data Table 4. Depressive symptoms improved most in patients with elevated baseline scores (Fig. 3f(i)): the estimated 12-month change was -2.52 points (95% CI -5.01 to -0.04) with baseline PHQ-9 ≥ 5 ($n = 263$) and -6.01 points (-9.28 to -2.74) with baseline PHQ-9 ≥ 10 ($n = 117$), where intervals excluded zero from 9 months onward. In the overall PHQ-9 population ($n = 625$), intervals overlapped zero at all time points.

General pain intensity showed the same baseline-severity gradient (Fig. 3f(ii)). With baseline pain ≥ 4 ($n = 266$), intervals overlapped zero at every time point; with baseline pain ≥ 7 ($n = 160$), reductions reached -0.97 points at 12 months (95% CI -1.93 to -0.01), with intervals excluding zero from 6 months onward.

Waist circumference declined at every time point ($n = 99$; Fig. 3f(iii)), by -2.68 inches at 12 months (95% CI -4.60 to -0.76). No clear longitudinal changes were observed for alcohol use or muscle strength.

Discussion

We describe an end-to-end approach using LLMs and recursive, self-improvement AI agents to accurately convert EHR medical text into computable clinical elements, organize them within patient-centered KGs and support individual patient journey exploration and large population analyses. The principal advantage of this approach is broad capture and analyses of clinically salient information at scale. Clinical history, disease context, adherence barriers, nuanced symptom descriptions and family history details are often incompletely represented in coded fields; making these elements computable and linking them to standardized concepts greatly expands the range of questions that clinicians and researchers can reliably ask of routine-care data at large scale across multiple clinical conditions.

For the pipeline to be clinically useful, it must be trustworthy, transparent and traceable to source documentation. This is essential for AI generative models, which may introduce unsupported statements or hallucinate false information. In clinical text, both false inclusion and precision matter: the distinction between ‘chest pain’ and ‘denies chest pain’ is central to valid inference. We therefore designed the extraction framework around verifiability, provenance-linked atomic statements, adjudication against source text and calibrated thresholds to make outputs reviewable rather than opaque.

Structuring extractions as minimal, provenance-linked ‘atomic statements’ is critical for making LLM-generated outputs verifiable and auditable. Atomic statements (single propositions with explicit evidence spans, timestamps and assertion status) impose a disciplined interface between generation and verification. Evaluation metrics and verifiers operate more reliably on atomic units, as evidenced by per-claim assessments of fidelity between a generated proposition and its source evidence, such as FactScore or dedicated verification workflows (for example, Verifact)²² and processes to atomic parts can achieve perfect accuracy over millions of steps²³.

Combining per-statement judgments with probability calibration and conformal thresholds provides confidence measures for each extracted concept and finite-sample guarantees¹⁰. Our approach enables safer deployment of generative AI in healthcare by providing clinicians with interpretable confidence measures and formally verifying that the proportion of incorrect accepted extractions is statistically bounded.

We demonstrated our model-extracted entities were validated through a blinded, multiple expert physician-adjudicated references yielding high accuracy confirming the clinical validity of the framework at speed and cost efficiency required for large longitudinal datasets.

Our use of graph database organization and storage, combined with self-improvement AI agents, allows navigation across the KG computable elements at scale and speed for hypothesis testing and exploration. The KG approach emphasizes explicit relationships and

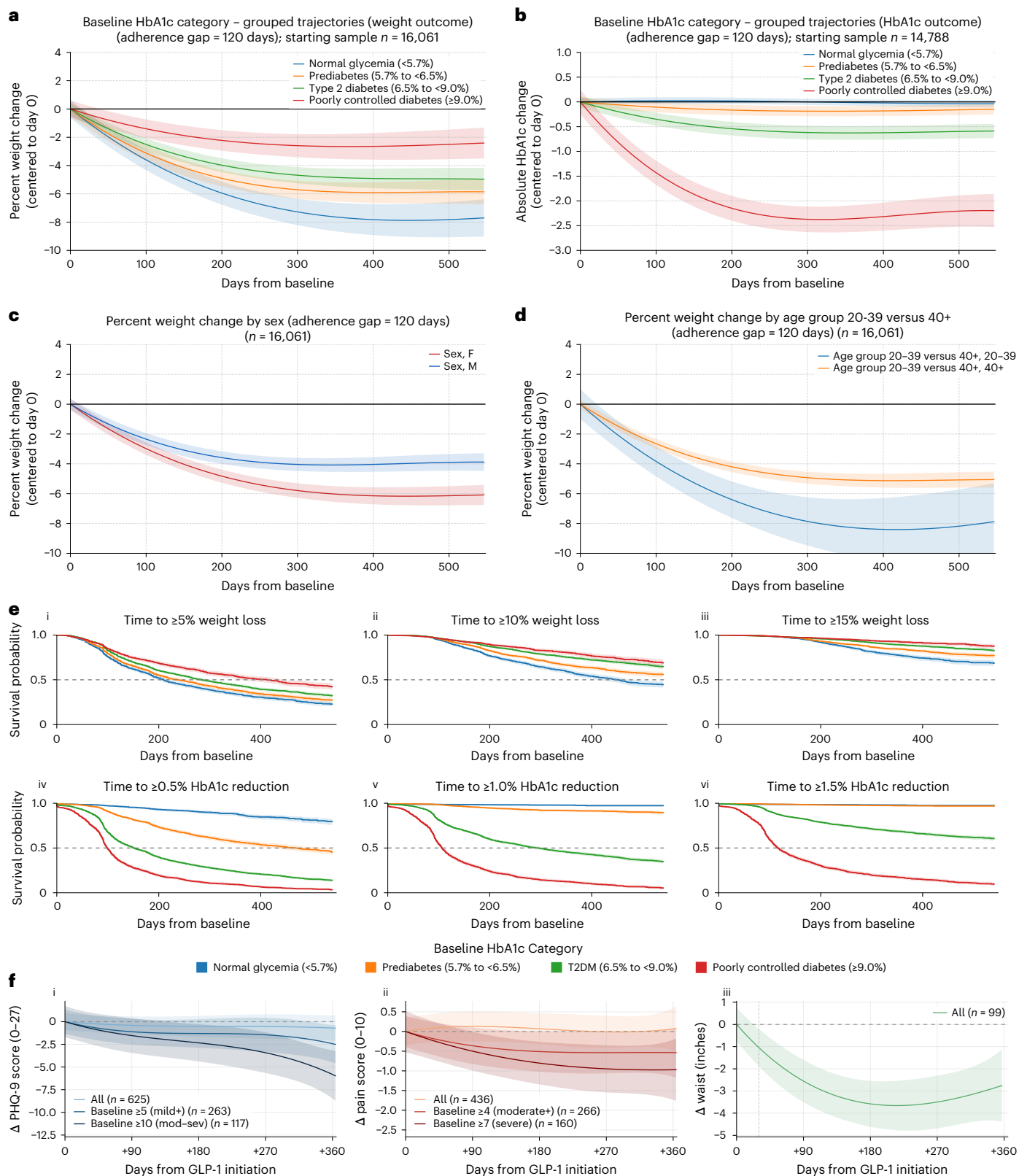
Fig. 3 | Longitudinal weight, HbA1c and note-derived assessment trajectories following GLP-1 receptor agonist initiation. **a**, Weight change (%) from day 0 stratified by baseline HbA1c category (normal glycemia, $<5.7\%$; prediabetes, 5.7 to $<6.5\%$; T2DM, 6.5 to $<9.0\%$; and poorly controlled, $\geq 9.0\%$). **b**, Absolute HbA1c change (percentage points) from day 0 by the same baseline categories. **c**, Weight change stratified by sex (female, $n = 9,787$; male, $n = 6,274$). **d**, Weight change stratified by age group (20–39 years, $n = 1,103$; ≥ 40 years, $n = 14,958$). **a–d** show model-predicted mean trajectories over 540 days of follow-up, estimated in separate generalized estimating equation models fitted within each stratum with robust s.e. and a B-spline basis for time (3 d.f.). **e**, Unadjusted Kaplan–Meier estimates of time to clinically meaningful weight loss (i, $\geq 5\%$; ii, $\geq 10\%$; iii, $\geq 15\%$) and HbA1c reduction (iv, ≥ 0.5 point; v, ≥ 1.0 point; vi, ≥ 1.5 points), stratified by baseline HbA1c category; adjusted hazard ratios from Cox

proportional hazards models are given in Extended Data Table 2. **f**, Model-predicted mean change from each patient’s own baseline over 12 months in note-derived clinical assessments, estimated by generalized estimating equations with a B-spline basis for time (3 d.f.): i, depressive symptoms (PHQ-9; all patients, $n = 625$; baseline PHQ-9 ≥ 5 , $n = 263$; baseline PHQ-9 ≥ 10 , $n = 117$); ii, general pain intensity (all patients, $n = 436$; baseline pain ≥ 4 , $n = 266$; baseline pain ≥ 7 , $n = 160$); iii, waist circumference ($n = 99$); in **f**, n denotes patients with an eligible baseline assessment and at least one follow-up assessment. Primary analysis used a GLP-1 RA persistence gap of 120 days (**a–e**) and exploratory analyses in **f** were conducted without a persistence requirement (Methods). Shaded bands in **a–d** and **f** are 95% CI for the model-predicted mean at each time point, referenced to day 0; bands in **e** are 95% exponential Greenwood intervals. All tests are two-sided Wald tests and no adjustment was made for multiple comparisons.

temporality. By integrating validated extraction from clinical notes with native structured EHR data, grounding elements to standard ontologies and organizing them within a patient-centered KG, we reconstruct clinical histories in a form that is analytically useful and traceable to source documentation over time.

Medication and symptom events are linked not only to patients and visits, but also to medication concepts, enabling direct reasoning

about drug–symptom associations. These design choices facilitate multi-hop queries that traverse linked events, concepts and time-points, and downstream machine learning in ways that respect clinical pathways. The resulting process preserves the logic of conventional clinical review workflows across narrative text at scale, with extracted elements, intermediate outputs and analytic code available for investigator or clinician review. The result is not a black-box system that



substitutes for clinical or analytic judgment, but a human-guided environment in which labor-intensive tasks are accelerated, while query design, analytic decisions and interpretation remain under investigator or clinician control.

Clinical application of this KG approach to our GLP-1 RA cohort revealed longitudinal patterns in weight and glycemic response that would be difficult to detect using structured data alone or single-timepoint analyses. Information needed to understand treatment exposure, symptoms, adherence, clinical assessment and disease evolution often resides in narrative text rather than structured fields²⁴. In our cohort, over two-thirds of HbA1c values, over one-third of cohort patients and over one-third of GLP-1 RA medication usage were only identified only by unstructured data.

GLP-1 RA induced weight loss was greater and more rapid in individuals with lower baseline glycemic burden (HbA1c levels); conversely, GLP-1 RA induced glycemic control (HbA1c levels) improved faster in patients starting with higher glycemic burden. Although previous studies have reported differential GLP-1 RA effects across glycemic groups, most relied on selected populations or baseline-to-follow-up comparisons at fixed timepoints^{25,26–30}. These longitudinal patient analyses could be clinically employed to develop individualized care pathways to guide GLP-1 RA dose escalation, switching among GLP-1 RAs and intensive complementary interventions.

Furthermore, we were able to examine clinically relevant longitudinal patterns in depression, pain scores and waist circumference after GLP-1 RA initiation, variables that would otherwise require large-scale manual human chart review. Patients with moderate to severe depression improved after GLP-1 RA initiation, consistent with an effect currently under evolving research^{31,32}. Pain intensity was reduced, especially among patients with higher pain burden, supporting potential benefits of GLP-1 RAs on neuroprotective, metabolic and inflammatory regulation³³.

This work should be interpreted considering several limitations. The GLP-1 RA study example is observational and inherits the familiar constraints of routine-care data, including irregular measurement, incomplete capture and variable follow-up. Although the extraction pipeline was extensively validated, the strength of this approach lies in traceability and reviewability. The analysis was also conducted across multiple care systems and environments, and the cohort reflects differing local documentation practices, patient mixes and prescribing patterns. Estimates reflect an on-treatment estimand with follow-up censored at loss of evidence of ongoing therapy. Medication discontinuation may depend on treatment response, yet sensitivity analyses show results were stable across persistence definitions. Threshold attainment could be observed only at clinical encounters, so differences in measurement frequency across strata could in principle influence apparent timing. It should be noted that these exploratory estimates were conducted as single-arm within patient comparisons; however, measurement frequency was similar across strata, and the observed differences were too small to account for the differences reported. Finally, richer measurement available from unstructured data does not by itself resolve confounding or other inferential challenges; causal interpretation still depends on study design, assumptions and appropriate statistical methods.

Real-world evidence nevertheless complements rigorously conducted clinical trials by showing how treatments are prescribed, experienced and monitored in routine care. Routine-care data can also reveal patterns of care, including outcomes measured and overlooked. By making these patient journeys computable, reviewable and queryable, this platform can help refine hypotheses, sharpen subgroup definitions and inform future prospective studies.

As AI is increasingly used to develop more sophisticated tools for clinical research and care, including simulation, causal and predictive modeling, and point-of-care decision support, the promise will depend on the fidelity of the patient histories from which they learn. If narrative

text is not extracted accurately and easily computable, then model systems will have incomplete or misleading accounts of care that shapes the models themselves, and the quality of their inferences, predictions and recommendations. By making note-derived clinical information accurately computable, reviewable and traceable to source evidence, this platform offers a more complete foundation for both translational research and future clinical analyses at scale.

There are a myriad of opportunities for facilitated EHR review. It has been estimated that clinicians spend an average of 5–16 min in chart review of each new patient³⁴, 30 min for clinical research patient chart reviews³⁵ and greater than 4 h per patient for regulatory compliance or medicolegal review³⁶. In our study, the five expert physician reviewers required approximately 70 h to complete the review of the 120 charts across all of the clinical domains. In contrast, our KG-based AI model accurately reviewed each of the GLP-1 RA charts in seconds.

