Computable longitudinal patient journeys derived from structured and unstructured EHR data: development, validation, and clinical application in GLP-1 receptor agonist therapy

Edward Kim[1,2]*, Richard Foty[1]*, Avnesh S. Thakor[3], Jay S. Skyler[4], Lucy F. Robinson[5], Charles B. Cairns[1,6]†, Vicki Seyfert-Margolis[1]†

[1] RespondHealth, Bethesda, MD, USA
[2] Department of Computer Science, Drexel University, Philadelphia, PA, USA
[3] Department of Radiology, School of Medicine, Stanford University, Palo Alto, CA, USA
[4] Diabetes Research Institute, University of Miami Miller School of Medicine, Miami, FL, USA
[5] Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, Philadelphia, PA, USA
[6] College of Medicine, Drexel University, Philadelphia, PA, USA

* These authors contributed equally to this work and are considered co-first authors
† These authors jointly supervised this work

Abstract

Most electronic health record (EHR) data reside in unstructured text, yet real-world evidence studies rely almost exclusively on structured fields. We present a framework using large language models to extract computable elements from unstructured EHR data and integrate them with structured data into provenance-linked knowledge graphs for reviewable longitudinal analysis. Physician adjudication confirmed high precision (95.5–99.5%) and near-perfect inter-rater agreement (κ = 0.917–1.000) across 49 million entities from 31,235 patients. Applied to 16,061 adults initiating GLP-1 receptor agonist therapy, weight loss was greater with lower glycemic burden while HbA1c reduction was greatest in poorly controlled diabetes. Unstructured documentation expanded analytic reach: nearly 70% of HbA1c observations came exclusively from notes, and over 6,200 patients would have been missed using structured data alone. The framework also enabled analysis of outcomes absent from structured fields, including depressive symptoms, pain, and waist circumference. This provides a trustable, transparent, and traceable foundation for converting narrative EHR data into analyzable longitudinal patient journeys at scale.

Introduction

The widespread adoption of electronic health records (EHRs) has created unprecedented opportunities for learning from patient clinical care. Traditionally, EHR-based research and risk prediction has focused predominantly on structured fields, including diagnosis codes, medication orders, laboratory results, and procedure codes. Yet most data in EHRs are unstructured, specifically in free-text documentation that captures symptom narratives, subtle adverse effects, treatment rationales, adherence challenges, family history, and nuanced assessments of disease status that are either not coded at all or are only partially reflected in structured fields. The EHR is analogously referred to as an iceberg, where structured data (codes, labs) represent only the visible 20%, leaving the rich clinical narrative submerged and hidden.[1]

Large-scale real-world clinical data analyses published in literature historically have not utilized unstructured EHR data. A systematic review[2] found that less than 1% of studies utilized unstructured text for case detection, resulting in a 40% loss in sensitivity. In a recent survey of unstructured data use in pediatrics, only 9 out of the 984 research studies utilized unstructured data from EHRs to identify adverse events in children.[3] Importantly, less than 2-4% of social determinants from the unstructured clinical narratives are captured. Failing to extract these unstructured insights blinds researchers to many of the primary drivers of healthcare outcomes. This renders the majority of a patient's journey invisible, causing the derived predictive models to be fundamentally flawed.[4]

To address the problem of unlocking unstructured EHR data, classical natural language processing (NLP) pipelines were developed. These rule-based systems and concept matchers such as cTAKES[5] and MetaMap,[6] demonstrated that notes could be mined for additional diagnoses, problems, and symptoms beyond those visible in billing codes; however, with limitations analogous to key word searching. Biomedical AI has remained largely "unimodal" and fails to integrate the narrative context essential for precision medicine.[7] More recently, transformer-based models, trained on clinical corpora, have further improved extraction performance[8–15] and additional approaches have been developed to minimize AI induced hallucinations.[16] Despite these important advances, there are challenges to making EHR data accessible, extractions accurate, and outputs analyzable for clinical use. Indeed, recent FDA guidance reiterates that, "large amounts of key clinical data are unstructured data within EHRs",[17] and extra care must be taken when converting this data into computable formats.

In this manuscript, we present a novel approach that accurately extracts both structured and unstructured EHR data, provides high agreement with the manual extraction by trained specialty physicians, and is scalable to large patient datasets across all diseases and

clinical conditions. Our approach uses large pre-trained language models to extract clinical entities, which are subsequently organized into a knowledge graph (KG). The result is not an AI driven black-box system that substitutes for clinical or analytic judgment, but a human-guided query environment in which labor-intensive tasks are accelerated while the study design, analytic decisions and interpretation remain under investigator control.

The first section of the paper focuses on analytical validity of the methods surrounding the extraction of data from the EHR's. Additionally, it focuses on how to make the process feasible (time and cost) and scalable, all necessary for large scale data analyses. We incorporate verifiability and principles (atomicity, auditability, decomposition, adjudication, and uncertainty quantification) to ensure accuracy of extractions and analytic outputs.

The second section focuses on clinical utility and actionability of the data by placing the extracted concepts into a KG (multi-dimensional) space to allow all relationships between variables to be interrogated and to allow for scalability of the data.

KGs are a structured representation where entities (nodes) are connected by explicit semantic relationships (edges).[18] This architecture allows for graph traversal, enabling users to "walk" across the data, e.g. navigating from a patient node to their medication history, and then to specific adverse events extracted from notes, thereby revealing connections across disparate data sources. A single agentic-friendly programmatic interface navigates the large-scale data in the KG to make data interrogation feasible with rapid cost-effective, analyses at scale allowing for both exploration and hypothesis testing.

This combination of extracted structured and unstructured data and the agentic framework allows downstream users to query multiple clinical parameters at the level of the individual or at large scale (millions of patient visits). This framework turns previously opaque narrative text into an analyzable graph structure and enables a more faithful representation of real-world clinical complexity.

Finally, we demonstrate the clinical utility of our approach by using this framework to conduct a large-scale longitudinal analysis of treatment responses among individuals initiating glucagon-like peptide-1 receptor agonists (GLP-1 RAs). While there is an abundance of evidence about the effectiveness and safety of GLP-1 RAs for the management of both type 2 diabetes mellitus (T2DM) and obesity from clinical trials,[19–24] there are wide variations in obesity management standards and coverage of obesity medications.[25,26] In addition, there are multiple avenues for direct to consumer access to GLP-1 RAs outside of traditional clinician-initiated prescribing pathways.[27]

Using this framework, we identify large numbers of patients initiating a given medication with high confidence, reconstruct patient-level trajectories over time, and model longitudinal changes in weight and hemoglobin A1c (HbA1c) following GLP-1 RA initiation. We further characterize outcomes captured only in unstructured clinical notes, describe continuous trajectories and time-to-event outcomes to capture how treatment effects evolve, and incorporate baseline markers of glycemic status alongside clinically verified note-derived phenotypes. Together, these capabilities move beyond fixed-timepoint endpoints to provide a more complete characterization of real-world treatment response.

Methods

Platform architecture and technical framework

We developed a framework that integrates structured and unstructured EHR data into a computable, reviewable, longitudinal representation of care (Supplementary Fig. S1). The framework combined structured clinical data with information extracted from narrative clinical notes, mapped these elements to existing biomedical ontologies where possible, and organized them within a patient-centered KG. Across this workflow, self-improving AI research agents were used for note extraction, ontology grounding, graph assembly, retrieval, cohort construction, and generation of reviewable analysis code under investigator oversight.

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, a large language model-based extraction pipeline processed unstructured clinical notes to derive computable elements, including medication events (side effects, discontinuations, switches, dose changes, and current therapies), self-reported symptoms, phenotypes, adherence barriers, family history, hospitalizations, assessments, and disease-status events. Extracted elements were mapped to standard concept systems where possible. When existing ontologies did not provide adequate coverage, recurrent unmapped concepts in the study notes were organized into taxonomic categories, including gastrointestinal side effects, cost-related barriers, and dose-change patterns, and incorporated into the classification workflow.

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.

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 switches. Verification was implemented at multiple stages, as described below (Supplementary Fig. S2). 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 named entity recognition benchmarks.[11] 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. 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 prior 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 introduced.[28] 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 prior 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 with the immediately preceding note and with the prior 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 (Supplementary Table 1; Fig S3).

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.

The procedure began by aggregating short text spans from verified extraction events. These spans were embedded in a semantic vector space and clustered using K-means, with the number of clusters selected by silhouette analysis. An instruction-tuned LLM then assigned labels to each cluster and aggregated clusters into higher-level parent categories, e.g., grouping nausea and vomiting under gastrointestinal side effects. A two-level hierarchy was empirically selected for this taxonomy, although we note that the depth could also a tunable parameter. Taxonomy stability was assessed using UMAP projection. The resulting categories were incorporated into the KG as reusable derived concepts for filtering, classification, and query. Detailed descriptions of embedding, clustering, labeling, parameter optimization, and visualization procedures are provided in the supplementary material (Supplementary Note 1; Supplementary Figs. S4 and S5). To enable interoperability with established biomedical vocabularies, we additionally trained a text span classifier capable of mapping clinical extraction snippets to existing ontologies where applicable.

Verification of extracted statements

Unstructured EHR text was converted into structured atomic statements across multiple clinical domains, including symptoms, phenotypes, diseases and conditions, medications, assessments, family history, adherence-related events, and medication switching. 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 (Supplementary Note 2).

Physician adjudication of extracted statements

Clinical validity of extracted statements was assessed by structured physician review of model outputs using an interactive chart-review interface. Extracted outputs were sampled across 20 patients, yielding 2,710 reviewed extractions. The review set included outputs from the primary LLM-based extraction pipeline and comparator models for medication and phenotype extraction, including PhenoTagger as a non-LLM baseline. 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.

For each review instance, the interface displayed the source clinical note together with the corresponding extracted outputs in a split-screen format. Selecting an extracted item revealed the supporting text span within the source note, enabling direct verification of source linkage. 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. Interrater agreement was quantified using percent agreement and Cohen's kappa. 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.

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.

To evaluate optimization strategies quantitatively, we benchmarked model variants on the BiolarkGSC+ validation set, a manually annotated corpus labeled with Human Phenotype Ontology (HPO) terms.[29] Performance was assessed using document-level macro metrics based on HPO name matching, as described previously.[30] These experiments informed three complementary optimization strategies: targeted fine-tuning or distillation, task-specific post-processing, and multi-model aggregation (Supplementary Note 3).

KG 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 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-1 receptor agonist 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.

Data source

Data for this study were obtained from the Sidus Insights (Niagara Falls, NY) real-world data platform, a large de-identified ambulatory EHR data resource derived from multiple healthcare provider systems spanning four distinct EHR systems, 1,337 physician practices, and 8,871 unique providers. The source data included structured clinical records and longitudinal free-text documentation. 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.

Study population and analytic cohorts

The study population was derived sequentially from the Sidus Insights longitudinal EHR resource. 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 hemoglobin A1c information within a prespecified window, spanning 60 days pre to 14 days post 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 naïve to prior GLP-1 receptor agonist therapy at baseline. We excluded pregnancy during the study window, prior 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 hemoglobin A1c 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 standards.[31] Height, weight, body mass index (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 (Supplementary Note 4).

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.

Longitudinal cohort definition and persistence of therapy

Ongoing GLP-1 receptor agonist 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 receptor agonist therapy exceeded a prespecified duration. Primary analyses used a 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 in Supplementary Note 5.

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 (540 days), beyond which sample size and documentation density were insufficient for reliable longitudinal inference.

Statistical analysis of primary outcomes

Longitudinal trajectories of weight and HbA1c following GLP-1 receptor agonist 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 non-linear change over follow-up. The spline specification was selected a priori using the quasi-likelihood under the independence model criterion (QICu); 3 degrees of freedom minimized QICu for both weight and HbA1c trajectories.

Change in weight and change in HbA1c were analyzed as continuous outcomes. Baseline glycemic category was included as an effect modifier to evaluate differences in trajectories across glycemic states. Models were adjusted for age, sex, baseline BMI category, race, and concomitant use of medications known to affect weight. Model-based predicted trajectories were generated for interpretation.

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 confidence intervals were reported where appropriate. Sensitivity analyses examined alternative definitions of GLP-1 receptor agonist persistence, cohort restrictions, and model specifications (Supplementary Notes 5 and 6).

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 initiation, and observations below the 1st percentile and above the 99th percentile were excluded within each assessment domain.

Two complementary longitudinal approaches were used. 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 degrees of freedom. 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 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 initiation, models for these four outcomes were additionally adjusted for time-varying percentage weight change from baseline, derived from the primary weight analysis dataset and merged to each domain observation by nearest date (within ±60 days).

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 therapy between March 11, 2020 (WHO pandemic declaration) and May 11, 2022 (end of the U.S. 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.

Results

Technical validation and KG construction

Patient Data Pipeline and Cohort Identification for Extractions and Analyses

We analyzed unstructured EHR text from a prespecified random sample of approximately 1,000,000 clinical visits to ensure computational feasibility at scale, capturing 31,235 patients (Fig. 1; Table 1a). These visits were drawn from a broader cohort of approximately 213,000 individuals with GLP-1 exposure in a longitudinal ambulatory EHR resource with up to 10 years of available clinical history.

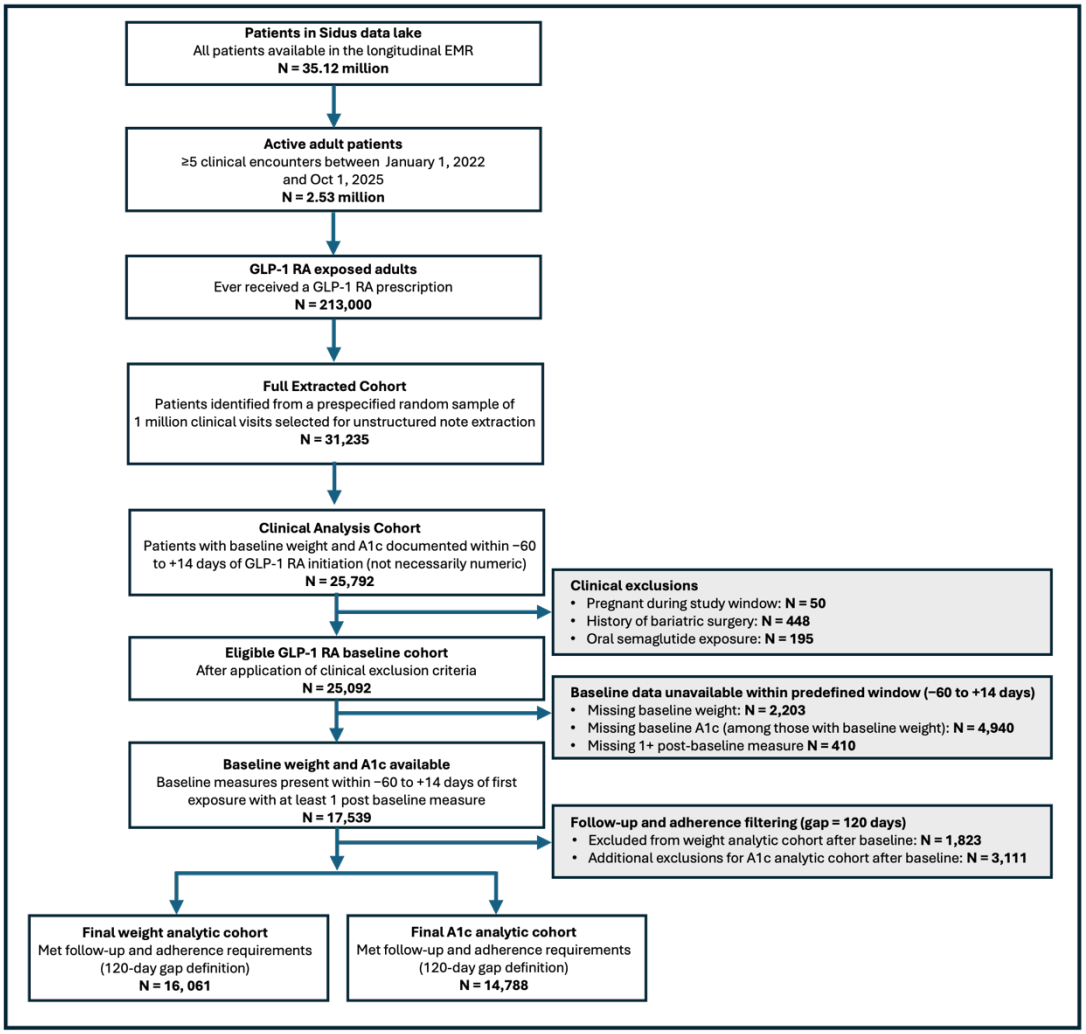


Fig. 1: Flow diagram summarizing cohort selection, baseline assessment pathways, and definition of the final analytic cohorts for longitudinal weight and HbA1c analyses**.**

Quantification and Validation of the Extraction Process

The extraction pipeline recovered 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.

Most extracted entities were fully supported by source evidence, with full-support rates of 94.1–98.4% across major categories, but slightly lower (75.5%) in more ambiguous categories like medication switching and the reasons behind the switch.

Physician adjudication confirmed high clinical validity for both phenotype and medication extraction (Table 1b). Three board-certified physicians from multiple institutions independently reviewed extracted entities. Physicians 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 (κ = 0.955–1.000 across pairs) and substantially lower for the non-LLM baseline PhenoTagger (κ = 0.600–0.800), reflecting the latter's inconsistent handling of clinical language. Across 3,736 physician phenotype reviews, our pipeline achieved 99.5% precision; 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 (κ = 0.917–1.000) compared to BERT Large Med NER[32,33] (κ = 0.781–0.964) and BioNER[34] (κ = 0.788–1.000). Across 744 physician medication reviews, our pipeline achieved 95.5% precision, versus 81.6% for BioNER, and 65.4% for BERT Large Med NER. Together, these findings indicate that the extraction framework recovered clinically relevant information from narrative documentation with high validity in real-world notes.