In conclusion, our study presents an end-to-end approach that enables the medical record to be reliably interrogated across multiple dimensions at speed and scale. The GLP-1 RA analyses are a concrete clinical example of broader trajectory-level insights. More broadly, the platform highlights the importance of extraction accuracy and data provenance as AI becomes increasingly embedded in clinical research and care. The longer-term value of this trustable, transparent and traceable approach may extend beyond the description of real-world care experience, providing a stronger foundation for future predictive, causal and decision-support applications within clinical research and care.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-026-04695-x>.

References

- Kong, H.-J. Managing unstructured big data in healthcare system. *Healthc. Inform. Res.* **25**, 1–2 (2019).
- Ford, E., Carroll, J. A., Smith, H. E., Scott, D. & Cassell, J. A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J. Am. Med. Inform. Assoc.* **23**, 1007–1015 (2016).
- Golder, S., O'Connor, K., Lopez-Garcia, G., Tatonetti, N. & Gonzalez-Hernandez, G. Leveraging unstructured data in electronic health records to detect adverse events from pediatric drug use: a scoping review. *Annu. Rev. Biomed. Data Sci.* **8**, 227–250 (<https://doi.org/10.1146/annurev-biodatasci-111224-124530>) (2025).
- Mohan, G. & Gaskin, D. J. Social determinants of health and US health care expenditures by insurer. *JAMA Netw. Open* **7**, e2440467 (2024).
- Savova, G. K. et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **17**, 507–513 (2010).
- Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* **2001**, 17–21 (2001).
- Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
- Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
- Yang, X. et al. A large language model for electronic health records. *npj Digit. Med.* **5**, 194 (2022).
- Kim, E., Foty, R., Shrestha, M. & Seyfert-Margolis, V. Conformal prediction and verification of large language model extractions in EHR data. In *Proc. AAAI 2025 Fall Symposium Series*, Vol. 7 (eds Petrick, R. & Geib, C.) 539–546 (PKP Publishing Services, 2025).

11. Kim, E., Shrestha, M., Foty, R., DeLay, T. & Seyfert-Margolis, V. Structured extraction of real world medical knowledge using LLMs for summarization and search. In *Proc. 2024 IEEE International Conference on Big Data (BigData)* 3421–3430 (IEEE, 2024).
12. Wei, Y., Li, Q. & Pillai, J. Structured LLM augmentation for clinical information extraction. *Stud. Health Technol. Inform.* **329**, 971–976 (2025).
13. Arzideh, K. et al. From BERT to generative AI - comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Comput. Biol. Med.* **195**, 110665 (2025).
14. Schaye, V. et al. Large language model-based assessment of clinical reasoning documentation in the electronic health record across two institutions: development and validation study. *J. Med. Internet Res.* **27**, e67967 (2025).
15. Shinohara, E. & Kawazoe, Y. Efficient medical NER with limited data: enhancing LLM performance through annotation guidelines. *Int. J. Med. Inform.* <https://doi.org/10.1016/j.ijmedinf.2025.106230> (2025).
16. Thio, S., Lewis, M., Denaxas, S. & Dobson, R. J. Unlocking electronic health records: a hybrid graph RAG approach to safe clinical AI for patient QA. *Front. Digit. Health* **8**, 178077 <https://doi.org/10.3389/fdgh.2026.1780700> (2026).
17. Girman, C. J., Ritchey, M. E. & Re, V. L. III Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiol. Drug Saf.* **31**, 717 (2022).
18. Uzuner, Ö., Solti, I. & Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* **17**, 514–518 (2010).
19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, (eds Burstein, J. et al.) 4171–4186 (ACL, 2019).
20. Helios9/BioMed_NER. *Hugging Face* https://huggingface.co/Helios9/BioMed_NER
21. Hogan, A. et al. Knowledge graphs. *ACM Comput. Surv.* **54**, 1–37 (2021).
22. Chung, P. et al. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI* **3**, AIdbp2500418 <https://doi.org/10.1056/AIdbp2500418> (2025).
23. Meyerson, E. et al. Solving a million-step LLM task with zero errors. Preprint at <https://arxiv.org/abs/2511.09030> (2025).
24. McMurry, A. J. et al. Large language model symptom identification from clinical text: multicenter study. *J. Med. Internet Res.* **27**, e72984 (2025).
25. Kahn, S. E. et al. Effect of Semaglutide on Regression and Progression of Glycemia in People With Overweight or Obesity but Without Diabetes in the SELECT Trial. *Diabetes Care* **47**, 1350–1359 (2024).
26. Cerchi, E. et al. Effects of tirzepatide on weight management in patients with and without diabetes: a systematic review and meta-analysis. *Int. J. Obes.* **49**, 2415–2425 (2025).
27. Rodriguez, J. et al. Semaglutide vs tirzepatide for weight loss in adults with overweight or obesity. *JAMA Intern. Med.* <https://doi.org/10.1001/jamainternmed.2024.2525> (2024).
28. McGinnis, T. et al. Clinical effectiveness of semaglutide for weight loss in a veterans affairs' anti-obesity pharmacotherapy clinic. *J. Gen. Intern. Med.* <https://doi.org/10.1007/s11606-025-09943-3> (2025).
29. Frías, J. P. et al. Tirzepatide versus semaglutide once weekly in patients with type 2 diabetes. *N. Engl. J. Med.* **385**, 503–515 (2021).
30. Misra, S. Deciphering the effects of semaglutide across the glycemic spectrum. *Diabetes Care* **47**, 1322–1324 (2024).
31. Chen, X., Zhao, P., Wang, W., Guo, L. & Pan, Q. The antidepressant effects of GLP-1 receptor agonists: a systematic review and meta-analysis. *Am. J. Geriatr. Psych.* **32**, 117–127 (2024).
32. Anderberg, R. H. et al. GLP-1 is both anxiogenic and antidepressant; divergent effects of acute and chronic GLP-1 on emotionality. *Psychoneuroendocrinology* **65**, 54–66 (2016).
33. He, Y. et al. Advances in GLP-1 receptor agonists for pain treatment and their future potential. *J. Headache Pain* **26**, 46 (2025).
34. Agatstein, K. Chart review is dead; long live chart review: how artificial intelligence will make human review of medical records obsolete, one day. *Popul. Health Manag.* **26**, 438–440 (2023).
35. Siems, A. et al. Structured chart review: assessment of a structured chart review methodology. *Hosp. Pediatr.* **10**, 61–69 (2020).
36. Bradshaw, R. W. Using peer review for self-audits of medical record documentation. *Fam. Pract. Manag.* **7**, 28–32 (2000).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹RespondHealth, Bethesda, MD, USA. ²Department of Computer Science, Howley College of Engineering and Computing, Drexel University, Philadelphia, PA, USA. ³Department of Radiology, School of Medicine, Stanford University, Palo Alto, CA, USA. ⁴Diabetes Research Institute, University of Miami Miller School of Medicine, Miami, FL, USA. ⁵Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, Philadelphia, PA, USA. ⁶Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁷Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁸Department of Emergency Medicine, College of Medicine, Drexel University, Philadelphia, PA, USA. ⁹These authors contributed equally: Edward Kim, Richard Foty. ¹⁰These authors jointly supervised this work: Charles B. Cairns, Vicki Seyfert-Margolis. ✉e-mail: vicki@respondhealth.net

Methods

Platform architecture and technical framework

Our technical framework is composed of several key elements: an extraction process that utilized LLM's to capture unstructured data from EHR notes; assembly of a KG that integrates both structured and unstructured data elements; and mapping of the data elements to known ontologies when possible; creating of data-driven ontologies; and use of agentic AI for longitudinal data analysis at scale.

Structured inputs included patient demographics, visits, diagnoses, procedures, laboratory measurements, vital signs and medication prescriptions, which were grounded to standard terminologies before graph construction. In parallel, an LLM-based extraction pipeline processed unstructured clinical notes to derive current computable elements. Extraction entities in following categories were defined as follows: medication (current and prescribed medications and medication events such as side effects, discontinuations, switches and dose changes); self-reported symptoms (patient-reported); phenotypes (physician-observed symptoms and disease characteristics); adherence (patient adherence to diet, exercise and medications as well as their barriers to adherence); family history (relation to patient and diseases); hospitalizations (mentions of hospitalizations and reason for hospitalization); quantitative assessments (labs, surveys, vitals and any other quantitative instrument); and disease-status (diseases grounded in ICD-10 codes). Extracted elements were mapped to standard concept systems, see the 'Ontology mapping' section below.

When existing ontologies did not provide adequate coverage, data-driven concept-discovery methods were used, see the 'Concept mapping' section below. The KG served as the shared computational representation linking atomic statements, higher-level concept groupings, patients and curated ontology concepts. Within this graph, nodes encoded atomic statements and higher-level concept groupings, whereas edges captured temporal and semantic relationships together with links to external ontologies. Patient nodes represented longitudinal patterns over time, enabling reconstruction of individual care trajectories from structured data and note-derived events within a single framework. The framework was designed to support trust, transparency and trackability in downstream research use. Accordingly, source-linked provenance was retained throughout so that extracted statements, graph elements and downstream analytic variables could be traced to supporting note text, encounter context, ontology mappings and investigator-reviewed intermediate outputs and derived code before interpretation.

The sections that follow describe data harmonization, note extraction, ontology mapping, KG construction and longitudinal analytic use.

Data source

Data for this study were obtained from the Sidus Insights (Niagara Falls, NY, USA) real-world data platform. Upstream normalization to standard clinical terminologies was applied to support cross-site aggregation. All analyses were conducted on de-identified data in accordance with applicable data-use agreements.

Agent-assisted longitudinal note extraction

A core component of the framework was agent-assisted extraction from longitudinal clinical notes with preservation of source-linked evidence for review. LLMs were used with structured generation and domain-specific prompting to derive computable clinical information from unstructured text. Targeted extraction schemas were defined for each domain. For medication-related data, the extraction process captured medication names together with associated side effects, adverse events, dose changes, initiations, discontinuations and switches³⁷. Verification was implemented at multiple stages, as described below (Extended Data Fig. 2). The named entity extraction pipeline used LLMs to identify clinical information from unstructured notes and map extracted elements to standardized ontologies. In previous work,

we benchmarked encoder-based BERT models against decoder-based LLMs and found that specialized encoder models outperformed more general LLMs on conventional NER benchmarks¹¹. For the present study, however, LLMs were used for extraction from real-world EHR notes, particularly physician documentation, because they more reliably resolved contextual meaning in narrative text.

This was particularly important for negation handling. For example, in the statement 'He denies vision problems, chest pain, trouble breathing, stomach or bowel problems, urine problems, muscle aches, new extremity swelling, sleep problems, or mood or energy problems,' the extraction process needed to assign the full coordinated list to the negated context to avoid spurious entities and false-positive phenotypes. Time-limited extractions (medical entities that are nonhistoric in nature or described at the current visit date) were also challenging for NLP NER models especially given the large quantity of copy paste in the visit notes. LLMs were also used to resolve shorthand notation and abbreviations common in clinical documentation and to identify clinically relevant information implied by context rather than stated using standardized terminology. Because some extracted elements were explicitly documented whereas others were inferred from context, each output was assigned an evidentiary support level to distinguish directly recorded findings from context-based inferences.

Temporal modeling: from snapshots to trajectories

Because EHR notes are longitudinal, extraction was designed to account explicitly for change over time. In previous work, we quantified the amount of new information contributed across sequential notes and found that, after the fourth visit, approximately 13% of note content was newly introduced³⁸. In practice, repeated text may reflect either copy-forward documentation or persistence of clinically relevant findings. The extraction pipeline therefore prioritized differences between the current and previous visits so that newly documented or materially updated information could be distinguished from persistent findings. Extracted events were aligned to the encounter timeline, with onset, duration and resolution captured when stated. Contradictory or regressing assessments were retained to represent evolving clinical states.

Temporality was implemented using a visit-to-visit change-tracking approach. At each encounter, the current note was compared to the immediately preceding note and with the previous stored extraction. Findings were labeled as added when newly documented, modified when there was a meaningful change in presence, severity or context, and removed when no longer documented; stable items were carried forward without duplication. Explicit negation was retained by preserving the underlying concept label and recording negation status separately. When severity was stated, it was recorded as an attribute (mild, moderate or severe). Source-linked provenance was maintained by retaining the supporting evidence string from the source note, enabling direct traceability of each extracted value to its origin (Extended Data Fig. 3).

Ontology mapping

To support interoperability and standardized representation of extracted events, each atomic statement was grounded to established biomedical ontologies, including UMLS, SNOMED CT, HPO, ICD, LOINC and RxNorm, with assignment of stable identifiers and explicit semantic relationships. Ontology grounding served three purposes. First, it linked note-derived assertions to curated concepts while preserving source-linked provenance. Second, it connected patient-level observations to external knowledge resources for downstream interpretation and evidence synthesis. Third, it enabled reuse across datasets through standard concept identifiers and semantic relationships. This alignment of extracted statements with reference terminologies supported consistent querying, validation and integration of note-derived and structured EHR data within the patient-centered KG.

Concept discovery

Because standard medical ontologies did not adequately capture some patient-centered concepts, including adherence barriers, reasons for medication titration and nuanced symptom descriptions, we developed a data-driven taxonomy discovery pipeline to organize free-text extractions into hierarchical categories. This approach was applied to seven event domains: medication symptoms, medication discontinuations, medication switches, dose change reasons, adherence barriers, adherence recommendations and hospitalization reasons.

Concept discovery: computational pipeline for unsupervised taxonomy generation

To operationalize concept discovery, we implemented a computational pipeline to discover and organize unstructured clinical motifs into interpretable taxonomies. We processed seven distinct event categories where free-text extractions were allowed by the LLM: medication symptoms, medication stops, medication switches, dose change reasons, adherence barriers, adherence recommendations and hospitalization reasons. The pipeline begins by extracting short text spans from verified note-derived events (for example, 'upset stomach' from a medication side-effect event). These text spans are vectorized using OpenAI text embeddings to capture semantic similarity.

We then applied k -means clustering to these high-dimensional vectors. To determine the optimal number of clusters (k), we performed a sweep over a range of k values (typically 15–50), selecting the optimal k based on the maximum silhouette score, while also monitoring Inertia (the 'elbow method') to ensure compact and well-separated clusters.

Once clusters were defined, we used an LLM to generate clinically meaningful labels. For each cluster, the top 50 most frequent text examples were prompted to the LLM to generate a concise, hierarchical category label (for example, 'gastrointestinal side effects'). A second LLM pass then grouped these fine-grained cluster labels into higher-level parent categories forming a two-level taxonomy.

Finally, to visualize and validate the semantic structure, we projected the high-dimensional embeddings into two-dimensional (2D) space using Uniform Manifold Approximation and Projection (UMAP) with cosine metric. We performed extensive parameter sweeps on n neighbors (varying from 0.1% to 10% of the dataset size) and min dist (0.0–0.4) to assess the stability of the projections and ensure that the visual clusters corresponded to the discovered semantic groupings. Centroids and representative examples for each cluster were serialized to enable runtime classification of new data points. This procedure yields concise, clinically meaningful groupings that can be reused as code-like filters in a downstream KG. Figure 2a shows an example taxonomy generated for medication switching. In summary, the taxonomic building stage enables a reproducible path to concept discovery, surfacing data-driven patient-reported motifs and barriers that complement structured concepts, while preserving traceable provenance to the originating notes.

Verification of extracted statements

Unstructured EHR text was converted into structured atomic statements across multiple clinical domains, including self-reported symptoms, phenotypes, diseases, medications, quantitative assessments, family history, adherence and hospitalizations. Each statement was linked to supporting source-note evidence and assigned verification metadata.

Evidentiary support was assessed using a four-level framework: fully supported, partially supported, inferred or unsupported. Fully supported statements were explicitly justified by the cited text; partially supported statements were substantively correct but incomplete or imprecise; inferred statements were plausible from context but not explicitly stated; and unsupported statements lacked adequate source evidence. Support rates were summarized by entity category.

A verification score of 3 (fully supported) denoted a non-ambiguous extraction with clear support in the source text.

Verification details: partial and inferred support in extracted statements

To better understand the verification of score <3 outputs, we clustered verifier rationales using the same concept-discovery pipeline described above (embedding, clustering and cluster labeling). Two recurring patterns dominated.

For score 2 (partially supported) statements, the underlying clinical finding was typically present in the note, but the exact modifier required by the atomic claim was missing or only weakly implied. Most often, documentation supported the symptom or diagnosis itself, while the extracted statement additionally asserted a qualifier such as severity (for example, mild/moderate/severe), status (for example, stable/improved/worsened), laterality, temporality (new versus chronic) or a precise linkage between a finding and its presumed cause or condition. A second common source of partial support was ambiguous or conflicting documentation, where the record contained mixed signals across sections (for example, review of systems versus HPI/assessment) or across adjacent visits, making an absolute assertion only conditionally true. Finally, a subset of partially supported cases reflected duplication and representation mismatches: near-duplicate extractions or discrepancies between problem-list codes and encounter text (or between coding systems) that preserved the clinical concept but weakened support for the specific coded or qualified phrasing.

For score 1 (inferred) statements, the most frequent pattern was absence by implication rather than explicit negation. In these cases, the extraction asserted a negative finding (for example, 'denies X') even though the note provided only indirect support (for example, 'psych normal,' 'no abnormalities reported' or omission of the symptom from a Review of Systems (ROS) list). Inferences were also triggered by specificity gaps in negation (for example, denial of a general symptom without explicitly denying a more specific variant) and by internal or longitudinal inconsistencies that undermined blanket negative statements (for example, an ROS denial paired with affirmative documentation elsewhere in the same note, or a denial at the current visit contradicting findings recorded in nearby encounters). Together, these findings indicate that most lower-scoring outputs arise not from spurious concept identification, but from over-specific or over-confident modifiers and negation phrasing that are not stated verbatim in the source documentation.

Expert validation of extracted statements

To assess the clinical validity of the extraction framework, expert physician reviewers judged model-extracted entities against the original clinical note using a purpose-built interface to determine how faithfully our extraction process captured facts documented at each current patient visit³⁹. The interface displayed the full source note alongside each extracted statement and its supporting text span. Correctness was judged against the source note under a rule set fixed before evaluation, with criteria specific to each clinical domain, including symptoms, phenotypes, diseases and conditions, medications and quantitative assessments (Supplementary Note 1).

Evaluation visits were sampled from the Sidus Insights real-world data platform, spanning four ambulatory EHR systems, further described in the data source section below. Five visits were reviewed first to calibrate interpretation of the scoring rules; calibration met the prespecified thresholds (entity F1 ≥ 0.90 and Gwet's AC1 ≥ 0.80).

Physician reviewers manually scored the statements produced by our extraction model against the source note, in three independent domains: medications, phenotypes and quantitative assessments. A primary physician reviewer scored 100 randomly selected visits, spanning 96 providers and 86 practices across the four distinct EHR systems. A second expert physician reviewer independently scored a 20-visit subset to establish reliability of the clinical chart review of 3,303

extracted elements⁴⁰. Each statement was judged correct or incorrect against the note, and reviewers additionally flagged documented findings the model had missed, supporting a recall estimate. Within the 20-visit subset, 32 correct statements were withheld before review to test for physician reviewer automation bias^{41,42}.

Disagreements on the 20-visit subset were resolved by a third expert physician, blinded to reviewer identity, producing an adjudicated reference. Reviewer concordance was summarized using percent agreement, positive and negative specific agreement, and Gwet's AC1; because correct verdicts far outnumbered incorrect ones, we did not rely on Cohen's κ , which is deflated under high agreement⁴³. Model performance was reported against the primary reviewer against the adjudicated reference on the subset, and the primary reviewer's independent subset in each case as precision, recall and F1, with phenotype recognition additionally scoring assertion polarity. Reviewers and the adjudicator were all board-certified physicians from independent institutions.

Comparison with other extraction models

As a secondary analysis, we benchmarked our extraction pipeline against existing NLP extraction models, existing publicly available gold standard datasets, and different distributions of patient visits. On 20 additional charts (2,710 reviewed extractions), using the same review interface and adjudication process described above with two additional expert reviewers. The review set included outputs from the primary LLM-based extraction pipeline and comparator models for medication and phenotype extraction, including PhenoTagger and BERT (large) as non-LLM baselines. Review outputs were presented model by model so that adjudication could be performed on the extracted statements generated by each system for the same underlying note set.

Reviewers evaluated each extracted statement against the underlying chart evidence and then advanced to the next note. This process was repeated across extraction domains and across the candidate models under comparison.

Reviewed phenotype statements were assessed for correctness of the phenotype concept and correctness of assertion status, including present versus negated or absent status. Reviewed medication statements were assessed for support in the note, including attributes such as current therapy, discontinuation, switch, dose change, side effect or adverse event where applicable. Reviews were performed independently.

Agreement and model performance were summarized after adjudication. Inter-rater agreement was quantified using percent agreement and Cohen's κ , appropriate here because cross-model verdicts were more balanced than in the single-model validation, where we used Gwet's AC1. Model-level performance was evaluated using strict accuracy as the primary metric. For phenotype extraction, concept recognition was also evaluated with assertion polarity ignored to isolate errors related to negation and assertion handling.

On existing datasets, we benchmarked model variants on the BiolarkGSC+ validation set, a manually annotated corpus labeled with HPO terms (Extended Data Table 5)⁴⁴. Performance was assessed using document-level macro metrics based on HPO name matching, as described previously⁴⁵. These experiments informed three complementary optimization strategies: targeted fine-tuning or distillation, task-specific post-processing and multimodel aggregation (described below; Extended Data Fig. 4 and Extended Data Table 5).

Optimization of LLM-based extraction

Because the extraction workflow required repeated model calls across longitudinal patient records, optimization was performed to improve scalability while maintaining extraction quality. In our implementation, the pipeline averaged 300–400 LLM calls per patient record, reflecting 20–30 visits, seven modular extractors and atomic verification. At the scale of the full dataset, use of the largest reasoning models would have imposed substantial computational time and token cost, making

direct deployment impractical. Optimization was therefore required not only for extraction performance, but also for throughput and cost control, which is for large-scale extractions, we utilize the 'mini' versions of proprietary models and quantify our supervised fine tuning on the distilled open weight models.

Extraction optimization: fine-tuning, post-processing and multimodel aggregation

Optimization 1: extraction via fine-tuning. To improve scalability, we fine-tuned compact decoder-based extraction models on task-specific annotation sets. Instruction-following decoder models, including GPT-4.1 Mini, GPT-4.1 Nano and the open-source Qwen 2.5 4B model, were adapted using low-rank fine-tuning on a Fine-Tuning Cohort of 100,000 de-identified visits drawn from the 1 million-visit extraction sample (Fig. 1). Distillation labels for extraction were generated using GPT-5 Mini.

Azure OpenAI fine-tuning. Fine-tuning of the Azure GPT models was performed in three sequential phases of continual adaptation. All phases used the same task hyperparameters: batch size 16, learning rate multiplier 0.1 and one epoch. Because the underlying model weights were closed, parameter-efficient adaptation was implemented through the provider's fine-tuning interface rather than direct control of low-rank adaptation (LoRA). Cumulative billed training totaled 354 million tokens, and cumulative wall-clock training time was 13 h 09 min for the Nano model and 22 h 58 min for the Mini model. Fine-tuned models showed improved performance over the corresponding base models for phenotypic entity extraction.

Extended Data Fig. 4 illustrates the training dynamics for the GPT-4.1 Mini and Nano models during the LoRA fine-tuning process. Both models exhibit a steep initial reduction in training loss, effectively converging within the first 7,500 steps. This rapid stabilization suggests that the low-rank adaptors were able to efficiently learn the task-specific distribution of the clinical extraction data without requiring extensive epochs. The 50-step moving average (blue curve) highlights the consistent downward trend despite expected mini-batch variance. These loss trajectories align with our observation that targeted fine-tuning on a high-quality, verified subset of 100,000 visits is sufficient to adapt general-purpose reasoning models to the specific nuances of clinical entity extraction, balancing computational efficiency with performance gains.

Open-source Qwen 4B fine-tuning. In parallel, we fine-tuned an open-source 4B-parameter Qwen instruction model for phenotype extraction using parameter-efficient LoRA adaptors. Training was performed on four A40 GPUs with mixed bfloat16 precision. LoRA was applied to all target layers with rank 16, scaling factor 32 and dropout 0.05. Optimization used AdamW with a cosine learning rate schedule and an initial learning rate of 2×10^{-4} . Total wall-clock training time was 3 days and 1 h.

On the BiolarkGSC+ benchmark, supervised fine-tuning using labels distilled from GPT-5 Mini improved recall and overall F1 in compact models intended for high-throughput deployment (Extended Data Table 5). In this setting, the principal effect of fine-tuning was reduction of systematic false negatives while maintaining acceptable precision. Fine-tuned compact models also reduced inference time and token cost relative to larger reasoning models, supporting practical deployment of the extraction workflow at dataset scale.

Listings 1 and 2 in Supplementary Note 2 show the prompts used for HPO phenotype extraction and longitudinal delta extraction of HPO-mapped phenotypes across sequential clinical notes, respectively.

Optimization 2: task-based post-processing. Task-specific prompt refinement and post-processing were used to improve benchmark-aligned extraction performance from off-the-shelf models. In our

implementation, a lightweight post-processing algorithm was applied to filter selected LLM outputs that were clinically meaningful but fell outside the annotation scope of the BiolarK^{GSC+} benchmark. These included terms such as syndromes, mutations, deletions, genes and allelic variants, which could be extracted appropriately in clinical text but were not always counted as valid outputs under benchmark labeling rules. For this benchmark task, filtering improved precision with minimal effect on recall. Detailed benchmark comparisons are provided in Extended Data Table 5.

Optimization 3: multimodel aggregation. Different LLMs, trained on different data mixtures and optimized with different objectives and alignment procedures, exhibited distinct failure modes. We therefore evaluated aggregation across strong but diverse models as an additional optimization strategy. As shown in Extended Data Table 5, the combined Sonnet 4.5 + GPT-5.2 output increased coverage relative to individual off-the-shelf models, achieving recall of 0.95. When coupled with task-specific post-processing to filter benchmark-specific false positives, this multimodel approach achieved an F1 of 0.90, second only to single-model Sonnet 4.5 (F1 = 0.92). In practical terms, aggregation improved recall because different models tended to miss different subsets of phenotypes, including omissions related to shorthand, template-based phrasing or negation scope. Aggregation was therefore used to improve robustness without relying on the error profile of any single model. Different optimization strategies involved different trade-offs among accuracy, throughput and cost, and should be selected according to the priorities of the clinical application.

Knowledge graph integration and analytic environment

Following extraction, temporal reconciliation, ontology grounding and taxonomy generation, structured EHR data and note-derived computable elements were integrated into a patient-centered KG linking patients, encounters, clinical concepts and extracted entities with attributed atomic statements within a single longitudinal representation. This graph served as the shared computational substrate for reconstruction of individual patient trajectories and for downstream cohort-level analysis across structured and note-derived domains. Study-specific cohorts were subsequently defined within this graph-based analytic environment, as described in the following sections.

The KG was made available within a VS code-based analytic environment for reviewable end-to-end clinical research over integrated EHR data. Within this environment, AI research agents assisted with cohort identification, variable assembly, longitudinal summarization, retrieval from the graph, and generation of analysis-ready code and outputs. Linkage to source evidence, temporal context and ontology-grounded concepts was preserved throughout. Investigators reviewed intermediate outputs and generated code before interpretation. This environment allowed movement between patient-level review and cohort-level analysis within a shared computational representation. Agentic skills were optimized in a self-reflective manner, where analysis was performed by a combination of users and agents, and the associated thought process and response was reflected upon. Initial improvements were made by subject-matter experts for content, and then further improved by agentic AI in a self-improvement loop. The resulting skills are both optimal for analysis and technical capabilities.

Clinical application: longitudinal study of GLP-1 receptor agonist treatment response

We applied the framework described above to a longitudinal study of individuals initiating GLP-1RA therapy. Cohort construction, variable derivation and statistical analyses were performed within the KG and analytic environment described above, with AI research agents assisting retrieval, cohort specification and generation of reviewable analysis code under investigator oversight.