Table 1a: Cohort summary and extraction verification statistics.

Total Cohort Summary of 1M visits from 31,235 patients.						
Category	**Records**	**Entities**	**Score 3**	**Score 2**	**Score 1**	**Score 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).

Table 1b: Physician adjudication of extraction precision and inter-rater agreement across three independent reviewers

Panel A. Inter-rater Agreement				
Phenotypes	**Rater Pair**	**N**	**Agreement (%)**	**K**
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
Ours (GPT)	MD1 vs MD2	596	98.5	0.955
	MD1 vs MD3	596	100.0	1.00
	MD2 vs MD3	596	98.5	0.955
Medications	**Rater Pair**	**N**	**Agreement (%)**	**K**
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	1.00
	MD1 vs MD3	8	100	1.00
	MD2 vs MD3	35	94.3	0.788
Ours (GPT)	MD1 vs MD3	22	100	1.00
	MD1 vs MD2	22	100	1.00
	MD2 vs MD3	48	95.8	0.917*
Panel B. Extraction Performance				
Model	**Phenotypes (N)**	**Precision (%)**	**Negation Error (%)**	**Other Error (%)**
PhenoTagger	1,847	15.3	83.4	1.3
Ours (GPT)	1,889	99.5	0.0	0.5
Model	**Medications (N)**	**Precision (%)**	**Negation Error (%)**	**Other Error (%)**
Bert Large Med NER	344	65.4	3.4	15.1
BioNER	179	81.6	3.8	30.8
Ours (GPT)	221	95.5	2.3	2.3

* Indicates PABAK measure given Cohen's kappa degeneration with no variance. *N in Panel A* denotes the number of double-rated overlapping reviews used for inter-rater agreement; extraction performance metrics in Panel B were computed over 3,736 physician phenotype reviews and 744 physician medication reviews.

Concept Discovery

When standard medical terminology already exists, we map the extracted terms to the associated vocabulary, e.g. ICD10, HPO, UMLS, etc. For custom terms, we organized note-derived events into data driven taxonomies including 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

then 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 showed stable semantic structure across parameter sweeps, supporting the consistency of the discovered groupings.

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

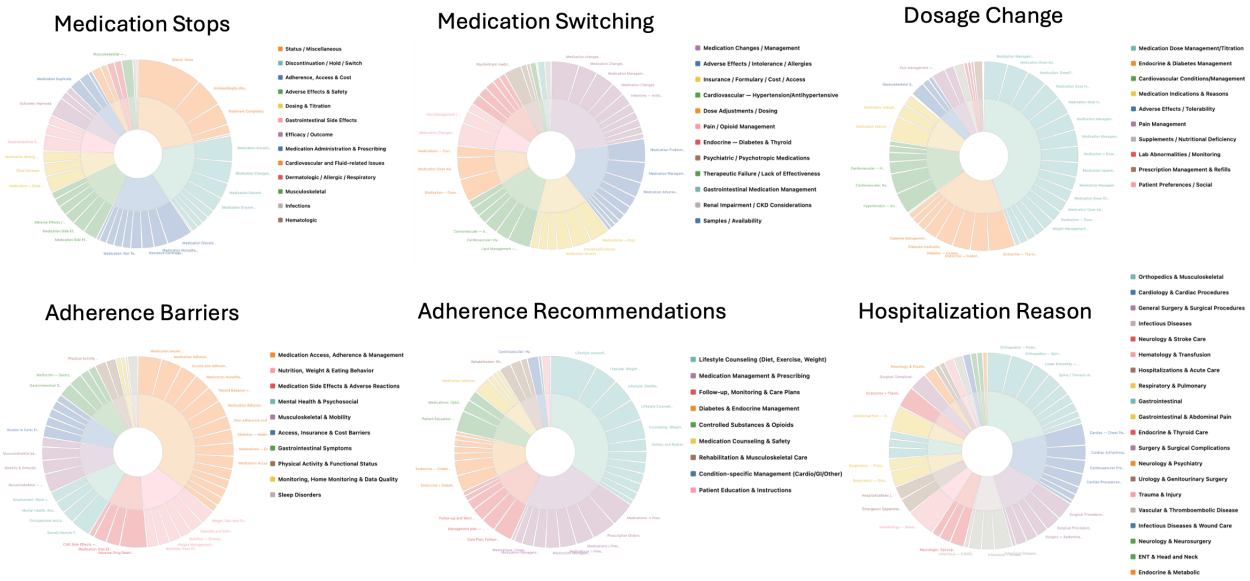
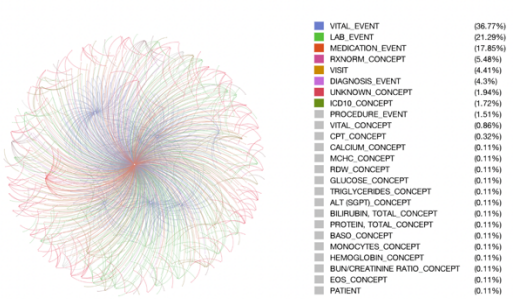
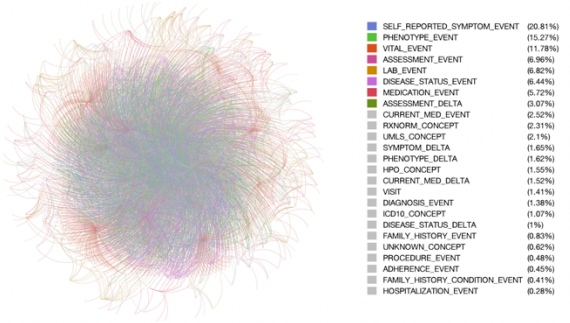


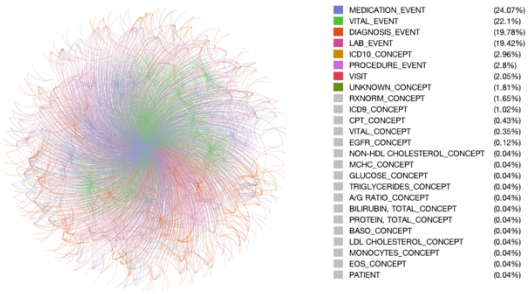
Fig. 2a: Data-driven taxonomy of note-derived events. Short text motifs extracted from clinical notes are embedded and clustered, labeled into interpretable groups, and consolidated into a practical hierarchy for downstream analytics.



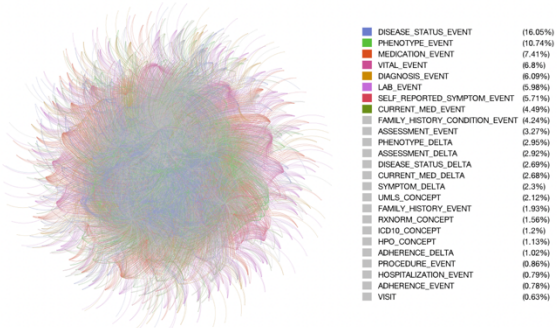
i) Patient 1, structured only



ii) Patient 1, structured + unstructured



iii) Patient 2, structured only



iv) Patient 2, structured + unstructured

Fig. 2b KG representations for two individual patients with and without unstructured note-derived data. KGs for Patient 1 (43 visits; **i, ii**) and Patient 2 (52 visits; **iii, iv**). Left panels (**i, iii**) show graphs constructed from structured EHR data only (930 and 2,538 nodes, respectively); right panels (**ii, iv**) show graphs augmented with note-derived entities (2,902 and 8,249 nodes, respectively). The expanded node diversity and density in the augmented graphs reflects the additional clinical information recovered from unstructured documentation.