Study population and analytic cohorts

For development and evaluation of the extraction framework, we first restricted the database to active adult patients and then identified those with evidence of semaglutide or tirzepatide exposure during the study period using structured prescriptions and medication mentions in clinical documentation. Among these patients, we further identified those with baseline weight and HbA1c information within a prespecified window, spanning 60 days before to 14 days after treatment initiation, and sampled visits from this eligible population for extraction development and evaluation. For the clinical application, we then constructed a longitudinal cohort of adults initiating injectable semaglutide or tirzepatide who were naive to previous GLP-1 RA therapy at baseline. We excluded pregnancy during the study window, previous bariatric surgery and oral semaglutide exposure. Eligibility for outcome analyses further required numeric baseline values and at least one post-baseline measurement for the outcome under analysis, yielding separate analytic cohorts for weight and HbA1c analyses.

Data extraction and preprocessing

Within the eligible subset used for extraction development and downstream longitudinal analysis, structured clinical records and note-derived clinical concepts were generated using the integrated extraction and harmonization framework described above. These data were then used to define baseline characteristics, treatment exposure, follow-up and outcomes; downstream analytic variables were incorporated into the patient-centered KG and analytic environment described above. Anthropometric and laboratory measurements were evaluated for biological plausibility and internal consistency before analysis using conservative, prespecified criteria informed by population-based reference standards⁴⁶. Height, weight, BMI and HbA1c values outside physiologically plausible ranges, or inconsistent with internal measurement relationships, were treated as missing. When recorded BMI values were implausible but contemporaneous height and weight measurements were plausible, BMI was recalculated from height and weight. To reduce the influence of documentation artifacts and data-entry errors, short-interval weight excursions inconsistent with plausible physiological change were removed using conservative temporal consistency rules (described below).

Derived variables and covariates

Baseline HbA1c was categorized as normal glycemia (<5.7%), prediabetes (5.7% to <6.5%), T2DM (6.5% to <9.0%), and poorly controlled diabetes (≥9.0%). These categories reflect observed baseline HbA1c ranges using standard clinical cutpoints and are used for descriptive stratification; they do not represent clinical diagnoses and do not account for the effect of concurrent glucose-lowering therapy on baseline HbA1c. Baseline BMI was categorized using standard clinical cut points as underweight (<18.5 kg m⁻²), normal weight (18.5 to <25.0 kg m⁻²), overweight (25.0 to <30.0 kg m⁻²), obesity class I (30.0 to <35.0 kg m⁻²), obesity class II (35.0 to <40.0 kg m⁻²) and obesity class III (≥40.0 kg m⁻²). Age was defined at baseline and categorized for descriptive and stratified analyses.

Concomitant medication use known to affect weight was identified using a prespecified RxNorm-based code set and included as an adjustment covariate in multivariable models. Race was retained as recorded in the EHR and was included in adjusted models but was not used for primary stratification.

Patients with missing age or sex were excluded. For other covariates, missingness was retained and analytic cohorts were restricted to individuals with complete baseline covariate data; no statistical imputation was performed.

Data preprocessing and plausibility criteria

Anthropometric and laboratory measurements were evaluated for biological plausibility and internal consistency before analysis. Values

outside prespecified plausibility ranges or internally inconsistent, were treated as missing and excluded from analysis.

Weight

Weight measurements were evaluated in two passes. In the first pass, per-visit weight values were flagged as implausible if they fell below 45% or exceeded 200% of the patient's own median weight across all recorded visits (applied in patients with at least three weight measurements). The upper bound of 200% of the patient's median was selected to detect unit-conversion errors, specifically, readings recorded in kilograms in a field expecting pounds, which appear as approximately half the patient's true weight. A second rule flagged isolated spike readings: a weight measurement was removed if it deviated by more than 40% from the mean of its immediate preceding and following observations for the same patient.

In the second pass, baseline weight values were re-evaluated. A baseline measurement was flagged as implausible if it was less than 60% of the patient's median visit weight, provided the patient's median visit weight exceeded 130 lbs (to avoid flagging low-weight individuals). Flagged values were set to missing, and the corresponding percent weight change and BMI derived from those measurements were also set to missing. Patients whose baseline weight remained missing after this process were excluded from weight analyses.

BMI

When a recorded BMI value was implausible but contemporaneous height and weight measurements were both within plausible ranges, BMI was recalculated from the recorded height and weight rather than excluded. This approach recovered BMI for patients where the discrepancy arose from transcription rather than physiologically impossible values.

HbA1c

HbA1c values were evaluated against accepted physiological ranges. Values outside this range were treated as missing. Patients without a valid baseline HbA1c or without at least one post-baseline HbA1c measurement were excluded from glycemic analyses.

Data cleaning, variable derivation and cohort construction

Derived HbA1c and diabetes classification. A unified HbA1c variable was constructed by combining laboratory HbA1c and derived HbA1c values calculated from estimated average glucose (EAG) using the standard conversion:

$$\text{HbA1c}_{\text{derived}} = \frac{\text{EAG} + 46.7}{28.7}$$

For each patient and date, laboratory and derived HbA1c values were merged into a single record. Exact duplicates were removed. Values outside the physiological range (<3% or >20%) were set to missing.

Demographics, adult restriction and baseline eligibility. Demographic data were merged with longitudinal vitals. The analytic cohort was restricted to adults (≥ 18 years) at baseline. Records with missing age or missing/unknown sex were excluded.

Physiological plausibility filters were applied as follows:

- Height <48 or >90 inches
- Weight <60 or >700 pounds
- BMI <10 or >80 kg m⁻²

BMI was recalculated when necessary, using:

$$\text{BMI} = 703 \times \frac{\text{weight (lb)}}{\text{height (in)}^2}$$

GLP-1 RA medication identification and classification. GLP-1 RA medication events were mapped using curated RxNorm identifiers and classified by:

- Brand (Ozempic, Wegovy, Mounjaro and Zepbound)
- Active ingredient (semaglutide and tirzepatide)

Medication events predating US FDA approval of the corresponding agent were considered data-entry errors and excluded. Longitudinal medication histories were constructed for each patient, preserving event timing and formulation.

Baseline definition and baseline window

Baseline GLP-1 RA exposure was defined as the first qualifying GLP-1 RA medication event. A baseline window spanning -60 to +14 days around this date was used to derive baseline measurements.

Within this window, the value closest in time to the baseline date was selected for each variable (HbA1c, weight, BMI and height), with ties resolved in favor of values occurring on or before baseline.

Anthropometric cleaning and Centers for Disease Control and Prevention outlier detection

Reference-based outlier screening. Height, weight and BMI were evaluated against age- and sex-specific reference distributions derived from Centers for Disease Control and Prevention growth and adult reference data as outlined in the manuscript. Values outside extreme percentiles (approximately <1st or >99th) were flagged as implausible.

Short-interval weight spike detection. Weight trajectories were evaluated for abrupt changes. For any pair of measurements ≤ 30 days apart, a change $\geq 10\%$ triggered a spike flag. If the spike was isolated (not supported by adjacent values), the later value was set to missing.

HbA1c spike detection and cleaning. HbA1c trajectories were evaluated for implausible spikes using a rule-based approach:

For each patient, HbA1c values were sorted chronologically. For any candidate HbA1c >15%, 'neighboring values' were defined as all other HbA1c measurements within ± 365 days.

A value was classified as a spike and removed only if all of the following conditions were met:

No other HbA1c >15% existed within the ± 365 -day window, and
The candidate value exceeded the local context by:

- ≥ 5 percentage points above the median of neighboring values, or
- ≥ 4 percentage points above the second-highest value in the window (which itself must be $\leq 12\%$), or
- ≥ 5 percentage points above the nearest temporal neighbor.

If no other HbA1c values existed within ± 365 days, the value was retained (not considered a spike).

All removed values were flagged, and downstream HbA1c-based variables were recomputed accordingly.

Exclusion criteria and final analytic cohort. Sequential exclusion criteria were applied as follows:

Age <18 years or missing age/sex
Pregnancy at baseline or during GLP-1 RA exposure
History of bariatric surgery
Baseline exposure to oral semaglutide (Rybelsus).

The final analytic cohort consisted of adults initiating injectable semaglutide or tirzepatide. Patient counts were recorded at each step to support transparent cohort flow reporting. All exclusion criteria were prespecified before analysis.

Derived variables and outcome definitions. Derived variables included:

Baseline HbA1c categories (from baseline_a1c_final if available, else baseline_a1c):

- 0 to <5.7% → 'Normal glycemia'
- 5.7 to <6.5% → 'Prediabetes'
- 6.5 to <9.0% → 'Type 2 diabetes'
- ≥9.0% → 'Poorly controlled diabetes'

Baseline BMI categories:

- <25 kg m⁻² → normal/overweight reference
- 25–29.9 kg m⁻² → overweight
- 30–34.9 kg m⁻² → class I obesity
- 35–39.9 kg m⁻² → class II obesity
- ≥40 kg m⁻² → class III (severe) obesity.

Weight outcomes: percent change and achievement of ≥5%, ≥10%, ≥15% and ≥25% weight loss

Glycemic outcomes: absolute HbA1c reduction and attainment of HbA1c <6.5%

Medication exposure: GLP-1 RA monotherapy, switching and persistence

Longitudinal cohort definition and persistence of therapy

Ongoing GLP-1 RA use was defined from prescription records together with supporting clinical documentation. Continued therapy was operationalized using a gap-based persistence approach, in which follow-up was censored when the interval between successive indications of GLP-1 RA therapy exceeded a prespecified duration. Primary analyses used a prespecified 120-day gap threshold, selected to balance conservative inference of ongoing therapy with sufficient longitudinal follow-up, for stable estimation of treatment trajectories through 1 year in routine clinical practice. The persistence operationalization and alternative thresholds evaluated in sensitivity analyses are described below.

To ensure consistent temporal alignment across patients, all individuals contributed a baseline observation at day 0 corresponding to treatment initiation. Follow-up was restricted to post-baseline observations and was censored at the earliest occurrence of loss of evidence for ongoing therapy or the end of the analytic window. Maximum follow-up was 18 months (~1.5 years), beyond which sample size and documentation density were insufficient for reliable longitudinal inference.

Operationalization of GLP-1 receptor agonist persistence of therapy

Continued GLP-1 receptor agonist use was operationalized using a backward-looking gap-window approach. At each follow-up day, a binary persistence indicator was set to 1 if the time elapsed since the most recent evidence of GLP-1 RA therapy (prescription, medication order or clinical documentation of ongoing use) was ≤*g* (gap) days and no explicit discontinuation code had been recorded, and 0 otherwise. Evidence events included pharmacy dispense records, prescriber orders and structured or note-derived mentions of active GLP-1 RA use; documented discontinuation was treated as a definitive censoring event regardless of the gap threshold.

For each patient and each candidate *g*, the first day on which the persistence indicator was 0 defined the censoring date; all subsequent observations for that patient were excluded from the analytic dataset for that gap definition. This process was applied separately for each of the eight gap thresholds evaluated: 30, 60, 90, 120, 150, 180, 365 and 730 days. Primary analyses used *g* = 120 days. Sensitivity analyses were conducted across all eight threshold values; estimates were consistent in direction and magnitude, and no conclusion depended on the choice of gap threshold.

Statistical analysis of primary outcomes

Longitudinal trajectories of weight and HbA1c following GLP-1 RA initiation were analyzed as the primary outcomes using generalized estimating equation models to account for repeated measurements within individuals and irregular follow-up intervals. Time since baseline was modeled using spline functions to allow for nonlinear change over follow-up. The spline specification was selected a priori using the quasi-likelihood under the independence model criterion (QICu); 3 d.f. minimized QICu for both weight and HbA1c trajectories.

Change in weight and change in HbA1c were analyzed as continuous outcomes using two complementary models. A single pooled model included baseline glycemic category as a covariate and estimated the adjusted mean difference in each outcome for each category relative to normal glycemia; these differences are constant over follow-up and are reported in Extended Data Table 2. Separate models were then fitted within each baseline glycemic category, and within each sex and age stratum, to generate the model-predicted trajectories presented in Fig. 3 and the model-predicted values at 3, 6, 9 and 12 months in Extended Data Table 3. All models were generalized estimating equations with a Gaussian family and an independence working correlation structure clustered on patient, with robust standard errors, adjusted for age, sex, baseline BMI category, race and concomitant use of medications known to affect weight, and used a B-spline basis for time with 3 d.f.

Complementary time-to-event analyses were performed to evaluate time to clinically meaningful thresholds of weight loss and HbA1c reduction. Kaplan–Meier curves were used to visualize time-to-event distributions stratified by baseline glycemic category. Cox proportional hazards models were used to estimate adjusted hazard ratios. Proportional hazards assumptions were evaluated using standard diagnostic procedures.

All statistical tests were two-sided and CIs were reported where appropriate. Sensitivity analyses examined alternative definitions of GLP-1 RA persistence, cohort restrictions and model specifications (described above).

To quantify the analytic contribution of note-derived data, the primary generalized estimating equation and Cox proportional hazards models were repeated in a structured-data-only subcohort, from which patients who entered the cohort solely through note-derived measurements were excluded. Estimates from this sensitivity analysis are reported alongside the primary estimates in Extended Data Table 2.

Software. Cohort construction and the generalized estimating equation trajectory models were performed in Python v.3.13.5 with pandas v.2.3.2, NumPy v.2.3.2, SciPy v.1.16.1, statsmodels v.0.14.5, patsy v.1.0.1, matplotlib v.3.10.6 and openpyxl 3.1.5. Cox proportional hazards models were fitted with lifelines v.0.30.0 under Python v.3.12 with pandas v.2.2.3, NumPy v.2.2.1.3, SciPy v.1.15.2 and matplotlib v.3.10.1. Extraction validation, benchmark comparison and fine-tuning analyses were performed in Python v.3.12 with NumPy v.2.2.6 and matplotlib v.3.10.1. Concept discovery used OpenAI text embeddings with *k*-means clustering implemented in RAPIDS cuML and dimensionality reduction with UMAP.

Exploratory note-derived clinical assessments

We evaluated longitudinal changes in clinical assessments that were infrequently or inconsistently represented in structured EHR fields, including muscle strength, pain score, PHQ-9 score, waist circumference and alcohol use (drinks per day). These variables were extracted from narrative clinical documentation using the framework described above and retrieved from the longitudinal KG for downstream analysis. These analyses were prespecified as exploratory secondary outcomes and were conducted in all patients with available note-derived observations within the analysis window, without restriction to the adherence-defined cohort used in the primary weight and HbA1c

analyses. Because these assessments were documented in only a subset of patients, applying an additional persistence requirement would have further reduced already limited sample sizes. Results were therefore interpreted as descriptive and hypothesis-generating.

Before analysis, units were harmonized where needed, including conversion of waist circumference to inches and implausible values were excluded using predefined domain constraints: pain scores >10, muscle strength scores >5 and PHQ-9 scores >27. Analyses were restricted to events occurring from 180 days before to 365 days after GLP-1 RA initiation, and observations below the first percentile and above the 99th percentile were excluded within each assessment domain.

Two complementary longitudinal approaches were used (described below). Both used generalized estimating equation models with a Gaussian family, an independence working correlation structure and a B-spline for days from initiation with 3 d.f. All models were adjusted for age, sex, race or ethnicity, baseline HbA1c category and baseline BMI category. For PHQ-9 models, additional adjustment was made for antidepressant use at baseline, given its potential to confound the trajectory of depressive symptoms following GLP-1 RA initiation. To assess whether changes in depression, pain intensity, alcohol use, and muscle strength were independent of the concurrent weight loss observed after GLP-1 RA initiation, models for these four outcomes were additionally adjusted for time-varying percentage weight change from baseline, derived from the primary weight analysis dataset. Because weight and note-derived assessments are recorded on independent clinical schedules, each assessment observation was paired with the temporally nearest weight measurement from the same patient. No weight values were imputed or interpolated, and no observations were excluded for lacking a concurrent weight measurement. Waist circumference was not adjusted for concurrent weight change, because weight change lies on the causal pathway between GLP-1 RA exposure and abdominal adiposity and adjusting for it would remove the effect of interest.

To account for potential disruptions to care-seeking behavior and clinical measurement during the COVID-19 pandemic, a binary indicator variable was included in models for depression (PHQ-9), pain intensity, alcohol use and muscle strength; this variable was coded 1 for patients who initiated GLP-1 RA therapy between 11 March 2020 (WHO pandemic declaration) and 11 May 2022 (end of the US federal public health emergency), and 0 otherwise. The COVID-era covariate was omitted from waist circumference models due to the small analytic sample ($n = 54$ patients with both pre- and post-initiation observations) and disproportionate representation of COVID-era initiators in that subgroup (50%), which would preclude stable estimation.

Exploratory note-derived assessment analyses and subgroup definitions

The exploratory analyses of note-derived clinical assessments (PHQ-9, general pain intensity, waist circumference, alcohol use and muscle strength) are described in the main text. The quantitative results reported in the paper are from a baseline-anchored change-from-baseline approach, restricted to patients with a baseline assessment within ± 30 days of GLP-1 RA initiation and at least one repeat assessment between 30 and 365 days. Where more than one observation fell within the baseline window, their mean was used as the baseline value. Change from each patient's own baseline score was the outcome, modeled using GEE with a B-spline basis (3 d.f.) and the covariate set described in the main text. Estimated changes and 95% CIs were extracted at 3, 6, 9 and 12 months from the model-predicted trajectory.

Elevated-baseline subgroup analyses were restricted to patients meeting prespecified score thresholds at baseline: PHQ-9 ≥ 5 (mild or greater) and ≥ 10 (moderate–severe); pain intensity ≥ 4 (moderate or greater) and ≥ 7 (severe). These subgroups were defined before

analysis and the primary threshold for each domain was the lower of the two. No established clinical threshold exists for alcohol intake, so an elevated-baseline subgroup was instead defined empirically as the highest quartile of baseline consumption (≥ 10 drinks per day, $n = 47$).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Real-world data analyzed in this study were from expert-determined, de-identified patient EHRs, and therefore are governed under Health Insurance Portability and Accountability Act of 1996 (HIPAA), requiring licensing from the EHR owner Harris Computer Systems. These restrictions prohibit the authors from making them publicly available, and individual-level data will not be deposited in a public repository at any time. De-identified individual-level data may be made available to qualified researchers under a separate data license and data-use agreement; fees may apply. Requests should be directed to the corresponding author, will be evaluated jointly by RespondHealth and Harris Computer Systems for scientific merit and consistency with the terms of the license, and will receive a response within one month. Aggregate data supporting all findings, including cohort characteristics, model estimates and confidence intervals, are provided in the main text, Extended Data and Supplementary Information.

Code availability

The technical development and validation code, covering extraction evaluation, physician adjudication analysis, benchmark comparison, fine-tuning evaluation and concept-discovery clustering, is available on GitHub at https://github.com/MyOwnMed/glp1_nature2026_extraction_validation and archived with a permanent digital object identifier via Zenodo at <https://zenodo.org/records/22149576> (ref. 47), released under the Apache License 2.0. This release is self-contained: the physician adjudication records, the annotated benchmark corpus, the ontology release, the model outputs and the aggregate phenotype prevalence counts needed to reproduce every reported validation result are distributed with it, and the pipeline runs from the repository without access to individual-level EHR data.

Analysis code for the GLP-1 receptor agonist study, covering cohort construction, persistence-of-therapy censoring, the generalized estimating equation trajectory models and the Cox proportional hazards time-to-threshold models, is available on GitHub at https://github.com/MyOwnMed/glp1_nature2026_a1c_analysis and archived via Zenodo at <https://doi.org/10.5281/zenodo.22149500> (ref. 48), released under the MIT license. This code cannot be run against the data reported here, because the individual-level records it consumes are licensed from Harris Computer Systems and cannot be redistributed, as described in the Data availability statement. It is published so that every analytic decision behind the reported results can be read and checked, and it may be adapted for use with other datasets.

Each license covers the code in its own repository only. Between them, the two repositories contain the code that computes every value reported in the tables and figures of this paper and its Supplementary Information. The production extraction platform, comprising the deployed agentic extraction pipeline and the KG construction system, is proprietary and cannot be shared for commercial reasons; the validation code released here evaluates that platform's outputs, which are distributed with it.

References

37. Shao, C. et al. Scalable medication extraction and discontinuation identification from electronic health records using large language models. *J. Clin. Epidemiol.* <https://doi.org/10.1016/j.jclinepi.2025.112049> (2025).

38. Kim, E., Foty, R. & Seyfert-Margolis, V. Quantifying the unstructured narrative of patient care in EHR data. *Big Data Anal. Appl.* 70–74 (2025).
39. Lingren, T. et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J. Am. Med. Inform. Assoc.* **21**, 406–413 (2014).
40. Kaji, A. H., Schriger, D. & Green, S. Looking through the retrospectroscope: reducing bias in emergency medicine chart review studies. *Ann. Emerg. Med.* **64**, 292–298 (2014).
41. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127 (2012).
42. Khera, R., Simon, M. A. & Ross, J. S. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *JAMA* **330**, 2255–2257 (2023).
43. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **61**, 29–48 (2008).
44. Lobo, M., Lamurias, A. & Couto, F. M. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed. Res. Int.* **2017**, 8565739 (2017).
45. Yan, S. et al. PhenoRerank: A re-ranking model for phenotypic concept recognition pre-trained on human phenotype ontology. *J. Biomed. Inform.* **129**, 104059 (2022).
46. Fryar, C. D., Gu, Q., Afful, J., Carroll, M. D. & Ogden, C. L. Anthropometric reference data for children and adults: United States, August 2021–August 2023. *Vital Health Stat.* 3 <https://doi.org/10.15620/cdc/174595> (2025).
47. Kim, E. et al. GLP-1 technical development and validation code (Part I). *Zenodo* <https://doi.org/10.5281/zenodo.22149575> (2026).
48. Kim, E. GLP-1 clinical-impact analysis code (Part II). *Zenodo* <https://doi.org/10.5281/zenodo.22149500> (2026).

Acknowledgements

We thank Harris Computer Systems and Sidus Insights (Niagara Falls, NY, USA) for providing access to the de-identified real-world EHR data analyzed in this study, and Microsoft Corporation for technical collaboration and for the Azure compute and platform resources used for model development and extraction at scale.

Author contributions

E.K. and R.F. contributed equally. V.S.-M. and C.B.C. jointly supervised this work. E.K. and V.S.-M. jointly designed the extraction and KG architecture, which E.K., R.F. and V.S.-M. then developed into the full extraction, KG and agentic analysis framework, with E.K. leading its development and implementing the extraction pipeline, ontology grounding and KG construction. E.K., R.F., C.B.C. and V.S.-M. designed the physician validation study. E.K. ran the extraction validation, the benchmark comparison against existing extraction models and the fine-tuning experiments. R.F. analyzed the physician validation data and performed all of the GLP-1 RA clinical analysis, with statistical design and analysis review by J.S.S. and L.F.R. who also provided clinical and epidemiological subject-matter expertise. C.B.C., J.D.P. and M.K. served, respectively, as primary reviewer, second independent reviewer and blinded adjudicator in the chart review validation; C.B.C. also contributed clinical chart review expertise and interpretation of the results. A.S.T. provided clinical expertise, helped design the model benchmarking study, reviewed its results, participated in the chart review and recommended the blinded physician validation against an expert reference standard. V.S.-M. oversaw execution of the extraction and analytic work. V.S.-M. and C.B.C. together provided scientific, clinical expertise and oversaw interpretation and review of all results. C.B.C. and V.S.-M. jointly supervised the study. E.K., R.F., C.B.C. and V.S.-M. wrote the paper, and all authors reviewed, edited and approved the final version.

Funding

This work was supported in part by the National Institutes of Health (NIH) under contract 75N93024C00036. Remaining costs, including cloud compute purchased under a commercial agreement with Microsoft, were funded internally by RespondHealth. The NIH had no role in study design, data collection and analysis, the decision to publish or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Authors employed or contracted by RespondHealth include E.K., R.F. and V.S.-M.

Competing interests

R.F., E.K. and V.S.-M. are employees of RespondHealth. The work described in this manuscript involves methods developed by RespondHealth. M.K. receives grants and contracts from NIH, American Lung Association, Sanofi and GSK paid to Mount Sinai. She receives advisory board and consulting fees addressing treatments for asthma and chronic obstructive pulmonary disease from Sanofi, AstraZeneca, GSK, Genentech, Chiesi, Uniquity Bio, Kinaset and Apogee. She has received speaking fees for unbranded, non-promotional lectures at national meetings from AstraZeneca, Sanofi, GSK and Chiesi. She receives royalties from UpToDate as section editor for severe asthma. She holds patents related to the development of an asthma therapeutic. J.S.S. has served as an advisor to 4Immune, AbbVie, Abvance, ActoBiotics, Adocia, Aerami/Dance Biopharma, AiTA, Applied Therapeutics, Arecor, AstraZeneca, Avotres, Bayer, Biomea Fusion, COUR Pharmaceuticals, Dexcom, Eli Lilly, ICON, Immunothera, Integrated Nanotherapeutics, i2o, Kriya Therapeutics, Novo Nordisk, Oramed, Regor, Remedy Plan, RespondHealth, Sanofi, Shoreline Biosciences, Signos, Vertex, vTv Therapeutics and WiNK. He is a member of the boards of Applied Therapeutics, Elvinix and SAB Bio, and serves as Chair of the Strategic Advisory Board of the EU EDENT1FI consortium. He holds equity in 4Immune, Abvance, AiTA, Applied Therapeutics, Avotres, Dance Biologics, Dexcom, Elvinix, Immunomolecular Therapeutics, Oramed, SAB Bio, Signos, vTv Therapeutics and WiNK. C.B.C. reports grants or contracts from the NIH and the Bill & Melinda Gates Foundation (funds paid to Drexel University); consulting fees from bioMérieux (paid personally, through 2023); participation on data safety monitoring boards and advisory boards for the NIH and RespondHealth (unpaid); and leadership or fiduciary roles with several non-profit organizations. A.S.T. is a co-founder of, and holds stock options in, Teal Health, and is on the Scientific Advisory Board of, has received grants from, or is a consultant for RespondHealth, Cellular Vehicles, Nephrogen, ReThink64, AlloTRx, Inari, Aion Health, GE, Gondolabio and Genentech. L.F.R. declares no competing interests. J.D.P. declares no competing interests.