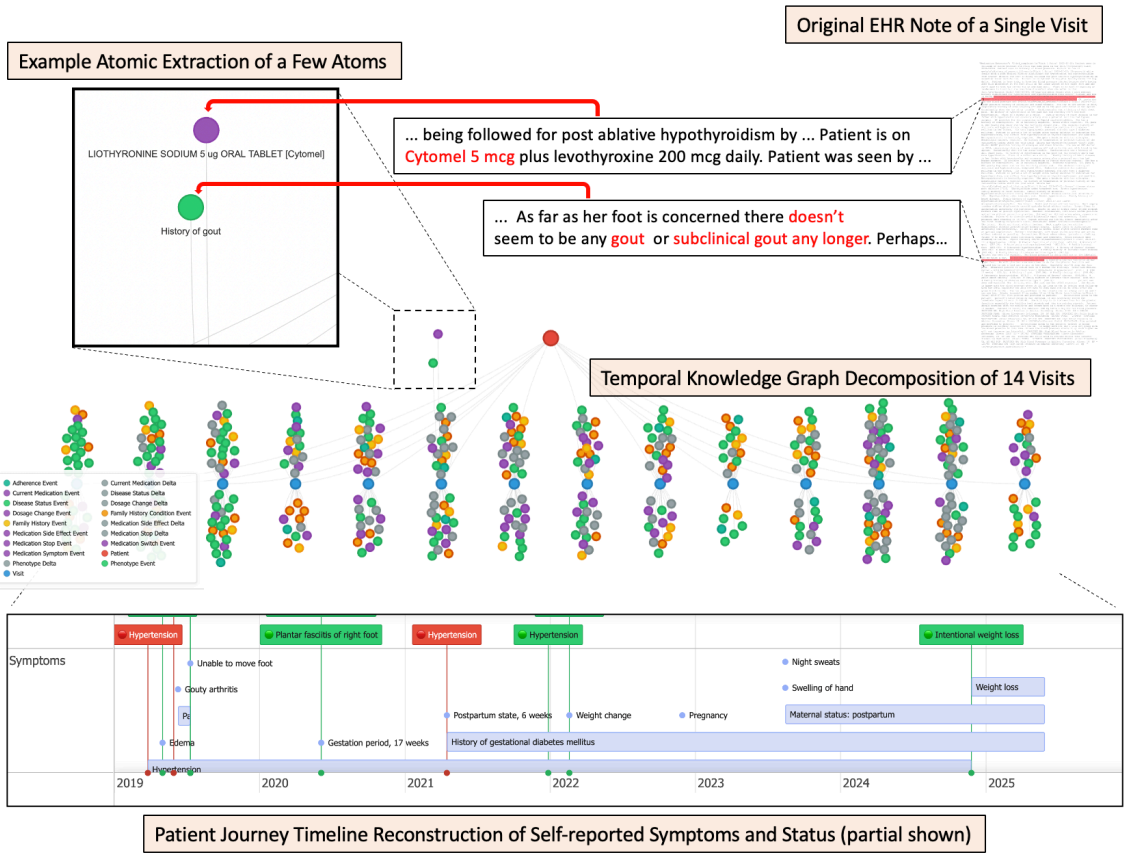


Fig. 2c: Patient Journey Reconstruction. The visualization combines a force-directed graph (top) showing the structural relationships between the patient, visits, and extracted clinical events, with a temporal timeline (bottom) aligning these events chronologically.

KG of Computable Elements for Clinical Analyses

Within the clinical analysis cohort defined in Fig. 1, comprising 25,792 patients with baseline weight and HbA1c documented from 60 days before to 14 days after GLP-1 initiation, we assembled a patient-centered KG for downstream longitudinal analyses. This graph comprised 88 million nodes and 241 million edges and mapped to more than 58,000 UMLS concepts, 8,736 HPO concepts, and 39,827 unique medications.

We represented each patient as a temporally linked KG that integrated note-derived phenotypes with structured EHR data, including diagnoses, laboratory values, vital signs and medications (Fig. 2b). Each element remained linked to source evidence, visit metadata, assertion or negation status and verification level, and was grounded to reference terminologies including HPO, UMLS and RxNorm, or our data driven taxonomy. These retrieval, grounding and graph-construction, steps were performed using agents that generated reproducible code and preserved provenance links to the underlying source data. This unified representation supported both cohort-level analytics and auditable reconstruction of longitudinal patient journeys. Shared visit anchors enabled temporal queries linking GLP-1 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 and uncertainty quantification metadata (Fig. 2c).

Clinical application in a GLP-1 initiation cohort

Description of the Analytic Cohort

Within this KG, agent-based workflows identified eligible patients, aligned note-derived and structured observations longitudinally around treatment initiation, and generated reviewable analytic code with source-linked provenance to assemble longitudinal cohorts for weight and glycemic analyses. Cohort assembly and attrition are summarized in Fig. 1.

After applying plausibility, follow-up, and persistence filters, the primary analysis cohorts included 16,061 individuals for longitudinal weight analysis and 14,788 for longitudinal HbA1c analysis. The smaller HbA1c cohort reflected less frequent testing in routine care. Under the 120-day persistence definition, the weight cohort decreased from 16,061 at baseline to 2,360 at 12 months and 367 at 18 months, reflecting expected attrition over time in the persistence-restricted cohort (Supplementary Table S2). Notably, 6,684 patients in the weight cohort and 6,238 in the HbA1c cohort were included on the basis of baseline

measurements identified from unstructured clinical notes; these patients would not have entered the analytic cohorts using structured baseline data alone.

Baseline characteristics varied across glycemic strata (Table 2). Individuals with normal glycemia were younger and had the lowest baseline HbA1c, whereas those with T2DM and poorly controlled diabetes were older and had progressively higher baseline HbA1c; more than 95% were aged ≥40 years in the latter groups. Baseline weight and 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 ≥10% and ≥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.

Analytic contribution of unstructured clinical documentation

Of the 14,788 patients included in the A1c analytic cohort, 42.2% had baseline HbA1c identified from unstructured clinical notes, whereas 57.2% had baseline HbA1c identified from structured EHR data. These source patterns persisted longitudinally. Patients with unstructured baseline HbA1c derived most subsequent HbA1c measurements from unstructured documentation (84.0%), whereas those with structured baseline HbA1c had most follow-up measurements captured in structured laboratory data (72.8%). Patients in the unstructured-baseline group also contributed a higher density of longitudinal glycaemic measurements (mean 7.31 vs 3.00 observations per patient).

Across all patients, 70.0% of HbA1c observations were identified exclusively from unstructured clinical notes, compared with 28.7% from structured laboratory data (1.3% from both sources). In contrast, weight was captured predominantly in structured vital-sign data (98.8%), with only 1.2% identified from unstructured documentation. GLP-1 RA exposure relied on both sources, with 47.9% of events supported by structured prescription records, 33.4% exclusively by clinical notes and 18.6% by both.

Integration of unstructured documentation affected the analyses in two ways. First, it increased cohort ascertainment and longitudinal measurement density, particularly for HbA1c, where patients identified through clinical notes contributed substantially more follow-up observations. Second, it enabled evaluation of clinical domains not consistently represented in structured EHR fields, including note-derived phenotypes and contextual clinical information. Together, these results illustrate both effects: increased precision for

conventional outcomes and access to clinical information not otherwise available for longitudinal analysis.

Table 2. 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)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
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					
	N (%)	N (%)	N (%)	N (%)	N (%)
<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	2938 (18.3%)	599 (17.3%)	599 (13.9%)	1197 (20.4%)	543 (22.5%)
Obese class I	4626 (28.8%)	1033 (29.8%)	1230 (28.5%)	1669 (28.5%)	694 (28.8%)
Obese class II	3777 (23.5%)	848 (24.5%)	1106 (25.6%)	1307 (22.3%)	516 (21.4%)
Obese class III	3364 (20.9%)	728 (21.0%)	1113 (25.8%)	1105 (18.8%)	418 (17.3%)
Unknown	662 (4.1%)	150 (4.3%)	156 (3.6%)	288 (4.9%)	68 (2.8%)
Sex					
Female	9787 (60.9%)	2671 (77.0%)	2861 (66.3%)	3040 (51.8%)	1215 (50.4%)
Male	6274 (39.1%)	797 (23.0%)	1455 (33.7%)	2824 (48.2%)	1198 (49.6%)
Race					
Caucasian	6543 (40.7%)	1418 (40.9%)	1731 (40.1%)	2485 (42.4%)	909 (37.7%)
Other	2432 (15.1%)	403 (11.6%)	695 (16.1%)	881 (15.0%)	453 (18.8%)
Unknown/Refused	7086 (44.1%)	1647 (47.5%)	1890 (43.8%)	2498 (42.6%)	1051 (43.6%)

Longitudinal weight and HbA1c trajectories stratified by baseline glycemic category, age and sex

Model-based longitudinal trajectories showed greater and more rapid weight loss after GLP-1 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, −4.9% (−5.5 to −4.2) in T2DM and −2.7% (−3.6 to −1.8) in poorly controlled diabetes (Supplementary Table S3a).

HbA1c trajectories showed the opposite gradient. Absolute HbA1c reduction increased with baseline glycemic severity, with the 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) (Supplementary Table S3a). 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 (Figs. 3c, 3d). 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 otherwise 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 (Supplementary Table S3a).

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 (Supplementary Table S3a).