Additional information

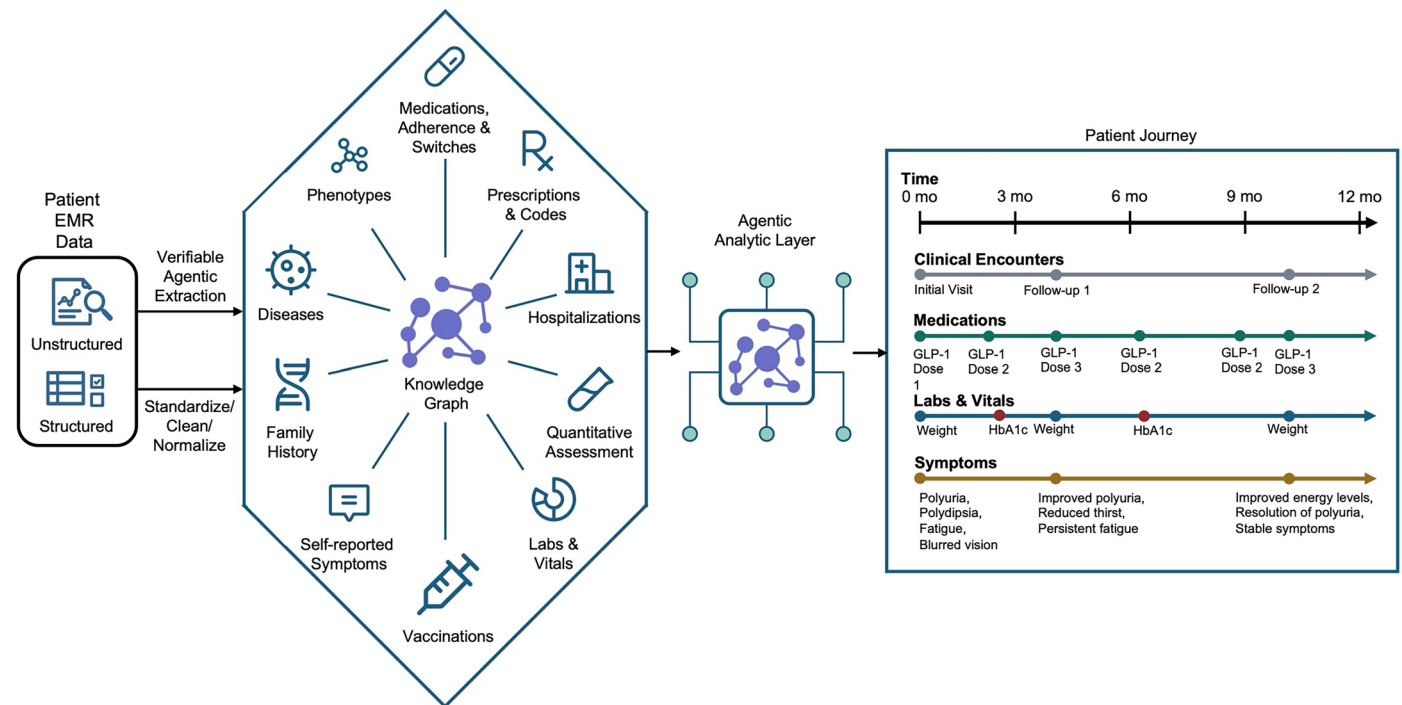
Extended data is available for this paper at <https://doi.org/10.1038/s41591-026-04695-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-026-04695-x>.

Correspondence and requests for materials should be addressed to Vicki Seyfert-Margolis.

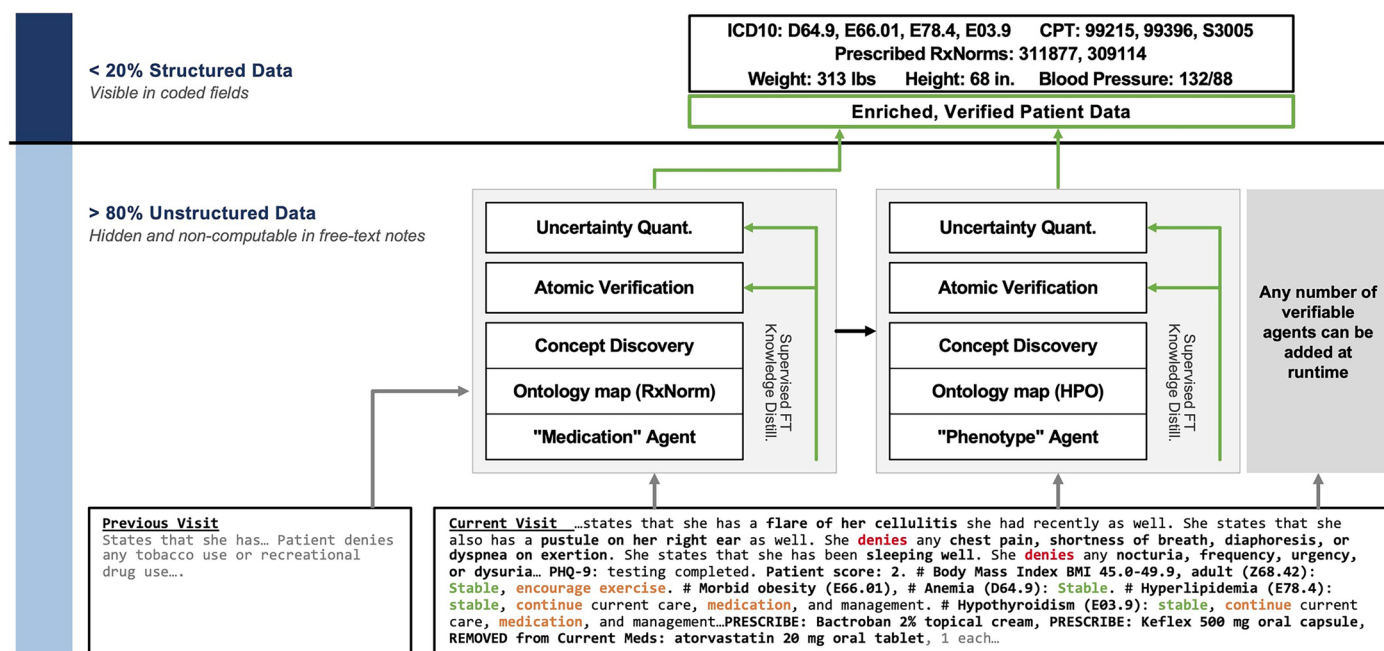
Peer review information *Nature Medicine* thanks Michael Nauck and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editors: Mattia Andreoletti and Joao Monteiro, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Framework for converting structured and unstructured EHR data into a computable longitudinal patient representation. Patient records enter as structured data, which is standardized, cleaned and normalized, and as unstructured clinical notes, which pass through verifiable agentic extraction. Both streams populate a patient-centered knowledge graph spanning phenotypes, diseases, family history, self-reported symptoms, vaccinations,

laboratory values and vital signs, quantitative assessments, hospitalizations, prescriptions and codes, and medications, adherence and switches. An agentic analytic layer queries the graph to reconstruct a patient journey, aligning clinical encounters, medications, laboratory values and vital signs, and symptoms on a common timeline.



Extended Data Fig. 2 | Longitudinal verifiable agentic extraction integrated with structured EHR data. Fewer than 20% of clinical data are visible in coded structured fields, whereas more than 80% reside in free-text notes and are not directly computable. Text from the current and previous visits is routed to domain-specific agents, here a medication agent and a phenotype agent, each combining ontology mapping to RxNorm and to the Human Phenotype

Ontology respectively, concept discovery, atomic verification and uncertainty quantification, with supervised fine-tuning and knowledge distillation. Verified outputs are merged with structured elements, including ICD-10, CPT and RxNorm codes, weight, height and blood pressure, to form an enriched, verified patient record. Additional verifiable agents can be added at runtime.

a

VISIT 1 – Current Medications

Patient here for followup of hypertension, now on low-dose amlodipine. 31-year-old white female with a past medical history significant for hyperthyroidism from Graves' disease and then is being followed for post ablative hypothyroidism. Patient is on Cytomel 10 mcg plus levothyroxine 112 mcg daily. Patient is here to have her blood pressure checked, now on amlodipine 2.5 mg daily Patient was seen and examined. Her blood pressure is still elevated, we'll increase the amlodipine to 5 mg.

Current Medication Extraction

```
"name": "Levothyroxine 112 MCG (0.112 MG) Oral Tablet",
"status": "current",
"dose": "112 mcg (0.112 mg)",
"route": "oral",
"dose_form": "tablet",
"atomic_statement": "Levothyroxine 112 mcg oral tablet is taken for thyroid replacement.",
"matched": true, "RXCUI": "966248",
"matched_name": "levothyroxine 112 mcg (0.112 mg) oral tablet",
"rxnorm_score": 1.00,

"verification": { "statement_score": 3, "confidence": 1.0,
"name_match_score": 3, "name_match_confidence": 1.0,
"original_name": "Levothyroxine 112 MCG (0.112 MG) Oral Tablet",
"matched_name": "levothyroxine 112 mcg (0.112 mg) oral tablet"}
```

Current Medications

- Cytomel 10 mcg
- Levothyroxine 112 mcg
- Amlodipine 2.5 mg

b

VISIT 2 – Medication Change

Patient here for followup of hypertension, now on a higher dose of amlodipine. 31-year-old white female with a past medical history significant for hyperthyroidism from Graves' disease and then is being followed for post ablative hypothyroidism. Patient is on Cytomel 10 mcg plus levothyroxine 112 mcg daily. Patient is here to have her blood pressure checked, one amlodipine 5 mg daily. Her specialist doctor fertility clinic, suggested we changed to labetalol... Patient was seen and examined. Her blood pressure is 120/90. We will stop the amlodipine and start labetalol 100 mg twice a day. She will return in 4 weeks ever blood pressure checked

Change Detected

Amlodipine 2.5 mg

Delta Medication Extraction

```
"medication_switched": [
{ "from": { "name": "Amlodipine 5 MG Oral Tablet", "status":
"discontinued",
"dose": "5 mg", "frequency": "one tablet daily", "route": "oral",
"dose_form": "tablet",
"RXCUI": "197361",
"matched_name": "AMLODIPINE BESYLATE 5 mg ORAL
TABLET",
"rxnorm_score": 0.9485848,

"to": { "name": "Labetalol 100 MG Oral Tablet", "status": "active",
"dose": "100 mg", "frequency": "one tablet twice a day", "route":
"oral",
"dose_form": "tablet",
"RXCUI": " 896758 ",
"matched_name": " labetalol hydrochloride 100 MG Oral Tablet ",
"rxnorm_score": 0.942958,
"atomic_statement": "Patient was switched from Amlodipine 5 mg
daily to Labetalol 100 mg twice daily."}
```

Dosage Change

- Amlodipine 2.5 mg
- Amlodipine 5 mg

Medication Switched

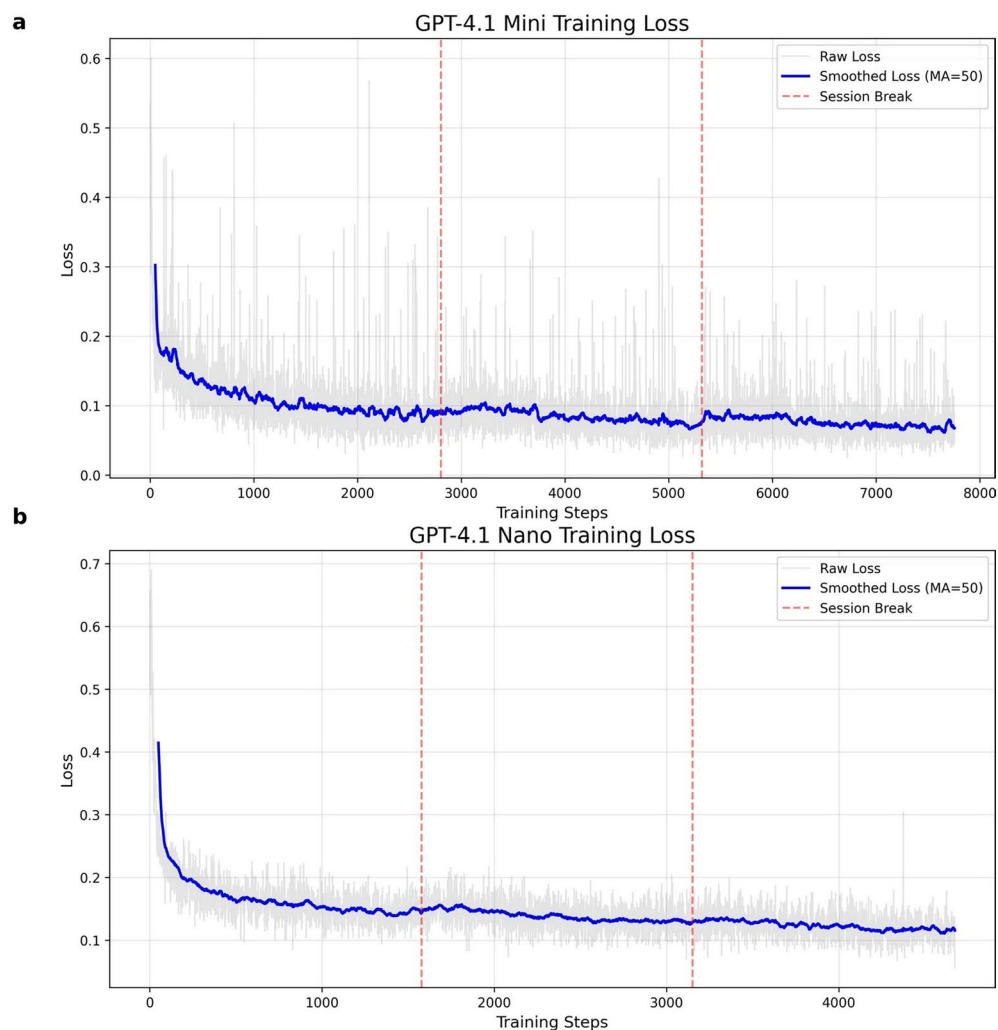
- Amlodipine 5 mg
- Labetalol 100 mg

Unchanged

- Cytomel 10 mcg
- Levothyroxine 112 mcg

Extended Data Fig. 3 | Visit-to-visit change tracking in the medication extraction agent. A worked example for one patient across two encounters. **a.** Visit 1 - Current Medications: Full structured extraction records the current medication list with source-linked evidence, an RxNorm mapping and a verification score. **b.** Visit 2 - Medication Change: Delta extraction compares

the note with the stored prior extraction and emits only what changed, distinguishing a dose change from a medication switch, while unchanged medications are carried forward rather than re-extracted. Each extracted element retains the supporting evidence string from the source note, so every change is traceable to the documentation that produced it.



Extended Data Fig. 4 | Training loss trajectories for low-rank adaptation fine-tuning of compact extraction models. a. GPT-4.1 Mini fine-tuned with low-rank adaptation (LoRA) on a 70/30 split of 100,000 de-identified visits. Gray traces show raw training loss, the solid curve shows a 50-step moving average and

dashed vertical lines mark session breaks. **b.** GPT-4.1 Nano fine-tuned under the same procedure. Both models show a steep initial decrease in training loss and stabilize by approximately 7,500 steps.