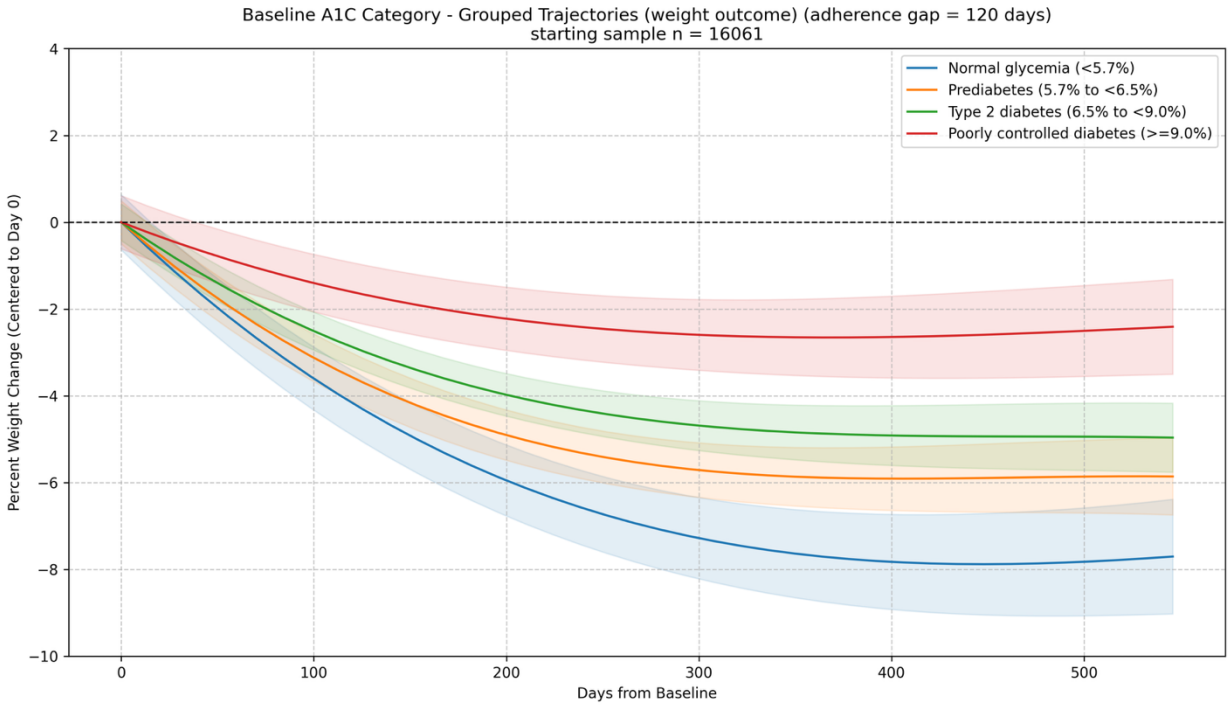


Fig. 3a: Weight trajectories by baseline HbA1c category (GLP-1 persistence gap = 120 days)

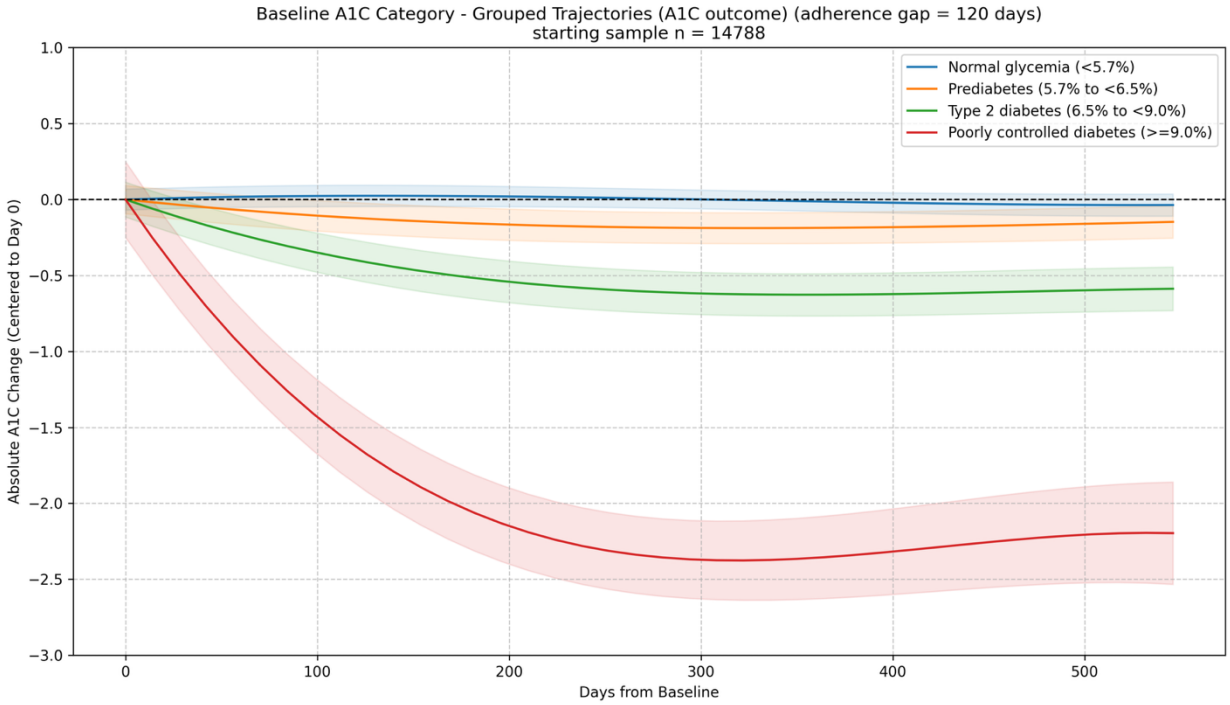


Fig. 3b: Longitudinal hemoglobin HbA1c change following GLP-1 initiation stratified by baseline HbA1c category (GLP-1 persistence gap = 120 days)

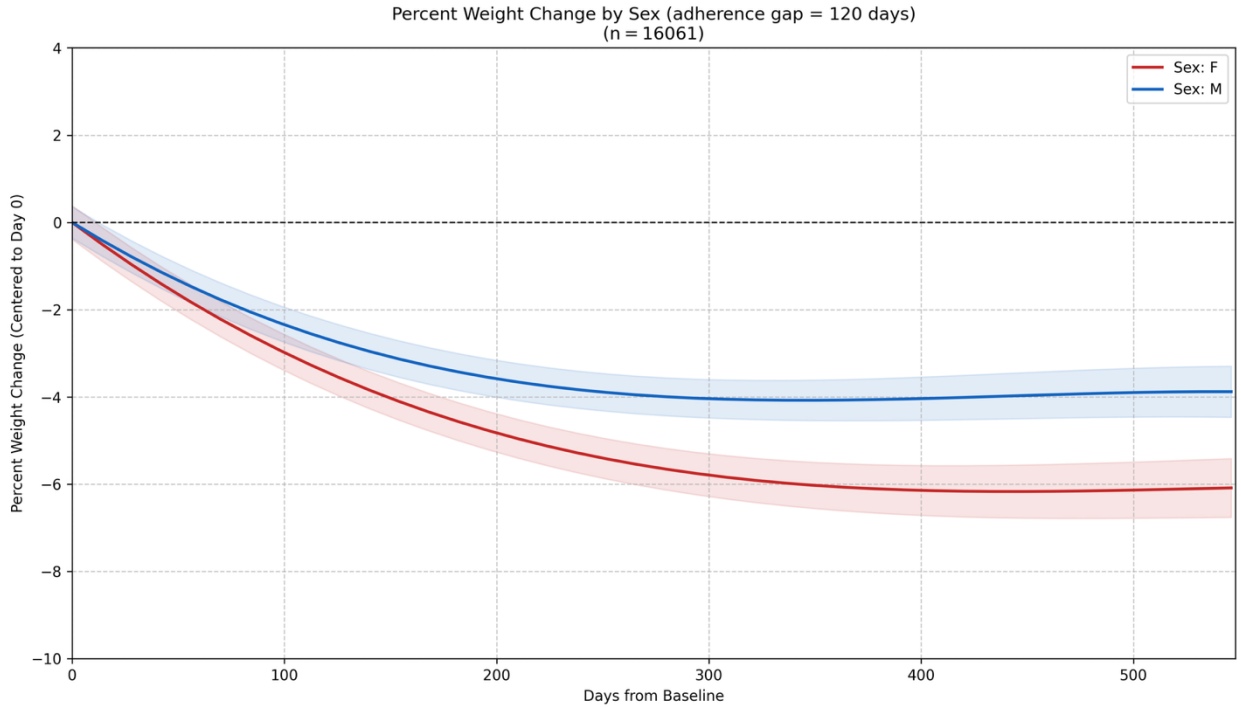


Fig. 3c: Longitudinal weight trajectories stratified by sex (GLP-1 persistence gap = 120 days)

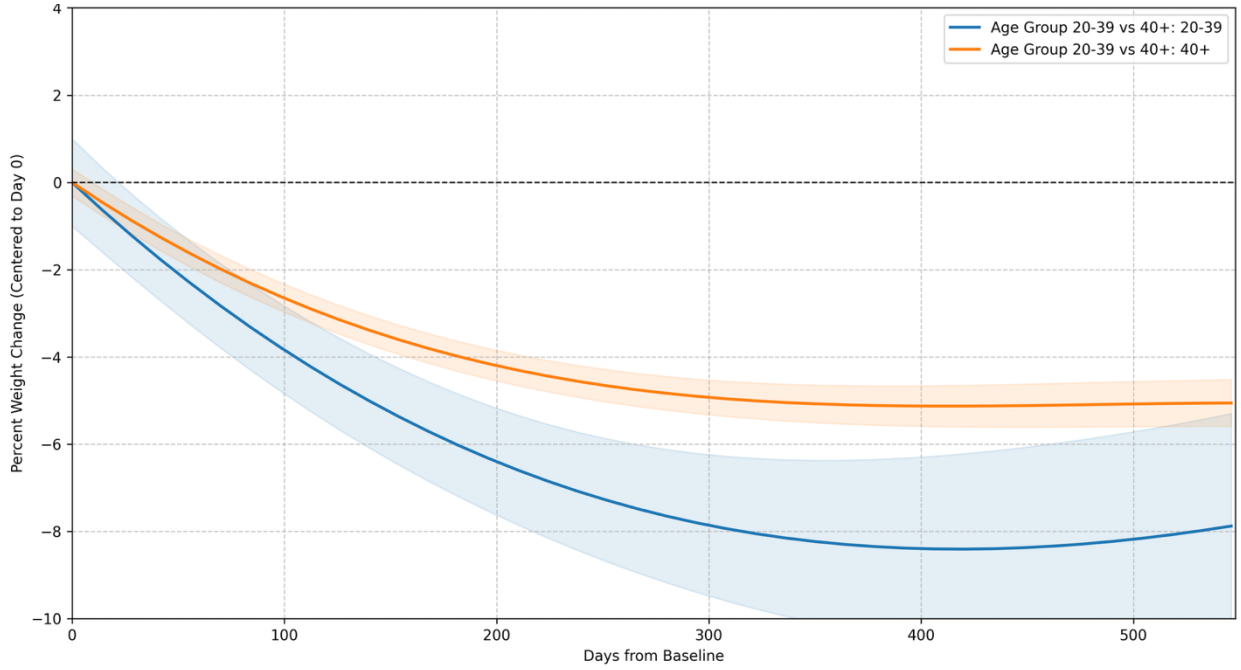


Fig. 3d: Longitudinal weight trajectories stratified by age group (GLP-1 persistence gap = 120 days)

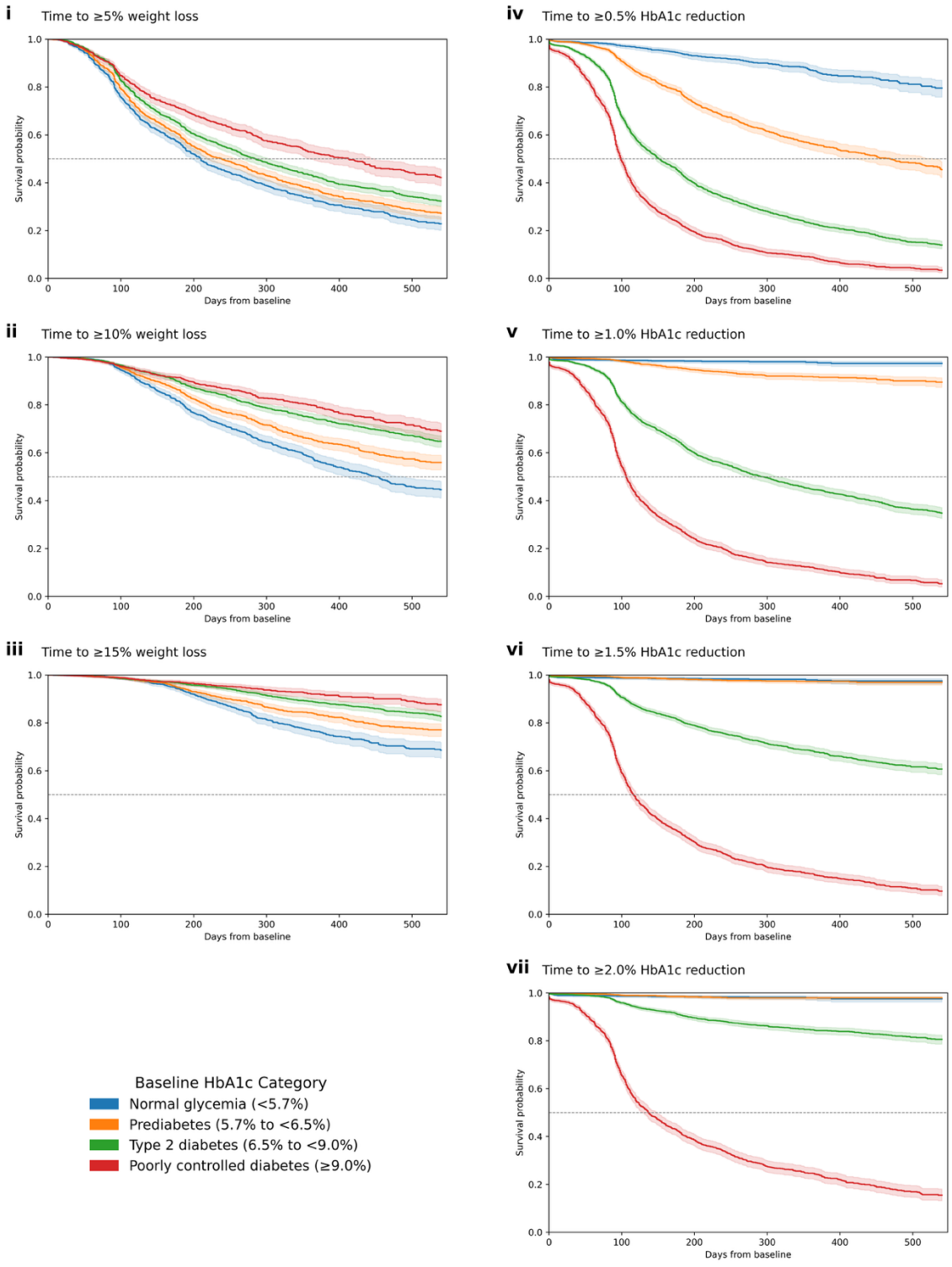


Fig. 3e: Time to clinically meaningful weight loss and HbA1c reduction following GLP-1 initiation, stratified by baseline glycemic category. Kaplan–Meier estimates of time to achieving weight loss (i-iii) and HbA1c (iv-vii). Curves represent the estimated probability of not yet reaching the specified threshold; participants were censored at the end of follow-up or loss of observation.

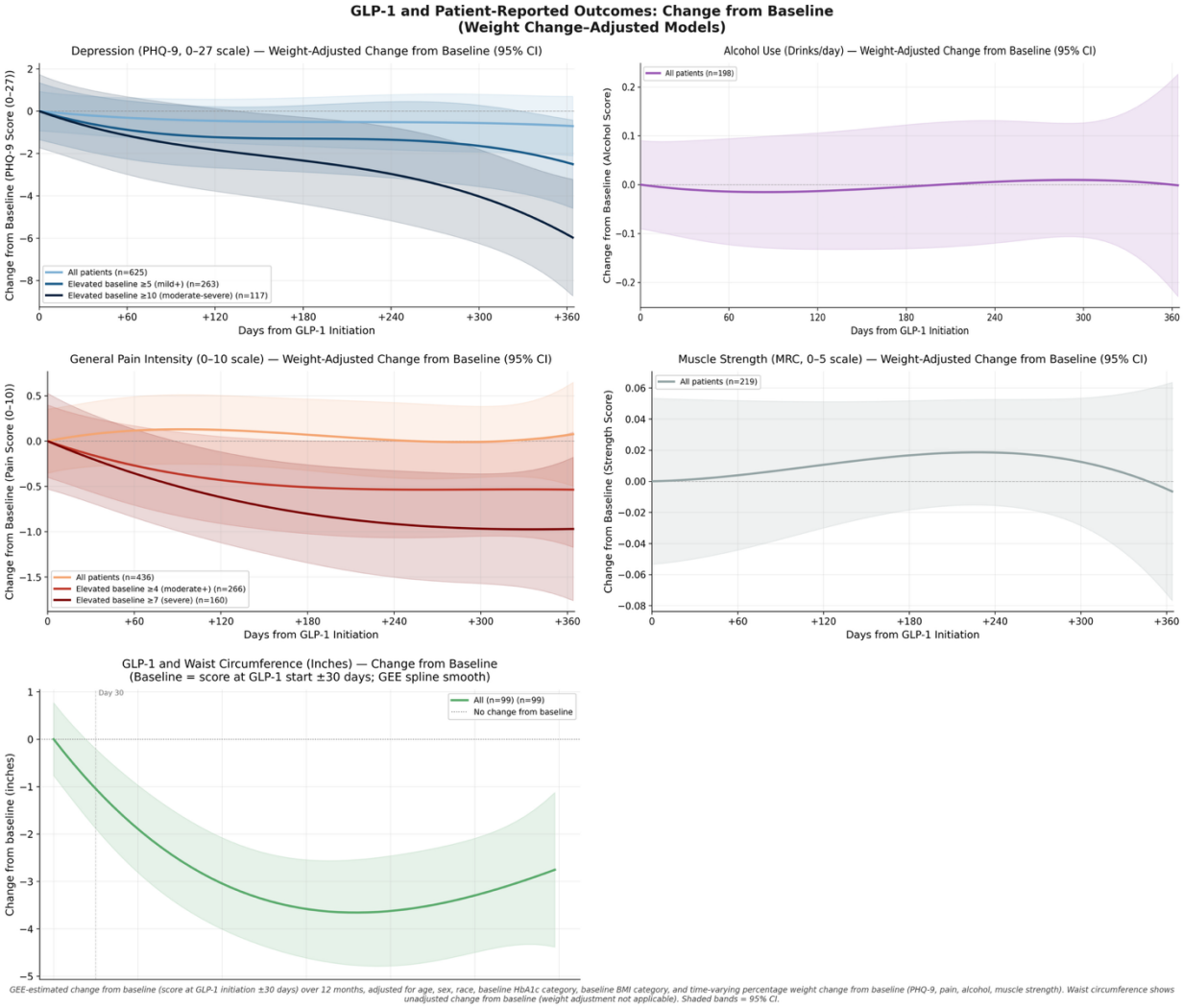


Fig. 3f: GLP-1 and Patient-Reported Outcomes: Change from baseline (score at GLP-1 initiation ±30 days) over 12 months. Shaded bands = 95% CI.

Time to clinically meaningful weight loss and HbA1c reduction

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 (Supplementary Table S3b).

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-vii). 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 with 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); median times to ≥1.5- and ≥2.0-point reductions were estimable only in poorly controlled diabetes (116 days [95% CI, 112–121] and 137 days [129–148], respectively).

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 (Supplementary Table S3b).

Sensitivity analyses

Longitudinal weight and HbA1c trajectories were similar across alternative GLP-1 persistence definitions, including gap thresholds from 30 days to 12 months.

Longitudinal trends in unstructured note-derived clinical assessments

Note-derived assessments were documented for 3,663 patients for PHQ-9, 1,827 for general pain intensity, 243 for waist circumference, 339 for alcohol use, and 642 for muscle strength (Supplementary Table S4a). Across domains, most patients were aged ≥40 years, obesity was common, and mean baseline HbA1c ranged from 6.4% in the waist circumference subgroup to 7.6% in the muscle strength subgroup.

A subset of patients met criteria for longitudinal change analyses, which required sufficient measurements near GLP-1 initiation and during follow-up. The clearest longitudinal signal was observed for depressive symptoms among patients with elevated baseline PHQ-9 scores. Among patients with moderate baseline PHQ-9 ≥5 (n = 263), the estimated change at 12 months was −2.52 points (95% CI, −5.01 to −0.04). Among those with more severe baseline scores, PHQ-9 ≥10 (n = 117), estimated changes were −2.36 points at 6 months (−5.05 to 0.33), −3.50 points at 9 months (−6.33 to −0.67), and −6.01 points at 12 months (−9.28 to −2.74). In the overall PHQ-9 analysis population (n = 625), confidence intervals overlapped zero at all time points (Fig. 3f; Supplementary Table S4b).

Waist circumference also declined over follow-up. Among patients included in the waist circumference analysis (n = 99), estimated changes were −2.57 inches at 3 months (95% CI, −3.80 to −1.35), −3.59 inches at 6 months (95% CI, −4.89 to −2.30), −3.48 inches at 9 months (−4.89 to −2.06), and −2.68 inches at 12 months (−4.60 to −0.76) (Fig. 3f).

General pain intensity showed larger decreases in patients with higher baseline pain scores. Among patients with baseline pain scores ≥4 (n = 266), estimated changes were −0.36 points at 3 months (95% CI, −0.98 to 0.25), −0.51 at 6 months (−1.16 to 0.14), −0.54 at 9 months (−1.20 to 0.12) and −0.54 at 12 months (−1.29 to 0.22), with confidence intervals overlapping zero at all time points. In those with baseline pain scores ≥7 (n = 160), reductions were more pronounced, with estimated changes of −0.50 at 3 months (95% CI, −1.23 to 0.22), −0.81 at 6 months (−1.57 to −0.05), −0.95 at 9 months (−1.75 to −0.15) and −0.97 at 12 months (−1.93 to −0.01); the 6-, 9-, and 12-month confidence intervals excluded zero, whereas the 3-month interval overlapped zero. No clear longitudinal changes were observed for alcohol use or muscle strength (Fig. 3f).

Discussion

In this study, we describe an end-to-end approach that uses 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 larger population analysis. The principal advantage of this approach is the 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 on a large scale across multiple diseases and clinical conditions.

For such a pipeline to be clinically useful, it must be trustworthy, transparent and traceable to source documentation. This is especially important 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 “no chest pain,” or between an absent and a documented history of myocardial infarction, is central to valid inference. We therefore designed the extraction framework around verifiability, using provenance-linked atomic statements, adjudication against source text and calibrated thresholds to make outputs reviewable rather than opaque.

Because this approach relies on LLMs to handle the complexity of clinical documentation, trust must also be built into the extraction process itself. We therefore designed the framework around three linked principles: atomicity, extraction adjudication, and calibration. Together, these form a multi-step verification process in which independent physician adjudication of model-extracted entities yielded near-perfect inter-rater agreement (inter-rater $\kappa = 0.917–1.000$) and high extraction precision (95.5–99.5%), confirming the clinical validity of the framework at the speed and cost efficiency required for large longitudinal datasets.

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: they (i) restrict the unit of claim so that factuality can be precisely measured, (ii) make provenance traceable to specific note spans for human or automated adjudication, and (iii) enable modular verification pipelines that score, filter, or abstain on a per-statement basis. 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 (e.g., Verifact[35]). Atomicity via massive decomposition of LLM agentic processes to atomic parts can achieve perfect accuracy over millions of steps.[36] Although LLM-based adjudication may appear circular, in both the literature[10,35] and in our experience in this study, it performs on par with medical expert determination.

Validation of extracted concepts therefore requires calibration against ground truth. We address the need for validation of extracted concepts by combining per-statement judgments with probability calibration and conformal thresholds to generate principled, risk-bounded acceptance decisions. In practice, this provides confidence measures for each extracted concept and finite-sample guarantees on the fraction of accepted, non-flagged outputs.[10]

The combination of atomicity, adjudication and uncertainty quantification 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.

Our approach advances not only improved extraction, but also the use of graph database organization and storage, combined with self-improvement AI agents, to navigate the KG across computable elements at scale and speed for hypothesis testing and exploration. This KG-approach emphasizes explicit relationships and 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.

For example, medication and symptom events are linked not only to patients and visits, but also to medication concepts, enabling direct reasoning about drug–symptom associations. Adherence, phenotype, and disease-status events are similarly anchored, and delta nodes make changes over time a first-class object of study. These design choices facilitate multi-hop queries that traverse linked events, concepts and timepoints, and downstream machine learning in ways that respect clinical pathways.

The resulting process thus preserves the logic of conventional clinical research workflows while reducing the manual burden of human chart abstraction. Records are still reviewed, variables defined, cohorts constructed, analyses executed and results interpreted by investigators. What changes is that these steps can be carried out systematically across narrative text at scale, with extracted elements, intermediate outputs and analytic code available for investigator review. The result is not a black-box system that substitutes for

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

The clinical application of this approach to our GLP-1 RAs 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 fields. In our cohort, nearly 70% of HbA1c values were identified only from unstructured notes, in addition note-derived data identified baseline HbA1c for more than 6,200 patients who would have been missed by a structured-data-only strategy. Likewise, complete ascertainment of GLP-1 exposure required both structured and unstructured data, with 47.9% of events captured from structured prescription records, 33.4% from unstructured documentation, and 18.6% supported by both sources.

Although prior studies have reported differential GLP-1 effects across glycemic groups, they have generally relied on selected populations or baseline-to-follow-up comparisons at fixed timepoints.[19,20,37–40] Here, we reconstructed longitudinal trajectories and time-to-event patterns within a routine-care cohort stratified by baseline glycemic status. Weight loss was greater and more rapid in individuals with lower baseline glycemic burden, with progressively attenuated weight loss across prediabetes, T2DM and poorly controlled diabetes. HbA1c reduction showed the opposite gradient: declines were larger and occurred earlier with increasing baseline glycemic burden, with the largest reductions observed in poorly controlled diabetes and most of the reduction occurring early after initiation before plateauing. These analyses could be clinically employed to develop pathways with early touch points after GLP-1 initiation for those with higher glycemic burdens. These early touch points could then be used to guide GLP-1 RA dose escalation, switching among GLP-1 agents and intensive complementary interventions.