Extended Data Table 1 | Inter-rater agreement among three independent physician reviewers

Phenotypes	Rater Pair	N	Agreement (%)	κ
PhenoTagger	MD1 vs MD2	712	87.8	0.600
	MD1 vs MD3	393	87.0	0.659
	MD2 vs MD3	393	91.3	0.800
GPT-5.2 Low	MD1 vs MD2	596	98.5	0.955
	MD1 vs MD3	596	100.0	1.000
	MD2 vs MD3	596	98.5	0.955
Medications	Rater Pair	N	Agreement (%)	κ
BERT Large Med NER	MD1 vs MD2	12	91.7	0.833
	MD1 vs MD3	26	88.5	0.781
	MD2 vs MD3	65	98.5	0.964
BioNER	MD1 vs MD2	8	100.0	1.000
	MD1 vs MD3	8	100.0	1.000
	MD2 vs MD3	35	94.3	0.788
GPT-5.2 Low	MD1 vs MD2	22	100.0	1.000
	MD1 vs MD3	22	100.0	1.000
	MD2 vs MD3	48	95.8	0.917*

N denotes the number of double-rated overlapping reviews used for inter-rater agreement. κ, Cohen's kappa; MD1 to MD3, the three independent physician reviewers. The asterisk indicates a prevalence-adjusted bias-adjusted kappa, reported where Cohen's kappa degenerates in the absence of variance. NER, named entity recognition.

Extended Data Table 2 | Baseline glycemic category and longitudinal weight and HbA1c change following GLP-1 initiation: full cohort versus structured-data-only sensitivity analysis

Structured + Unstructured				Structured Sources Only		
GEE — % weight change						
Baseline glycemic category	Estimate (95% CI)	N	P value	Estimate (95% CI)	N	P value
Normal Glycemia (reference)	reference	3,468	—	reference	2,998	—
Prediabetes	0.47 (0.14 to 0.81)	4,316	0.005	0.26 (−0.02 to 0.53)	3,520	0.068
T2D	0.56 (0.22 to 0.90)	5,864	0.001	0.22 (−0.08 to 0.53)	4,568	0.152
Poorly Controlled T2D	1.49 (1.12 to 1.85)	2,413	<0.001	0.82 (0.51 to 1.13)	1,838	<0.001
GEE — absolute HbA1c change						
Normal Glycemia (reference)	reference	2,846	—	reference	2,015	—
Prediabetes	−0.09 (−0.14 to −0.05)	3,811	<0.001	−0.06 (−0.12 to 0.01)	2,391	0.087
T2D	−0.39 (−0.44 to −0.34)	5,719	<0.001	−0.47 (−0.55 to −0.38)	3,380	<0.001
Poorly Controlled T2D	−1.69 (−1.79 to −1.59)	2,412	<0.001	−2.07 (−2.23 to −1.92)	1,463	<0.001
Cox — time to weight loss						
Baseline glycemic category	HR (95% CI)	Events / N	P value	HR (95% CI)	Events / N	P value
5% weight loss						
Normal Glycemia (reference)	reference	1,225 / 3,468	—	reference	871 / 2,998	—
Prediabetes	0.92 (0.85 to 0.99)	1,479 / 4,316	0.031	0.88 (0.80 to 0.97)	960 / 3,520	0.007
T2D	0.83 (0.76 to 0.89)	1,939 / 5,864	<0.001	0.78 (0.71 to 0.85)	1,127 / 4,568	<0.001
Poorly Controlled T2D	0.65 (0.59 to 0.72)	678 / 2,413	<0.001	0.64 (0.56 to 0.72)	380 / 1,838	<0.001
10% weight loss						
Normal Glycemia (reference)	reference	652 / 3,468	—	reference	400 / 2,998	—
Prediabetes	0.82 (0.74 to 0.92)	667 / 4,316	<0.001	0.82 (0.71 to 0.95)	382 / 3,520	0.006
T2D	0.66 (0.59 to 0.74)	752 / 5,864	<0.001	0.67 (0.58 to 0.78)	394 / 4,568	<0.001
Poorly Controlled T2D	0.56 (0.49 to 0.65)	279 / 2,413	<0.001	0.59 (0.48 to 0.71)	143 / 1,838	<0.001
15% weight loss						
Normal Glycemia (reference)	reference	312 / 3,468	—	reference	190 / 2,998	—
Prediabetes	0.80 (0.68 to 0.94)	299 / 4,316	0.007	0.81 (0.66 to 1.00)	174 / 3,520	0.053
T2D	0.63 (0.53 to 0.74)	321 / 5,864	<0.001	0.66 (0.53 to 0.81)	176 / 4,568	<0.001
Poorly Controlled T2D	0.46 (0.37 to 0.58)	103 / 2,413	<0.001	0.58 (0.43 to 0.77)	65 / 1,838	<0.001
Cox — time to HbA1c reduction						
0.5-point HbA1c reduction						
Normal Glycemia (reference)	reference	156 / 2,523	—	reference	94 / 1,589	—
Prediabetes	3.77 (3.17 to 4.48)	819 / 3,429	<0.001	3.49 (2.79 to 4.37)	430 / 1,871	<0.001
T2D	10.98 (9.32 to 12.93)	2,925 / 5,274	<0.001	9.84 (7.95 to 12.17)	1,476 / 2,709	<0.001
Poorly Controlled T2D	20.08 (17.00 to 23.72)	1,687 / 2,252	<0.001	16.13 (13.00 to 20.02)	910 / 1,184	<0.001
1.0-point HbA1c reduction						
Normal Glycemia (reference)	reference	33 / 2,523	—	reference	26 / 1,589	—
Prediabetes	3.00 (2.06 to 4.38)	150 / 3,429	<0.001	1.99 (1.27 to 3.11)	74 / 1,871	0.003
T2D	29.16 (20.64 to 41.21)	1,920 / 5,274	<0.001	19.79 (13.36 to 29.31)	941 / 2,709	<0.001
Poorly Controlled T2D	82.95 (58.66 to 117.29)	1,579 / 2,252	<0.001	52.10 (35.16 to 77.20)	844 / 1,184	<0.001
1.5-point HbA1c reduction						
Normal Glycemia (reference)	reference	31 / 2,523	—	reference	26 / 1,589	—
Prediabetes	1.12 (0.71 to 1.74)	52 / 3,429	0.631	0.87 (0.52 to 1.47)	32 / 1,871	0.610
T2D	15.36 (10.72 to 22.00)	1,036 / 5,274	<0.001	9.67 (6.49 to 14.40)	487 / 2,709	<0.001
Poorly Controlled T2D	74.98 (52.42 to 107.25)	1,446 / 2,252	<0.001	45.45 (30.62 to 67.47)	765 / 1,184	<0.001

Effect of each baseline glycemic category relative to the reference category, normal glycemia, at a GLP-1 persistence gap of 120 days. Generalized estimating equation (GEE) estimates are adjusted mean differences; Cox hazard ratios above 1 indicate faster attainment of the threshold. Baseline glycemic category enters the GEE models as a covariate rather than as an interaction with time, so each GEE estimate is a difference held constant across follow-up and not the difference at the end of follow-up; the stratified trajectories in Fig. 3 come from separate models fitted within each category. Both adjust for age group, sex, baseline body mass index category, race and concomitant weight-affecting medications, and the GEE models additionally for a B-spline basis for time with 3 degrees of freedom. For GEE, N is the baseline cohort per stratum, 16,061 for weight and 14,788 for HbA1c. For Cox, Events/N is events among patients at risk. Bold denotes $P < 0.05$. The 2.0-point HbA1c endpoint is omitted because it is non-estimable at short persistence gaps. The structured-only sub cohort excludes note-derived entrants. CI, confidence interval; HR, hazard ratio; T2D, type 2 diabetes. Estimates are tested with two-sided Wald tests, from generalized estimating equations with robust standard errors and from Cox proportional hazards models. No adjustment was made for multiple comparisons. Exact P values are given for every comparison in the table; values below 0.001 are shown as <0.001. Italic type marks the analysis blocks and the outcome thresholds within them, and carries no statistical meaning.

Extended Data Table 3 | Model-predicted weight and glycemic change at 3, 6, 9 and 12 months across clinical strata

Population (N)	3 mo	6 mo	9 mo	12 mo
Predicted % weight change following GLP-1 initiation				
Baseline glycemic category				
Normal glycemia (N = 3,468)	-3.1 (-3.8 to -2.4)	-5.6 (-6.4 to -4.8)	-7.1 (-8.0 to -6.2)	-7.7 (-8.7 to -6.7)
Prediabetes (N = 4,316)	-2.7 (-3.3 to -2.2)	-4.7 (-5.2 to -4.1)	-5.6 (-6.2 to -5.0)	-5.9 (-6.6 to -5.2)
T2D (N = 5,864)	-2.2 (-2.6 to -1.8)	-3.8 (-4.3 to -3.3)	-4.6 (-5.1 to -4.0)	-4.9 (-5.5 to -4.2)
Poorly controlled T2D (N = 2,413)	-1.2 (-1.9 to -0.6)	-2.1 (-2.8 to -1.4)	-2.6 (-3.3 to -1.8)	-2.7 (-3.6 to -1.8)
Sex				
Female (N = 9,787)	-2.5 (-2.7 to -2.3)	-4.5 (-4.7 to -4.2)	-5.6 (-6.0 to -5.3)	-6.1 (-6.5 to -5.7)
Male (N = 6,274)	-1.9 (-2.2 to -1.7)	-3.3 (-3.6 to -2.9)	-3.9 (-4.2 to -3.5)	-4.0 (-4.4 to -3.5)
Age group				
20–39 years (N = 1,103)	-3.0 (-3.6 to -2.5)	-5.6 (-6.5 to -4.8)	-7.3 (-8.4 to -6.3)	-8.1 (-9.5 to -6.7)
≥40 years (N = 14,958)	-2.2 (-2.4 to -2.1)	-3.9 (-4.1 to -3.7)	-4.8 (-5.0 to -4.5)	-5.1 (-5.4 to -4.8)
Predicted HbA1c change following GLP-1 initiation (percentage points)				
Baseline glycemic category				
Normal glycemia (N = 2,846)	0.02 (-0.05 to 0.09)	0.02 (-0.05 to 0.09)	0.01 (-0.06 to 0.07)	-0.01 (-0.08 to 0.05)
Prediabetes (N = 3,811)	-0.09 (-0.19 to 0.01)	-0.16 (-0.26 to -0.05)	-0.19 (-0.29 to -0.08)	-0.19 (-0.29 to -0.08)
T2D (N = 5,719)	-0.31 (-0.43 to -0.18)	-0.52 (-0.65 to -0.38)	-0.61 (-0.75 to -0.47)	-0.63 (-0.77 to -0.49)
Poorly controlled T2D (N = 2,412)	-1.26 (-1.50 to -1.01)	-2.06 (-2.31 to -1.81)	-2.36 (-2.61 to -2.10)	-2.36 (-2.63 to -2.08)
Sex				
Female (N = 8,890)	-0.35 (-0.39 to -0.32)	-0.60 (-0.65 to -0.55)	-0.72 (-0.77 to -0.66)	-0.74 (-0.81 to -0.68)
Male (N = 5,898)	-0.41 (-0.47 to -0.36)	-0.68 (-0.75 to -0.60)	-0.78 (-0.86 to -0.70)	-0.78 (-0.87 to -0.70)
Age group				
20–39 years (N = 883)	-0.43 (-0.56 to -0.31)	-0.66 (-0.83 to -0.48)	-0.67 (-0.86 to -0.48)	-0.59 (-0.80 to -0.39)
≥40 years (N = 13,905)	-0.38 (-0.41 to -0.35)	-0.64 (-0.68 to -0.59)	-0.75 (-0.80 to -0.70)	-0.77 (-0.82 to -0.71)

Model-predicted percent weight change and hemoglobin A1c (HbA1c) change following GLP-1 receptor agonist initiation, stratified by baseline glycemic category, sex and age group.

Predictions are based on the primary analytic cohort under a 120-day GLP-1 persistence gap definition and are centered on baseline (day 0). Values are point estimates with 95% confidence intervals. Three-month estimates use day 84 and nine-month estimates use day 280 from the pre-computed prediction grid. Bold denotes estimates whose 95% confidence interval excludes zero. T2D, type 2 diabetes.

Extended Data Table 4 | Model-predicted change in exploratory note-derived assessments following GLP-1 receptor agonist initiation

Population (N)	Estimate (95% CI)			
	3 mo	6 mo	9 mo	12 mo
Depression (PHQ-9, 0–27 scale)				
All patients (N = 625)	–0.41 (–1.76 to +0.95)	–0.51 (–2.00 to +0.97)	–0.54 (–2.18 to +1.10)	–0.71 (–2.40 to +0.98)
Baseline ≥5 (mild+) (N = 263)	–1.11 (–3.07 to +0.84)	–1.30 (–3.38 to +0.78)	–1.47 (–3.79 to +0.85)	–2.52 (–5.01 to –0.04)
Baseline ≥10 (mod-severe) (N = 117)	–1.55 (–4.13 to +1.04)	–2.36 (–5.05 to +0.33)	–3.50 (–6.33 to –0.67)	–6.01 (–9.28 to –2.74)
General Pain Intensity (0–10 scale)				
All patients (N = 436)	+0.13 (–0.39 to +0.65)	+0.07 (–0.46 to +0.60)	–0.01 (–0.55 to +0.53)	+0.08 (–0.60 to +0.76)
Baseline ≥4 (moderate+) (N = 266)	–0.36 (–0.98 to +0.25)	–0.51 (–1.16 to +0.14)	–0.54 (–1.20 to +0.12)	–0.54 (–1.29 to +0.22)
Baseline ≥7 (severe) (N = 160)	–0.50 (–1.23 to +0.22)	–0.81 (–1.57 to –0.05)	–0.95 (–1.75 to –0.15)	–0.97 (–1.93 to –0.01)
Waist Circumference (Inches)				
All patients (N = 99)	–2.57 (–3.80 to –1.35)	–3.59 (–4.89 to –2.30)	–3.48 (–4.89 to –2.06)	–2.68 (–4.60 to –0.76)
Alcohol Use (Drinks/day)				
All patients (N = 198)	–0.02 (–0.16 to +0.13)	–0.01 (–0.16 to +0.14)	+0.00 (–0.14 to +0.15)	–0.01 (–0.26 to +0.23)
Baseline ≥10 (top 25%) (N = 47)	–0.11 (–0.29 to +0.06)	–0.09 (–0.26 to +0.07)	–0.02 (–0.19 to +0.16)	+0.01 (–0.30 to +0.32)
Muscle Strength (MRC, 0–5 scale)				
All patients (N = 219)	+0.01 (–0.06 to +0.08)	+0.02 (–0.05 to +0.08)	+0.02 (–0.05 to +0.08)	–0.01 (–0.10 to +0.08)

Model-predicted change in note-derived clinical assessments following GLP-1 receptor agonist initiation. Outcomes were extracted from unstructured clinical notes using natural language processing. Analyses include patients with a baseline assessment within ±30 days of GLP-1 receptor agonist initiation and at least one repeat assessment between 30 and 365 days, without restriction to continuous GLP-1 users; results should be interpreted as exploratory and hypothesis-generating. Baseline subgroups are defined as: PHQ-9 ≥5 (mild or greater) and ≥10 (moderate–severe); pain score ≥4 (moderate) and ≥7 (severe); and, for alcohol use, the highest quartile of baseline consumption (≥10 drinks per day). All models are adjusted for age, sex, race, baseline glycemic category and baseline BMI category, and additionally for time-varying percent weight change, except for waist circumference, where weight change lies on the causal pathway between GLP-1 exposure and abdominal adiposity and was therefore not included. Values are shown as point estimates with 95% confidence intervals. Bold denotes estimates whose 95% confidence interval excludes zero.