Using note-derived unstructured data, we were able to examine clinically relevant longitudinal patterns in depression, pain scores, waist circumference, alcohol use and muscle strength after GLP-1 initiation, variables that would otherwise require large-scale manual human chart review. While these measures were available for only a minority of patients, reflecting real-world clinical documentation practices, our analysis revealed that PHQ-9 scores improved in patients with moderate to severe depression after GLP-1 initiation.[41] The use of GLP-1 RAs in clinical depression is a rapidly emerging area of research given evidence that GLP-1 activation central nervous system decreases behaviors associated with anxiety and depression in animal studies.[41,42] Waist circumference also

declined over time after GLP-1 initiation, consistent with other reports.[43] General pain intensity showed larger decreases among patients with higher baseline pain scores, with statistically significant reductions at 6, 9 and 12 months in the subgroup with higher-pain burdens. This finding is consistent with recent reports on benefits of GLP-1RAs on pain management via neuroprotective, metabolic and inflammatory regulation.[44] In contrast, no clear longitudinal changes were observed for alcohol use or muscle strength. These findings illustrate how making these previously inaccessible clinical assessments computable can enable rapid exploration of new, patient-centered questions for future study.

This work should be interpreted in light of several limitations. The GLP-1 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. Finally, richer measurement 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. Whereas trials test prespecified hypotheses in selected populations, routine-care data can also reveal patterns of care, including which outcomes are measured and which remain 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, their promise will depend on the fidelity of the patient histories from which they learn. If important parts of the clinical journey remain inaccessible in narrative text, then these systems are built on incomplete accounts of care. Such incompleteness is not a minor technical limitation; it shapes the models themselves, and therefore the quality of the inferences, predictions and recommendations they can support. By making note-derived clinical information computable, reviewable and traceable to source evidence, this platform offers a more complete foundation for both translational research and future clinical AI.

This study presents an end-to-end approach that transforms the medical record into a computable, reviewable and longitudinal representation of care that can be reliably interrogated across multiple dimensions at speed and scale. The GLP-1 analyses provide a concrete clinical example of the trajectory-level insights that become accessible when patient histories are reconstructed more completely, while also showing that clinically relevant outcomes long trapped in narrative notes can be studied at scale without the prohibitive costs and time required for manual chart review. More broadly, the platform highlights the importance of data completeness and provenance as AI becomes increasingly embedded in clinical research and care. The longer-term significance of this trustable, transparent and traceable approach may extend beyond the description of real-world experience, providing a stronger foundation for future predictive, causal and decision-support applications within clinical research, evaluation of the quality of clinical care and be scalable to large populations.

Competing interests

Richard Foty, Edward Kim, and Vicki Seyfert-Margolis are employees of RespondHealth Inc. The work described in this manuscript involves methods developed by RespondHealth Inc.

Jay S. Skyler 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 plc, Immunothera, Integrated Nanotherapeutics, i2o Inc., Kriya Therapeutics, Novo Nordisk, Oramed, Regor, Remedy Plan, RespondHealth Inc., 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.

Charles B. Cairns reports grants or contracts from the National Institutes of Health 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 National Institutes of Health and RespondHealth Inc. (unpaid); and leadership or fiduciary roles with several non-profit organizations.

Avnesh S. Thakor 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.

Lucy F. Robinson declares no competing interests.

References

1. Kong, H.-J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res* **25**, 1–2 (2019).
2. 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. *Journal of the American Medical Informatics Association* **23**, 1007–1015 (2016).
3. 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. *medRxiv : the preprint server for health sciences* 2025.03.20.25324320 Preprint at https://doi.org/10.1101/2025.03.20.25324320 (2025).
4. Mohan, G. & Gaskin, D. J. Social determinants of health and US health care expenditures by insurer. *JAMA Network Open* **7**, e2440467 (2024).
5. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* **17**, 507–513 (2010).
6. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. in 17 (2001).
7. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nature medicine* **28**, 1773–1784 (2022).
8. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
9. Yang, X. *et al.* A large language model for electronic health records. *NPJ digital medicine* **5**, 194 (2022).
10. Kim, E., Foty, R., Shrestha, M. & Seyfert-Margolis, V. Conformal prediction and verification of large language model extractions in EHR data. in vol. 7 539–546 (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 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. *International Journal of Medical Informatics* 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. *arXiv preprint arXiv:2602.00009* (2025).
17. Girman, C. J., Ritchey, M. E. & Re III, V. L. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiology and drug safety* **31**, 717 (2022).
18. Gao, Y. *et al.* Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *Jmir Ai* **4**, e58670 (2025).
19. Cerchi, E., Santo, P. A. do E., de Oliveira, M. C., Janovsky, C. C. P. S. & Halpern, B. Effects of tirzepatide on weight management in patients with and without diabetes: a systematic review and meta-analysis. *Int J Obes (Lond)* **49**, 2415–2425 (2025).
20. Frías, J. P. *et al.* Tirzepatide versus Semaglutide Once Weekly in Patients with Type 2 Diabetes. *N Engl J Med* **385**, 503–515 (2021).
21. Wilding, J. P. *et al.* Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine* **384**, 989–1002 (2021).
22. Jastreboff, A. M. *et al.* Tirzepatide once weekly for the treatment of obesity. *New England Journal of Medicine* **387**, 205–216 (2022).
23. Marso, S. P. *et al.* Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine* **375**, 311–322 (2016).
24. Marso, S. P. *et al.* Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine* **375**, 1834–1844 (2016).
25. Waldrop, S. W., Johnson, V. R. & Stanford, F. C. Inequalities in the provision of GLP-1 receptor agonists for the treatment of obesity. *Nat Med* **30**, 22–25 (2024).
26. Nunns, M. *et al.* The quantity, quality and findings of network meta-analyses evaluating the effectiveness of GLP-1 RAs for weight loss: a scoping review. *Health Technol Assess* 1–73 (2025) doi:10.3310/SKHT8119.
27. Eisenkraft Klein, D., Zenone, M. & Kesselheim, A. S. Glucagon-Like Peptide-1 Receptor Agonists and Pay-Per-Click Direct-to-Consumer Advertising. *JAMA Netw Open* **8**, e2538718 (2025).
28. Kim, E., Foty, R. & Seyfert-Margolis, V. Quantifying the Unstructured Narrative of Patient Care in EHR Data. *Big Data Analytics & Applications* 70 (2025).
29. Lobo, M., Lamurias, A. & Couto, F. M. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International* **2017**, 8565739 (2017).
30. Yan, S. *et al.* PhenoRerank: A re-ranking model for phenotypic concept recognition pre-trained on human phenotype ontology. *Journal of biomedical informatics* **129**, 104059 (2022).

31. 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. *National Health and Nutrition Examination Survey* **3**, (2025).
32. Uzuner, Ö., Solti, I. & Cadag, E. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association* **17**, 514–518 (2010).
33. google-bert/bert-large-uncased · Hugging Face. https://huggingface.co/google-bert/bert-large-uncased (2024).
34. Helios9/BioMed_NER · Hugging Face. https://huggingface.co/Helios9/BioMed_NER.
35. Chung, P. *et al*. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672* (2025).
36. Meyerson, E. *et al*. Solving a Million-Step LLM Task with Zero Errors. *arXiv preprint arXiv:2511.09030* (2025).
37. J Rodriguez, M Goodwin, Samuel Gratzl, Rajdeep Brar, Charlotte Baker, J Gluckman, L Stucky. Semaglutide vs Tirzepatide for Weight Loss in Adults With Overweight or Obesity.. JAMA internal medicine. 184. (9). 2024. doi:10.1001/jamainternmed.2024.2525.
38. 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) doi:10.1007/s11606-025-09943-3.
39. Misra, S. Deciphering the Effects of Semaglutide Across the Glycemic Spectrum. *Diabetes Care* **47**, 1322–1324 (2024).
40. 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).
41. Chen, X., Zhao, P., Wang, W., Guo, L. & Pan, Q. The antidepressant effects of GLP-1 receptor agonists: a systematic review and meta-analysis. *The American Journal of Geriatric Psychiatry* **32**, 117–127 (2024).
42. 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).
43. Wong, H. J. *et al*. Efficacy of GLP-1 receptor agonists on weight loss, BMI, and waist circumference for patients with obesity or overweight: a systematic review, meta-analysis, and meta-regression of 47 randomized controlled trials. *Diabetes Care* **48**, 292–300 (2025).
44. He, Y. *et al*. Advances in GLP-1 receptor agonists for pain treatment and their future potential. *The Journal of Headache and Pain* **26**, 46 (2025).