Extended Data Table 5 | BiolarkGSC+ validation set model performance

Model Type	Model	P	R	F1
Encoders & early LLMs	NeuralCR [Arbabi et al., 2019]	0.73	0.61	0.67
	PhenoTagger [Luo et al., 2021]	0.72	0.76	0.74
	PhenoRerank [Yan et al., 2022]	0.84	0.71	0.77
	Llama 3.1 (Nemo 70B) [Kim et al., 2024]	0.78	0.68	0.72
	PhenoGPT (GPT-3) [Yang et al., 2024]	0.82	0.79	0.81
	PhenoGPT (GPT-J) [Yang et al., 2024]	0.80	0.85	0.83
LLMs (base)	Claude Sonnet 4.5	0.95	0.89	0.92
	GPT 5.2 Reasoning Low	0.84	0.90	0.87
	GPT 5 Mini Reasoning Low	0.91	0.80	0.85
	GPT 4.1 Nano	0.90	0.46	0.61
	GPT 4.1 Mini	0.84	0.70	0.76
	Qwen 4B Instruct	0.87	0.62	0.73
Distilled SFT (higher speed, lower cost)	SFT GPT 4.1 Nano	0.72 (-0.18)	0.70 (+0.24)	0.71 (+0.10)
	SFT GPT 4.1 Mini	0.82 (-0.01)	0.80 (+0.10)	0.81 (+0.05)
	SFT Qwen 4B Instruct	0.79 (-0.09)	0.73 (+0.11)	0.76 (+0.03)
Post Proc (PP) (higher precision)	PP Claude Sonnet 4.5	0.96 (+0.01)	0.88 (-0.01)	0.92 (+0.00)
	PP GPT 5.2 Reasoning Low	0.88 (+0.04)	0.89 (-0.01)	0.88 (+0.02)
	PP GPT 4.1 Mini	0.86 (+0.02)	0.70 (-0.00)	0.77 (+0.01)
	PP+SFT GPT 4.1 Mini	0.88 (+0.05)	0.79 (-0.01)	0.83 (+0.02)
	PP Qwen 4B Instruct	0.89 (+0.02)	0.62 (-0.01)	0.73 (+0.00)
	PP+SFT Qwen 4B Instruct	0.87 (+0.08)	0.73 (-0.01)	0.79 (+0.03)
Multi-Agent (higher recall)	Sonnet 4.5 + GPT 5.2 Reasoning Low	0.82 (-0.12)	0.95 (+0.06)	0.88 (-0.03)
	PP (Sonnet 4.5 + GPT 5.2 Reason. Low)	0.87 (-0.10)	0.95 (+0.05)	0.90 (-0.01)
	GPT 4.1 Mini + Qwen 4B Instruct	0.80 (-0.07)	0.82 (+0.11)	0.81 (+0.05)
	PP+SFT (GPT 4.1 Mini + Qwen 4B Inst.)	0.81 (-0.06)	0.87 (+0.07)	0.84 (+0.01)

Class macro-average, document-level, from published literature and off-the-shelf large language models. P, precision; R, recall; F1, harmonic mean of macro-averaged precision and recall; SFT, supervised fine-tuning; PP, post-processing. Macro precision is averaged over classes with at least one prediction and macro recall over classes with at least one gold annotation, so the two averages are taken over overlapping but non-identical class sets. Parenthetical values are absolute deltas versus the corresponding base model, or, for combined models, versus the higher of the two. Rows in the Encoders and early LLMs block are quoted from their source publications, which scored on Human Phenotype Ontology concept identifiers rather than on phenotype names. Bold denotes the best-performing configuration for each metric; where two configurations agree to two decimal places, the bolded cell is the higher before rounding.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Data were obtained from the Sidus Insights (Niagara Falls, NY) de-identified ambulatory electronic health record platform, which aggregates structured clinical records and longitudinal free-text clinical documentation spanning four EHR systems, 1,337 physician practices and 8,871 providers. No software was used for primary data collection. Clinical entity extraction was performed with large language model pipelines served through Microsoft Azure OpenAI Service: GPT-5.2 (reasoning effort: low) for the extraction validation reported in Table 1, GPT-5 Mini Reasoning Low for the one-million-visit production extraction reported in Table 2, and GPT-4.1 Mini and GPT-4.1 Nano fine-tuned with low-rank adaptation, together with Qwen 2.5 4B, for the optimized extraction experiments. Claude Sonnet 4.5 was used in the multi-model aggregation comparison. Extracted statements were grounded against UMLS, SNOMED CT, the Human Phenotype Ontology, ICD-10, LOINC and RxNorm. Semantic embedding used OpenAI text-embedding models; taxonomy discovery used K-means clustering (RAPIDS cuML) and UMAP for dimensionality reduction.
-----------------	--

Data analysis

Cohort construction and the generalized estimating equation trajectory models were performed in Python 3.13.5 with pandas 2.3.2, NumPy 2.3.2, SciPy 1.16.1, statsmodels 0.14.5, patsy 1.0.1, matplotlib 3.10.6 and openpyxl 3.1.5. Cox proportional hazards models were fitted with lifelines 0.30.0 under Python 3.12 with pandas 2.2.3, NumPy 2.1.3, SciPy 1.15.2 and matplotlib 3.10.1. Extraction validation, benchmark comparison and fine-tuning analyses were performed in Python 3.12 with NumPy 2.2.6 and matplotlib 3.10.1. Analysis was carried out in Visual Studio Code. AI research agents assisted with cohort identification, variable assembly and generation of reviewable analysis code within the knowledge graph analytic environment; all intermediate outputs and generated code were reviewed and validated by the investigators before interpretation. All custom code is publicly available: technical development and validation code at https://github.com/MyOwnMed/glp1_nature2026_extraction_validation (Apache License 2.0) and clinical analysis code at https://github.com/MyOwnMed/glp1_nature2026_a1c_analysis (MIT licence). Both repositories are archived with permanent digital object identifiers, given in the Code availability statement of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Real world data analyzed in this study were from expert determined, deidentified patient electronic health records (EHR), and therefore are governed under HIPAA, requiring licensing from the EHR owner Harris Computer Systems. These restrictions prohibit the authors from making them publicly available, and individual-level data will not be deposited in a public repository at any time. De-identified individual-level data may be made available to qualified researchers under a separate data licence and data-use agreement; fees may apply. Requests should be directed to the corresponding author, will be evaluated jointly by RespondHealth Inc. and Harris Computer Systems for scientific merit and consistency with the terms of the licence, and will receive a response within one month. Aggregate data supporting all findings, including cohort characteristics, model estimates and confidence intervals, are provided in the main text, Extended Data and Supplementary Information.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Sex was recorded as documented in the electronic health record and was included as a covariate in adjusted analyses. No inference was made regarding gender identity beyond the recorded variable.
Reporting on race, ethnicity, or other socially relevant groupings	Race was recorded as documented in the electronic health record and reflects routine clinical reporting practices. Race categories were not assigned by the investigators. Race was included as an adjustment covariate in multivariable models but was not used for primary stratification or causal inference.
Population characteristics	The study population consisted of adults aged 18 years or older initiating injectable GLP-1 receptor agonist therapy in routine ambulatory clinical care. Baseline demographic and clinical characteristics, including age, sex, race, body mass index, and glycemic status, are summarized in the results section. Detailed inclusion and exclusion criteria are described in the Methods and Supplementary Methods.
Recruitment	This study used retrospectively collected, de-identified electronic health record data. Participants were not prospectively recruited. Inclusion and exclusion criteria were applied algorithmically to existing clinical records. Potential sources of selection bias inherent to routine clinical data are described in the Discussion.
Ethics oversight	All analyses were conducted using de-identified data in accordance with applicable data use agreements. Institutional review board approval and informed consent were not required for this study under applicable regulations governing research using de-identified secondary data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by data availability within the electronic health record dataset after application of prespecified inclusion and exclusion criteria. No formal a priori sample size or power calculation was performed, as the study was observational and exploratory in nature.
Data exclusions	All exclusion criteria were pre-established before analysis. Sequential exclusions were applied to the identified cohort as follows: age under 18 years or missing age or sex; pregnancy at baseline or during GLP-1 receptor agonist exposure; history of bariatric surgery; and baseline exposure to oral semaglutide (Rybelsus). Physiologic plausibility filters removed implausible measurements: height below 48 or above 90 inches, weight below 60 or above 700 pounds, and BMI below 10 or above 80 kg/m ² , with BMI recalculated from height and weight where necessary. Reference-based outlier screening against CDC anthropometric reference data and short-interval weight-spike detection removed implausible weight values, and a rule-based procedure removed implausible HbA1c spikes; all removed values were flagged and downstream derived variables recomputed. The full rules and thresholds are given in the Methods. No other data were excluded.
Replication	This is a retrospective observational study, so experimental replication does not apply. Four measures were taken to verify the reproducibility of the findings, and all four reproduced the reported direction and statistical significance of the primary results. First, the entire analysis was re-run across eight alternative definitions of GLP-1 receptor agonist persistence of therapy, from a 30-day to a 730-day permissible gap; cohort attrition under every definition is reported in Supplementary Table 1. Second, a structured-data-only sensitivity analysis was run in parallel to the primary analysis and is reported alongside it in Extended Data Table 2. Third, the extraction pipeline was validated against a blinded physician reference standard, with inter-rater agreement among three independent physician reviewers reported in Extended Data Table 1, and benchmarked against published and off-the-shelf models on an external validation set (Extended Data Table 5). Fourth, regression tests covering each analytic correction are distributed with the analysis code. No finding failed to reproduce.
Randomization	Participants were not randomized to treatment. This was an observational study of routine clinical care using existing electronic health record data. For computational feasibility, the extraction and analysis cohort was drawn from a prespecified random sample of approximately 1,000,000 clinical visits from a broader pool of approximately 213,000 individuals with GLP-1 exposure. Downstream analytic cohorts were then defined by applying prespecified clinical inclusion and exclusion criteria, as described in the Methods.
Blinding	Blinding was not applicable, as this was a retrospective observational study using de-identified electronic health record data.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing	
Data exclusions	
Non-participation	
Randomization	

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	

Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☒	☐ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☐	☒ Clinical data		
☒	☐ Dual use research of concern		
☒	☐ Plants		

Antibodies

Antibodies used	
Validation	

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	
Authentication	
Mycoplasma contamination	
Commonly misidentified lines (See ICLAC register)	

Palaeontology and Archaeology

Specimen provenance	
---------------------	----------------------

Specimen deposition

Dating methods

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	Not applicable. This study was a retrospective observational analysis of de-identified electronic health record data and was not a clinical trial requiring registration.
Study protocol	A formal prospective study protocol was not registered. The analytic design, cohort construction, and statistical methods are described in the Methods and Supplementary Methods.
Data collection	Data were obtained from a large de-identified ambulatory electronic health record data platform aggregating structured clinical records and longitudinal free-text documentation from multiple healthcare provider systems. Data reflect routine clinical care over the study period, with up to 10 years of available clinical history per patient, and were collected prior to analysis. No prospective data collection was performed for this study.
Outcomes	Outcomes were pre-defined prior to analysis and included longitudinal percent change in body weight and absolute change in hemoglobin A1c following initiation of injectable GLP-1 receptor agonist therapy, as well as time to clinically meaningful thresholds of weight loss and glycemic improvement. Exploratory secondary outcomes included longitudinal changes in note-derived clinical assessments: depressive symptoms (PHQ-9), pain intensity, waist circumference, alcohol use, and muscle strength. Outcome definitions and assessment procedures are described in the Methods and Supplementary Methods.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☐	☐ Public health
☐	☐ National security
☐	☐ Crops and/or livestock
☐	☐ Ecosystems
☐	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No Yes

- | | | |
|--------------------------|--------------------------|---|
| ☐ | ☐ | Demonstrate how to render a vaccine ineffective |
| ☐ | ☐ | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| ☐ | ☐ | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| ☐ | ☐ | Increase transmissibility of a pathogen |
| ☐ | ☐ | Alter the host range of a pathogen |
| ☐ | ☐ | Enable evasion of diagnostic/detection modalities |
| ☐ | ☐ | Enable the weaponization of a biological agent or toxin |
| ☐ | ☐ | Any other potentially harmful combination of experiments and agents |

Plants

Seed stocks

Not applicable. No plants were used in this study.

Novel plant genotypes

Not applicable. No plants were used in this study.

Authentication

Not applicable. No plants were used in this study.

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

Files in database submission

Genome browser session

(e.g. [UCSC](#))

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Acquisition

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI

☐ Used

☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See [Eklund et al. 2016](#))

Correction

Models & analysis

n/a | Involved in the study

- | | | |
|--------------------------|--------------------------|--|
| ☐ | ☐ | Functional and/or effective connectivity |
| ☐ | ☐ | Graph analysis |
| ☐ | ☐ | Multivariate modeling or predictive analysis |

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis