Original Paper

Clinical Free-Text Retrieval Across an Acute Physiological Derangement Composite in Intensive Care Unit Discharge Summaries: Stratified Evaluation Study

Yngve Mikkelsen, MD, MSc, DBA

Saïd Business School, University of Oxford, Oxford, England, United Kingdom

Corresponding Author:
Yngve Mikkelsen, MD, MSc, DBA
Saïd Business School
University of Oxford
Park End Street
Oxford, England, OX1 1HP
United Kingdom
Phone: 44 1865 288800
Email: yngve.mikkelsen@sbs.ox.ac.uk

Abstract

Background: Retrieval over clinical free text underpins record search, cohort identification, and retrieval-augmented language systems. Performance is usually summarized by a single corpus-level mean, which can obscure systematic heterogeneity across patients.

Objective: This study tests whether known-item retrieval performance varies across an early acute physiological derangement composite measured independently of the text, whether it reproduces under a validated severity-of-illness score, and whether it is robust to alternative outcome, retrieval, and explanatory specifications.

Methods: We linked 4078 MIMIC-IV (Medical Information Mart for Intensive Care IV) discharge summaries to an adult medical intensive care unit (ICU) cohort and built a pooled, length-matched index of 2956 documents drawn equally from 4 strata of a prespecified composite of early acute physiological derangement, comprising vital signs and Glasgow Coma Scale, mechanical ventilation, and age. A held-out 2-sentence passage had to retrieve its originating summary after removal. Eight contrastively trained dense retrievers, 5 masked language model controls, and BM25 were evaluated. Reciprocal rank at 10 (RR@10) was regressed on exposure stratum with patient-clustered SEs and Holm correction. The analysis was repeated under the Oxford Acute Severity of Illness Score (OASIS), a validated first-day score. Additional analyses examined categorical exposure shape, untruncated target rank, hit@10 (hit at 10), lexical representation, query characteristics, age removal, organ dysfunction, and measured covariates.

Results: Mean RR@10 across dense retrievers fell from 0.127 to 0.080 between the lowest and highest composite stratum (−0.047; mean relative decline 37.2%), with Holm-significant negative slopes in all 8; a pooled model gave −0.0145 per stratum (95% CI −0.0207 to −0.0082; $P<.001$; 23,648 observations). Under OASIS, it fell from 0.127 to 0.078 (−0.049; mean relative decline 38.7%); again, 8 of 8 were Holm-significant. The geometric mean rank of the correct document worsened by a factor of 1.17 per stratum under the composite and 1.24 under OASIS, 8 of 8 significant. BM25 also deteriorated, although from a higher baseline. It persisted across alternative query, indexing, rank, and hit@10 specifications; joint covariate adjustment attenuated the mean coefficient by 21.6%. Removing age attenuated but did not eliminate the pattern: all 8 retained negative slopes, the mean decline fell to 29.7%, and 3 of 8 remained Holm-significant.

Conclusions: Clinical free-text retrievability declined systematically with worsening early acute physiological derangement. The association held under an externally defined and validated severity-of-illness score as well as the study-specific composite and in lexical as well as dense retrieval. Corpus-level averages can therefore conceal clinically ordered heterogeneity, and clinical retrieval systems should be evaluated across patient strata rather than overall performance alone.

(JMIR AI 2026;5:e112610) doi: 10.2196/112610

XSL•FO
RenderX

KEYWORDS

clinical information retrieval; information storage and retrieval; electronic health records; natural language processing; patient heterogeneity; intensive care units; benchmarking; MIMIC-IV; Medical Information Mart for Intensive Care IV

Introduction

Retrieval over clinical free text underpins record search, cohort identification, and retrieval-augmented language systems. Retrieval performance over a corpus is conventionally summarized by a single mean estimate, which can conceal systematic heterogeneity within the population that corpus represents.

There is reason to test heterogeneity by physiological derangement. Documentation associated with greater physiological derangement may contain more competing problems, interventions, and shifts in clinical focus, any of which could alter how well a small passage identifies the larger record. Systematic reviews have synthesized the literature on retrieval-augmented generation in health care [1,2]. Within it, electronic health record (EHR) retrieval work has concentrated on models, query types, and retrieval settings, including EHR-specific dense retrievers, single- vs multipatient benchmarks, retrieval of related discharge reports for clinical question answering, benchmarking of pooling strategies, and fine-tuning on real-world clinical documents [3-7]. Heterogeneity has also been examined at the level of the document and its context, where corpus type and query format together explained as much retrieval variance as the choice of embedding model [8]. The present study holds the retriever fixed and varies the patient. We found no prior study asking whether retrieval performance over real clinical notes varies systematically with a derangement measure derived independently of the text.

We use known-item retrieval as a controlled assay of this question. A 2-sentence passage extracted from a discharge summary must retrieve its originating summary after that passage has been removed from the target. Ground truth is unambiguous, and all queries compete against the same candidate population. The task is not intended to reproduce a deployed clinical search; it isolates whether retrievability itself varies across patient strata.

We asked 3 questions: whether retrieval performance changes across quartiles of an early physiological-derangement composite; whether any gradient is shared by dense and lexical retrieval; and whether measured clinical, documentary, representation, or query characteristics attenuate the association.

Methods

Design and Data

We evaluated 13 transformer encoders and BM25 [9] on known-item retrieval over MIMIC-IV (Medical Information Mart for Intensive Care IV) discharge summaries. No model was trained or fine-tuned for this study. Analyses were conducted from August through October 2026. Discharge summaries came from MIMIC-IV-Note (version 2.2) [10], with intensive care unit (ICU) linkage from MIMIC-IV (version 3.1) [11,12]; both resources are distributed through PhysioNet [13]. Patient strata were derived from an independently constructed adult medical ICU cohort and derangement composite.

The primary design used one pooled mixed-derangement index: all 4 strata contributed to the same candidate population, so every query was evaluated against the same candidates, and BM25 used the same queries, targets, and index. Sensitivity analyses masked other documents from the query patient, repeated the earlier per-quartile-index design, used the unmatched sample, placed all usable query-target pairs in one index without subsampling or length matching, and rebuilt the exposure without its age component.

Exposure Characterization

The exposure is a prespecified composite of early acute physiological derangement: within-cohort quartile scores (0-3, worst quartile highest) for the largest deviation of heart rate from 80 beats/min and of respiratory rate from 18 breaths/min, and for the lowest mean arterial pressure, oxygen saturation, and Glasgow Coma Scale, plus age tertile (0-2) and mechanical ventilation (0-1). Vital signs were summarized over the first 24 hours after all 5 had first been recorded; all 5 were first available in the first ICU hour for 54.0% of 5878 stays, within the first 7 hours for 97.2%, and more than 24 hours after admission for 0.6%. Ventilation was timed from ICU admission, with a ventilation procedure starting within 24 hours of admission. Because Glasgow Coma Scale values are heavily tied, its quartile scoring yielded 3 levels, scored 1-3, so the attainable range is 1-18 rather than the nominal 0-18; scoring those levels 0-2 would lower every total by 1 point and leave stratum membership unchanged. Plausibility bounds applied before hourly aggregation, cut points, missing-data handling, and stratum boundaries are given in Multimedia Appendix 1.

The composite and its component definitions were specified before any retrieval outcome from the present study was examined, for a companion time-series analysis of the same cohort. It was used here because it was already defined and independent of the text, not because it was judged to measure severity better than a published score, and prior definition does not by itself justify a single exposure, which is why the analysis is also reported under a validated score below. Because the composite is not a published severity score, we characterized rather than validated it: hospital mortality increased monotonically across quartiles, from 17.0% (334/1967) to 34.3% (360/1049), while discrimination for hospital mortality was modest (area under the receiver operating characteristic curve [AUROC] 0.597, 95% CI 0.579 to 0.613). Discrimination is likely constrained by the cohort's requirement of an ICU stay of at least 72 hours, which excludes the early deaths from which first-day scores derive much of their discriminative power; Oxford Acute Severity of Illness Score (OASIS) reached 0.619 on the same records. A partial Sequential Organ Failure Assessment (SOFA) subscore using platelet and creatinine components was computable in 98.5% of 5878 stays and was

XSL•FO
RenderX

essentially uncorrelated with the composite (Spearman ρ=0.004) while showing similar discrimination (AUROC 0.598), which is consistent with their covering different organ axes rather than disagreement about severity. We therefore interpret the exposure as a measure of early acute physiological derangement as defined here, not as a validated severity-of-illness instrument, and we do not read the gradient as a general physiological-severity effect.

The vital-sign and ventilation terms are timed differently. The 5 vital-sign and Glasgow Coma Scale components are taken from the first 24 hours after a warm-up hour in which all 5 vital signs have been recorded at least once, whereas mechanical ventilation is assessed over the first 24 hours from ICU admission, so the 2 windows coincide for stays whose warm-up hour is short and diverge where vital-sign charting is delayed.

Replication With a Validated Severity Score

Because the composite is not a published severity score, the primary analysis was repeated under the OASIS, a validated first-day severity measure [14]. OASIS was chosen over the Simplified Acute Physiology Score II because every OASIS component can be reproduced from the raw release without a derived blood-gas table. OASIS also bears on how the present composite is constructed: its 10 components are the Glasgow Coma Scale, heart rate, mean arterial pressure, respiratory rate, temperature, urine output, mechanical ventilation, elective surgery, age, and pre-ICU length of stay, so a validated score named for acute severity of illness likewise contains age and mechanical ventilation, alongside terms that are not clinical measurements at all, namely elective surgery and pre-ICU length of stay. In such scores, the word “acute” denotes the first-day window over which the state is assessed rather than a property required of each component, and the composite follows that convention.

OASIS was computed for the same 5878 stays from the raw tables against the MIMIC-Code reference definition [15], pinned at release (version 3.0.1; commit c7e07560dc847e32cbb0b2890213e8e7cbd8bc7e). All 10 components were derived locally; the 29 item identifiers were verified against the data dictionaries before extraction, and coverage was at least 94.2% for every component. Two properties of the reference definition were replicated rather than corrected so that the result matches what the standard concept produces, and both are quantified in Multimedia Appendix 1: its oxygen-delivery view discards device observations charted without an oxygen-flow value, and its ventilation view discards single–time–point events. Invasive ventilation within the first 24 hours was identified for 59.0% of 5878 stays against 2857 of 5878 (48.6%) under the composite’s own ventilation flag, with 84.4% agreement (κ=0.69).

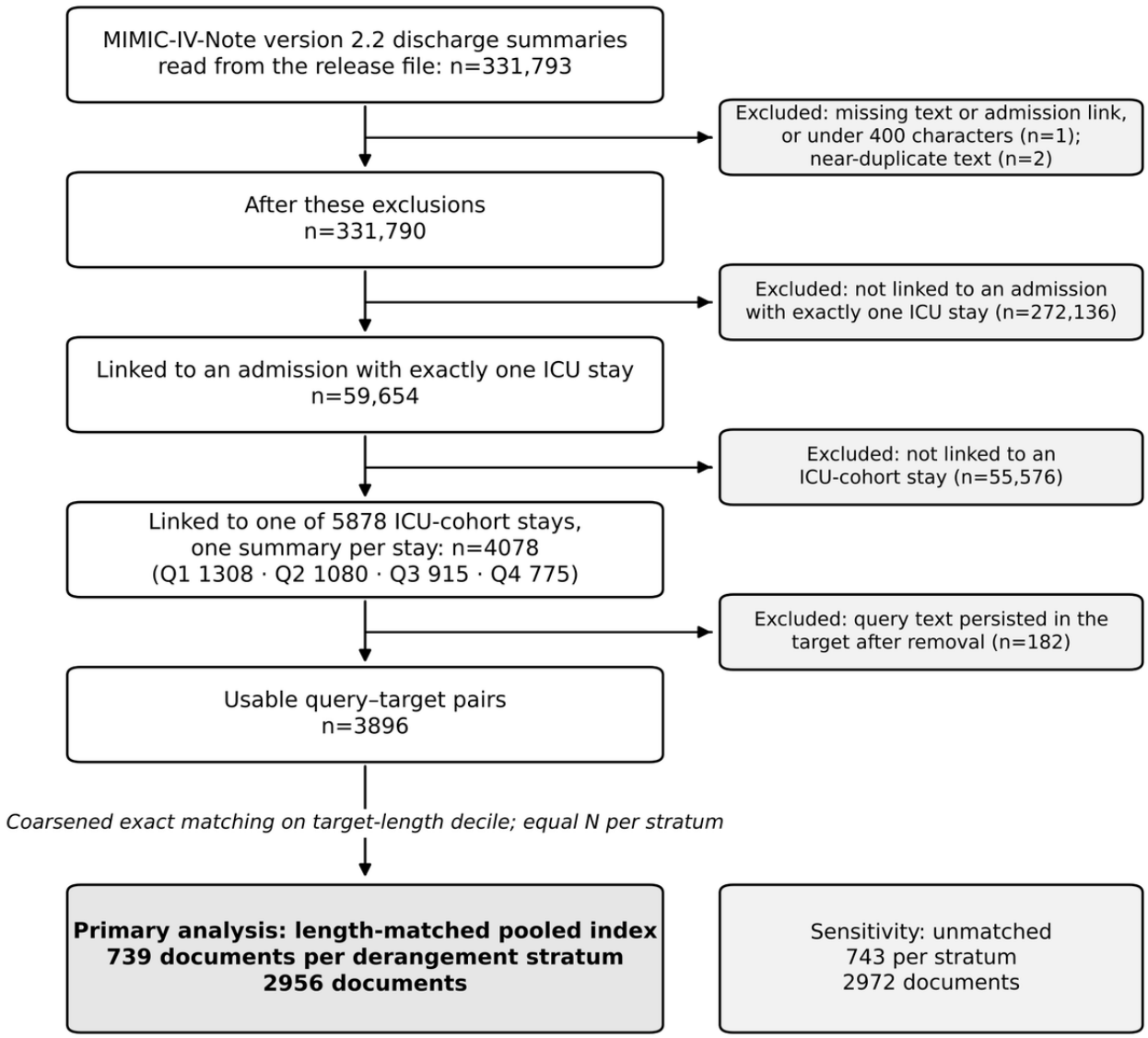
The two exposures also differ in two respects that bear on interpretation. OASIS follows the reference convention of setting the Glasgow Coma Scale to 15 when the verbal component cannot be assessed because of intubation, whereas the composite takes the lowest recorded value without that correction; and OASIS uses a fixed window from ICU admission, with a 6-hour look-back for vital signs and Glasgow Coma Scale and a strict first-24-hour window for urine output, rather than the composite’s warm-up-anchored window. Ventilation is timed from ICU admission under both. Strata were formed as quartiles of the OASIS total on the same cohort, and the resulting cells were built, indexed, and scored by the identical pipeline.

Cohort and Query Construction

The source cohort, constructed independently for a companion analysis, contained 5878 medical ICU stays of patients aged 18 years or older. From all 94,458 MIMIC-IV (version 3.1) ICU stays, it took the earliest stay of at least 12 hours in each admission, kept those whose first care unit was a medical ICU and which lasted at least 72 hours, and required at least 50 hours of observation after warm-up, truncated at 720 hours (Multimedia Appendix 1). After exclusion of admissions containing more than one ICU stay, 59,654 discharge summaries were linked to single-stay admissions, of which 4078 belonged to cohort stays, one summary per stay; query construction yielded 3896 (95.5%) usable query-target pairs. Because the composite takes integer values, its quartiles within the cohort are unequal in size (1967, 1566, 1296, and 1049 stays from the lowest to the highest), so the retained note counts (1308, 1080, 915, and 775) also fall across strata, while the proportion of stays retained rose slightly with the derangement score, from 66.5% (1308/1967) to 73.9% (775/1049; Figure 1).

XSL•FO
RenderX

Figure 1. Cohort construction. The release file contains 331,793 discharge summaries with unique note identifiers, a count confirmed with 2 independent parsers; the release documentation states 331,794. The primary analysis used the length-matched set of 739 documents per derangement stratum; the unmatched set is a sensitivity analysis. ICU: intensive care unit.



An eligible sentence contained at least 10 words, contained at least one token from a fixed, case-insensitive list used as a heuristic marker of running prose (for example, was, were, has, had, presented, denies), and appeared in no more than three documents in the 4078-note corpus, limiting templated boilerplate; the normalization used to decide when 2 sentences count as the same and the exclusion of header and field-value lines are given in Multimedia Appendix 1. The primary query was the first 2 sentences of the longest contiguous run of eligible sentences. This deterministic rule was preferred over a random eligible sentence, which yields higher variance, and over model-generated questions, which would introduce a generating model into the exposure-outcome path; the alternative-query analysis below tests sensitivity to the choice directly. The query span was removed from the target using token-space alignment, requiring at least 60% coverage in both directions; post hoc verification found no residual query-in-target leakage. Extraction success was similar across quartiles (95.3%-95.9%).

The resulting query is a verbatim passage from the target document rather than a clinician-authored information need. We use the term “query” throughout for the text submitted to the retriever, and the task is more precisely described as passage-to-source-document matching.

For alternative-query sensitivity, all nonoverlapping eligible 2-sentence windows were enumerated, and 5 were sampled without replacement (seed 42). All 5 were removed before indexing, fixing the candidate index across draws, and documents with fewer than 5 windows were excluded. Five-window eligibility among the 739 primary-analysis documents per quartile was 94.5%, 95.0%, 96.1%, and 94.2% from Q1 to Q4, with a median of 12 eligible windows in every stratum. The 5-draw analysis was built independently from the 4078 linked summaries, with 664 documents per stratum as the largest common cell size (Multimedia Appendix 2).

XSL•FO
RenderX

Indexing and Retrieval Models

Quartiles were coarsened-exact-matched on target-length decile and sampled to 739 documents each, producing a 2956-document pooled index; an unmatched set of 743 documents per quartile was retained as a sensitivity. The value of 739 was not derived from a power calculation: it is the largest cell size common to all 4 strata after length matching. Note that length proved essentially unrelated to exposure, but matching was retained in the primary design.

The 8 contrastively trained dense retrievers were BGE-base-en-v1.5 [16,17], GTE-base [18], E5-base-v2 [19], Nomic-embed-text-v1.5 [20], all-mpnet-base-v2 [21,22], all-MiniLM-L6-v2 [21,23], MedCPT [24], and BioLORD-2023 [25]. MedCPT and BioLORD are biomedically specialized but not pretrained on clinical documentation. Five nonretrieval masked language model encoders (BERT-base-uncased [26], BioBERT-v1.1 [27], ClinicalBERT [28,29], BiomedBERT [30], and SciBERT [31]) were included only as negative controls; they were standardized to CLS pooling because they have no canonical sentence-pooling convention, which may depress their absolute performance, and they are therefore read as a performance floor and not compared with the contrastive panel. The ClinicalBERT checkpoint is medicalai/ClinicalBERT [28], whose model card requests citation of the clinical application paper cited here [29] rather than a model-description paper; it is distinct from other published encoders that carry the same name. Model-specific prefixes and pooling conventions were used where documented; for MedCPT, its query encoder was applied to queries and its article encoder to target chunks. Embeddings were L2-normalized.

Targets were encoded as overlapping 510-token windows with stride 384. Where a model specifies a document-side instruction prefix, the prefix was tokenized separately and prepended to every window, with the content carried by each window reduced by the prefix length so that no encoded unit exceeded the model's maximum sequence length. Document score was the maximum cosine similarity between the query vector and any target chunk. Two-sentence queries were encoded as single vectors. BM25 was scored on the same target texts after query removal.

Outcomes and Inference

The primary retrieval outcome was reciprocal rank at 10 (RR@10); recall at 10 was a secondary, descriptive outcome. RR@10 was regressed on exposure quartile as an ordinal term (1-4) separately for each model using CR1 patient-clustered SEs [32]. Dense-model *P* values were Holm-adjusted [33]; 95% CIs and monotonicity across all 4 strata were also reported.

To locate where rank deterioration occurred, we recorded the target score, highest nontarget score, target-minus-best-distractor margin, and number of nontargets scoring above the target. Dense and BM25 slopes were compared both in absolute reciprocal rank (RR) units and after normalization by each method's Q1 mean. Paired patient-clustered bootstrap resampling (2000 replicates) preserved correlation across methods.

RR@10 denotes the per-query outcome and mean RR@10 its mean within a stratum; the distinction is kept throughout. The ordinal specification was supplemented by three further analyses of the same outcome. A categorical model entered quartile as three indicators against the lowest stratum, giving the absolute difference at each stratum; departure from linearity was tested by a 2-degree-of-freedom Wald restriction of those coefficients to the straight line through them, and a single contrast compared the lowest stratum against the remaining three. Because RR@10 is bounded and 0 for most queries, two further outcomes were analyzed so that information outside the top 10 contributes. The untruncated target rank was regressed on stratum as log(rank), whose coefficient exponentiates to the multiplicative change in the geometric mean rank, and in ordinal and categorical form, the latter giving each stratum's rank ratio against the lowest; hit@10 (hit at 10) was modeled by logistic regression. Both carried the same patient-clustered covariance as the primary model. The untruncated rank was also compared across strata by a Mantel-Cox log-rank test with a score-weighted test for trend, which assumes independent observations and is therefore reported only as a distributional check; with approximately one document per patient, the dependence it ignores is small. Holm correction across the 8 retrievers was applied to all three. Finally, a pooled model across all 8 retrievers estimated a single common slope, with retriever fixed effects and patient-clustered SEs; a Wald test of the quartile-by-retriever interaction assessed slope heterogeneity. Retriever was entered as a fixed effect because the 8 retrievers constitute the finite panel of interest rather than a sample from a wider population; the patient-clustered SEs account for the repeated scoring of the same query across retrievers.

Alternative Explanations and Sensitivity Analyses

The primary analysis is the Holm-corrected RR@10-on-quartile regression in the pooled length-matched index. That design was not prespecified: the earliest analyses scored each stratum against its own index, and the pooled design was adopted after those analyses had been run, because separate indexes make the candidate population vary with the exposure. Length matching within the pooled design was specified in advance. All remaining analyses reported below are sensitivity or alternative-explanation analyses and are labeled as such; Multimedia Appendix 3 summarizes which candidate each addresses.

Alternative explanations were grouped as design, exposure definition, patient complexity, documentation process, document and representation characteristics, and query construction. Patient-complexity variables were distinct *ICD* (*International Classification of Diseases*) diagnoses, distinct prescribed drug names, distinct *ICD* procedures, and the Charlson comorbidity index using the Quan *ICD-9/10* (*International Statistical Classification of Diseases, Ninth Revision/International Statistical Classification of Diseases, Tenth Revision*) mapping [34]; documentation variables were discharge-summary chart-to-store lag and note index within the admission; document and representation variables were target length and mean pairwise cosine among a document's own chunk vectors (semantic dispersion), with chunk count entered alongside

XSL•FO
RenderX

dispersion; and query variables were source-note character offset divided by note length and query length per stratum.

Approximately 12% of documents in the pooled index (362/2956) came from patients who contributed more than one, and because narrative blocks are frequently copied forward between notes for the same patient, a nontarget note could be retrieved instead of the exact source document, so every model was also scored with same-patient nontargets masked out of the candidate set.

As a further exposure sensitivity, patients were restratified using a laboratory-only organ-dysfunction score formed from the coagulation, liver, and renal components of the SOFA (range 0-12) over the first 24 hours after ICU admission. This score shares no input variable with the primary composite, so it tests a distinct axis rather than replicating the primary analysis; the cardiovascular and central nervous system components were excluded because they use mean arterial pressure and Glasgow Coma Scale, both composite inputs. Complete three-component scores were available for 4039 of 5878 (68.7%) stays, so this analysis uses a smaller complete-case population. RR was regressed on both the score quartile and the score value. Because that analysis differs from the primary one in sample and candidate index as well as in exposure definition, both exposures were additionally evaluated on identical records within a single index, with the per-model difference between the two slopes estimated by a paired patient-clustered bootstrap (2000 replicates).

Each of the 7 covariates (the 4 patient-complexity measures, the 2 documentation measures, and target length) was entered separately and then jointly into RR-on-quartile models on the same complete-case sample (n=2753), with all covariates standardized. Attenuation was the proportional reduction in the derangement coefficient relative to the unadjusted model. Variance inflation factors (VIFs) were calculated for the joint model. These adjusted models were exploratory; their *P* values were not Holm-corrected. Because several covariates can occur downstream of early physiological derangement, coefficient attenuation is not interpreted as a causal decomposition.

The Charlson index was rebuilt after two implementation errors were identified: weights had been summed per weight class rather than per condition, and Charlson hierarchical exclusions had not been applied. The corrected index sums distinct conditions and supersedes mild liver disease by severe, uncomplicated diabetes by complicated, and malignancy by metastatic disease; it spans 0-14 with a median of 3. Code sets were verified against the MIT Laboratory for Computational Physiology (MIT-LCP) MIMIC-Code Charlson reference implementation [35] of the Quan *ICD-9-CM* (*International Statistical Classification of Diseases, Ninth Revision, Clinical Modification*)/*ICD-10* mapping, with no reference code left uncaptured. The reference implementation also adds an age score of 0 to 4, which we omit because age is a component of this study's exposure.

Four further analyses were added. Retrieval was summarized by vital status within each stratum, and in-hospital death, discharge disposition, and hospital and ICU length of stay were added to the adjustment set; because all four are measured after the exposure window and may lie on downstream pathways, that model is reported as a postexposure sensitivity analysis, attenuation in it is given no causal interpretation, and the unadjusted coefficient remains the estimate of interest. Query length was also computed for the primary query set. The lexical baseline was re-scored on windowed documents matching the dense scoring geometry, each document taking its best window score, under both window-level and document-level inverse document frequency and at 2 window sizes, with differences between representations tested by a paired patient-clustered bootstrap on identical queries. Finally, the 3 retrievers whose documented query and document instructions differ were rescored with the document instruction applied to the query, and with none, holding document representations fixed so that only the query side varied.

Reproducibility

All random operations used seed 42. Analyses ran under Python (version 3.9.13) with torch (version 2.2.0), transformers (version 4.50.3), NumPy (version 1.26.4), pandas (version 2.2.3), and SciPy (version 1.13.1); the lexical baseline used rank_bm25 (BM25Okapi) with k1=1.5 and b=0.75. Encoder weights were loaded from pinned Hugging Face revisions, listed per model in Multimedia Appendix 1 together with the query and document instruction prefixes; the revisions are the commit hashes recorded in the refs/main file of the local Hugging Face cache. For 4 of the 5 masked language model controls, an earlier snapshot also remained in that cache, so for those 4 the revision used cannot be established from the cache alone; the ambiguity affects only the negative controls, which serve as a performance floor rather than as estimates. The cluster-robust covariance, Holm correction, paired bootstrap, log-rank, and logistic fits were implemented directly on NumPy and pandas, with no statistical-modeling package in the inferential path. Cell definitions, per-query result vectors, embedding caches, and result dumps are archived locally; code availability is described under the Data Availability section.

Ethical Considerations

This study is a secondary analysis of existing deidentified clinical data. No new data were collected, no participants were recruited, contacted, or compensated, and no identifiable private information was accessed. Only aggregate results are reported, and neither the manuscript nor its supplementary material contains images of individuals or material from which any individual could be identified.

MIMIC-IV (version 3.1) and MIMIC-IV-Note (version 2.2) are credentialed resources distributed under the PhysioNet Credentialed Health Data Use Agreement; their source collection and resource creation were reviewed by the Beth Israel Deaconess Medical Center IRB, which granted a waiver of informed consent and approved data sharing. All processing was performed locally by the credentialed author; no note text was transmitted to any third-party service, and no reidentification was attempted. Because the work used only existing deidentified data, no additional ethics-review application was submitted, and no application or case number therefore exists; the governance basis is the access controls,

XSL•FO
RenderX

data use agreement, deidentification, and original institutional approvals provided by the data custodian.

Results

Primary Retrieval Gradient

Cohort construction and the exclusions at each step are shown in Figure 1.

Discharge summaries ranged from 2157 to 50,223 characters, but length was unrelated to exposure (Spearman ρ=−0.005; n=3896), so the gradient is not explained by document length.

In the pooled 2956-document index, all 8 contrastively trained retrievers performed worse as derangement increased (Table 1). Mean RR@10 fell from 0.127 in the lowest quartile to 0.080 in the highest, a difference of −0.047, with Q1-to-Q4 relative declines from 29.1% to 43.9% and a mean of 37.2%; all 8 ordinal slopes remained significant after Holm correction, and 5 were strictly monotone. Model-level slopes and their CIs are shown in Figure 2 [9,16-25].

Recall@10 changed in the same direction. Nomic declined from 0.306 to 0.196 and BGE from 0.260 to 0.166; BM25 declined from 0.517 to 0.459 (Table 2).

Table 1. Mean reciprocal rank at 10 by derangement quartile in the pooled mixed-derangement index (2956 documents)[a].

Model	Q1	Q2	Q3	Q4	Δ%	β (95% CI)	*P* (Holm)	Mono
BGE	0.1439	0.1009	0.0923	0.0826	−42.6	−0.0192 (−0.0281 to −0.0104)	<.001	Yes
GTE	0.1512	0.1148	0.1024	0.0975	−35.5	−0.0173 (−0.0266 to −0.0081)	.001	Yes
E5	0.1136	0.0945	0.0770	0.0803	−29.3	−0.0117 (−0.0200 to −0.0035)	.01	No
Nomic	0.1689	0.1315	0.1296	0.1197	−29.1	−0.0150 (−0.0250 to −0.0049)	.01	Yes
MPNet	0.1248	0.0887	0.0952	0.0790	−36.7	−0.0131 (−0.0215 to −0.0047)	.009	No
MiniLM	0.1280	0.0920	0.0861	0.0719	−43.9	−0.0174 (−0.0256 to −0.0093)	<.001	Yes
MedCPT	0.0902	0.0473	0.0546	0.0544	−39.7	−0.0100 (−0.0169 to −0.0031)	.01	No
BioLORD	0.0961	0.0718	0.0695	0.0567	−41.0	−0.0121 (−0.0194 to −0.0047)	.006	Yes
Mean (dense)	0.1271	0.0927	0.0883	0.0803	−36.8	—[b]	—	—
BM25	0.3908	0.3480	0.3449	0.3384	−13.4	−0.0160 (−0.0304 to −0.0017)	.03	Yes

[a]Mono indicates strict monotonicity across all 4 strata. *P* values are Holm-adjusted across the 8 dense models; the BM25 row is contextual and was not included in the correction family. The Δ% in the mean (dense) row is the decline in the across-model mean (−36.8%); the mean of the 8 model-specific declines is −37.2%, and that is the quantity quoted in the text.

[b]Not available.

XSL•FO
RenderX

Figure 2. Change in reciprocal rank at 10 per derangement quartile, by model. Points are slopes of reciprocal rank at 10 on the derangement quartile in the pooled 2956-document index, with 95% CIs from patient-clustered SEs. The dotted line marks the mean dense slope; BM25 is shown below the rule.

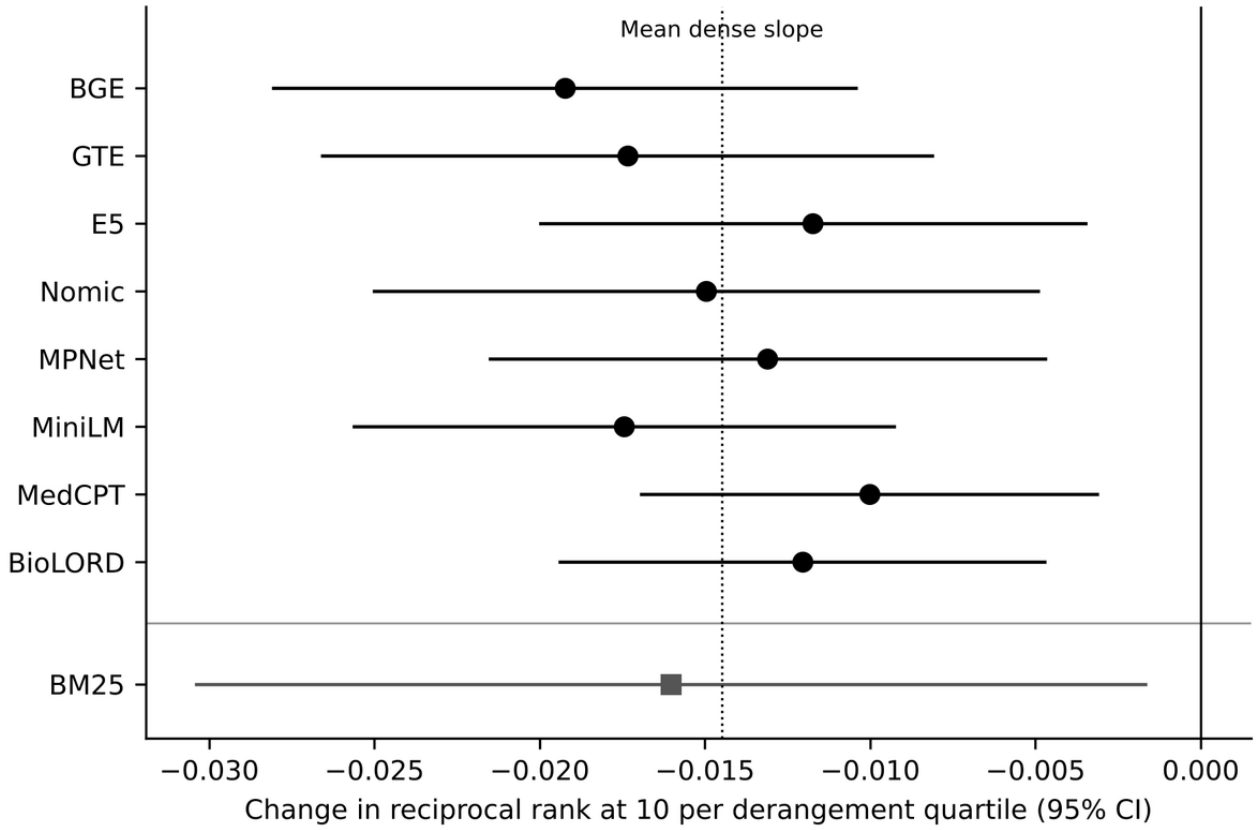


Table 2. Recall at 10 by derangement quartile in the pooled index[a].

Model	Q1	Q2	Q3	Q4	Δ%
BGE	0.260	0.204	0.192	0.166	−35.9
GTE	0.279	0.210	0.206	0.179	−35.9
E5	0.214	0.183	0.164	0.153	−28.5
Nomic	0.306	0.246	0.230	0.196	−35.8
MPNet	0.214	0.184	0.181	0.175	−18.4
MiniLM	0.245	0.184	0.176	0.154	−37.0
MedCPT	0.185	0.116	0.127	0.130	−29.9
BioLORD	0.195	0.133	0.137	0.108	−44.4
BM25	0.517	0.498	0.470	0.459	−11.3

[a]Percentage changes were calculated from unrounded values.

Replication With a Validated Severity Score

OASIS and the composite share 6 of OASIS's 10 input variables, namely age, Glasgow Coma Scale, heart rate, mean arterial pressure, respiratory rate, and mechanical ventilation, and correlate moderately with one another (Spearman ρ=0.488 between the OASIS total and the composite quartile). On identical records, OASIS discriminated against hospital mortality slightly better than the composite (AUROC 0.619 vs 0.597). Both values are modest, which may partly reflect the requirement for ICU stays of at least 72 hours and the consequent exclusion of early deaths, from which first-day severity scores derive much of their discrimination. The OASIS value therefore provides context for interpreting the composite's discrimination within this selected cohort rather than a comparison against an unrestricted population.

Under OASIS strata, all 8 dense retrievers again performed worse as the score increased (Table 3). Mean RR at 10 fell from 0.127 to 0.078, an absolute difference of −0.049, with a mean relative decline of 38.7% and all 8 slopes Holm-significant; 4 were strictly monotone. The unmatched OASIS sample gave

XSL•FO
RenderX

−39.1%, also with 8 of 8 Holm-significant. The composite's own estimate was −37.2%, so the 2 exposures agree to within 2 percentage points on the relative scale and to within 0.002 reciprocal-rank units on the absolute scale. Because those shared inputs include age and mechanical ventilation, this is a convergent replication under an externally defined and validated severity score rather than a component-isolation analysis; the age-free reconstruction below tests whether the gradient depends on the independently scored age component.

Table 3. Mean reciprocal rank at 10 by quartile of the Oxford Acute Severity of Illness Score in the pooled index (3448 documents)[a].

Model	Q1	Q2	Q3	Q4	Δ%	β (95% CI)	*P* (Holm)	Mono
BGE	0.1457	0.1003	0.0749	0.0872	−40.1	−0.0201 (−0.0283 to −0.0119)	<.001	No
GTE	0.1520	0.1153	0.0929	0.1035	−31.9	−0.0168 (−0.0256 to −0.0081)	.001	No
E5	0.1302	0.0973	0.0692	0.0693	−46.8	−0.0211 (−0.0288 to −0.0133)	<.001	No
Nomic	0.1714	0.1427	0.1080	0.1165	−32.0	−0.0199 (−0.0291 to −0.0108)	<.001	No
MPNet	0.1061	0.1001	0.0828	0.0789	−25.7	−0.0099 (−0.0173 to −0.0025)	.009	Yes
MiniLM	0.1310	0.1043	0.0784	0.0661	−49.6	−0.0221 (−0.0297 to −0.0145)	<.001	Yes
MedCPT	0.0879	0.0540	0.0509	0.0502	−42.9	−0.0116 (−0.0179 to −0.0053)	.001	Yes
BioLORD	0.0904	0.0815	0.0630	0.0536	−40.8	−0.0129 (−0.0197 to −0.0062)	.001	Yes
Mean (dense)	0.1268	0.0994	0.0775	0.0782	−38.4	—[b]	—	—
BM25	0.4052	0.3668	0.3402	0.3159	−22.0	−0.0295 (−0.0427 to −0.0162)	<.001	Yes

[a]Mono indicates strict monotonicity across all 4 strata. *P* values are Holm-adjusted across the 8 dense models; the BM25 row is contextual and was not included in the correction family. Cells contain 862 documents per stratum. As in Table 1, the Δ% in the mean (dense) row is the decline in the across-model mean (−38.4%), while the mean of the 8 model-specific declines, −38.7%, is the quantity quoted in the text.

[b]Not available.

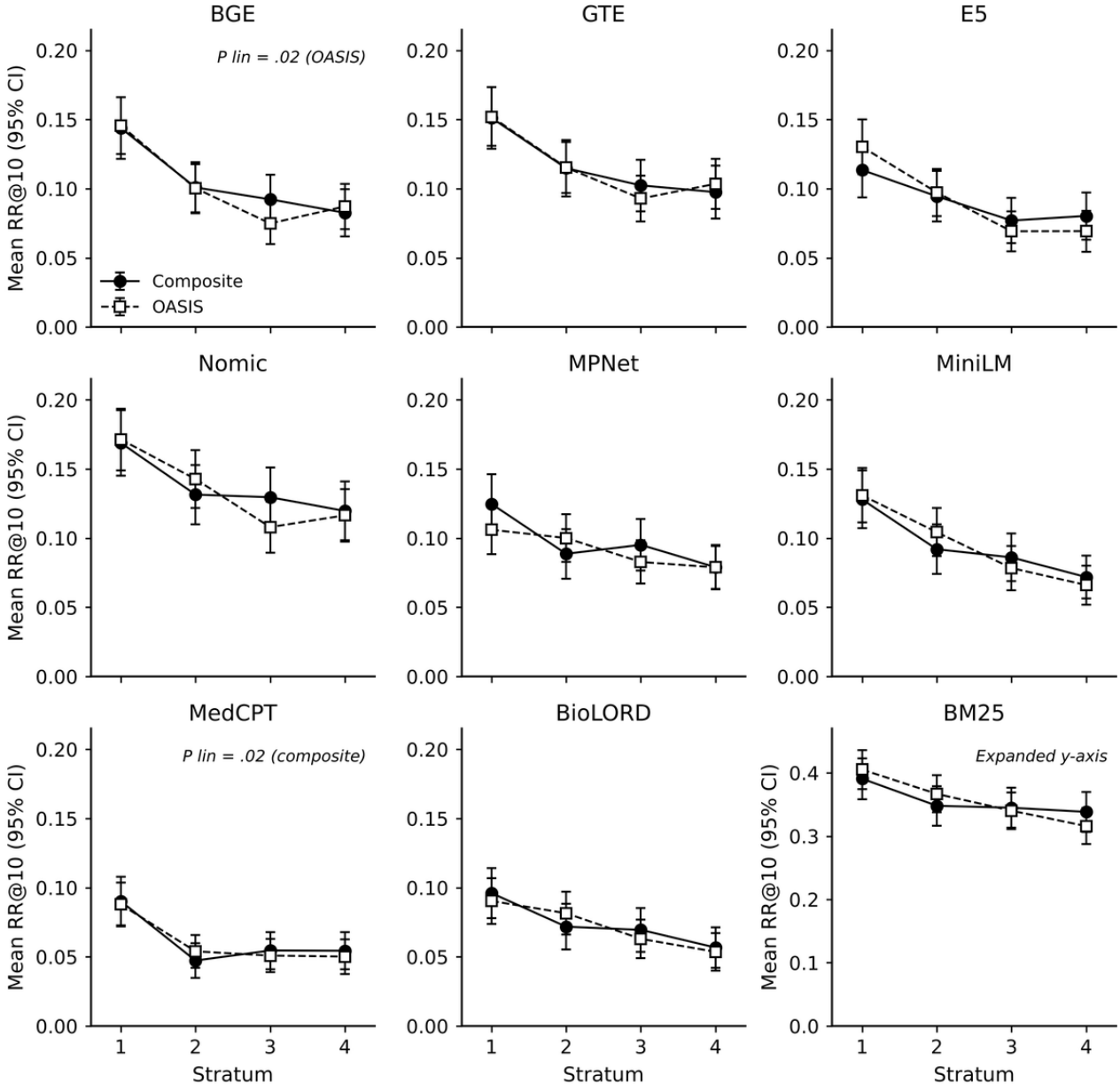
Shape of the Association

Entered categorically, the difference from the lowest stratum was negative at every higher stratum in every retriever, reaching −0.033 to −0.061 at the highest stratum under the composite and −0.027 to −0.065 under OASIS (Figure 3), and was largest at the highest stratum in 6 of 8 retrievers under the composite and 4 of 8 under OASIS. A 2-degree-of-freedom test of departure from linearity was rejected in 1 of 8 retrievers under each exposure after Holm adjustment, MedCPT under the composite (*P*=.02), and BGE under OASIS (*P*=.02).

The single contrast of the lowest stratum against the other three was −0.0520 to −0.0296 reciprocal-rank units across retrievers under the composite (mean −0.0400; all eight 95% intervals excluded 0, the widest being Nomic's −0.0686 to −0.0154) and −0.0582 to −0.0189 under OASIS (mean −0.0418), Holm-significant in 8 of 8 retrievers under the composite and 7 of 8 under OASIS, the exception being MPNet (−0.0189, 95% CI −0.0389 to +0.0011; *P*=.06). It recovered 76% to 106% of the full lowest-to-highest difference under the composite and 66% to 99.6% under OASIS, where a straight line implies 67%, so somewhat more of the decline occurs at the first stratum boundary than a linear term implies.

XSL•FO
RenderX

Figure 3. Mean reciprocal rank at 10 by stratum under both exposures, with 95% CIs. Points are per-stratum means of reciprocal rank at 10 with 95% CIs from patient-clustered SEs, one panel per contrastively trained retriever, and one for BM25. Filled circles and solid lines are quartiles of the acute physiological derangement composite; open squares and broken lines are quartiles of the OASIS. Each y-axis begins at 0; BM25's is expanded, with its baseline being about three times the dense baseline. The annotated dense panels are those in which the 2-degree-of-freedom test of departure from linearity was rejected after Holm adjustment. BGE: BGE-base-en-v1.5; BioLORD: BioLORD-2023; BM25: Best Match 25; E5: E5-base-v2; GTE: GTE-base; Nomic: Nomic-embed-text-v1.5; MedCPT: MedCPT biomedical retriever; MiniLM: all-MiniLM-L6-v2; MPNet: all-mpnet-base-v2; RR@10: reciprocal rank at 10. The single contrast of the lowest stratum against the other 3 was −0.0520 to −0.0296 reciprocal-rank units across retrievers under the composite (mean −0.0400; all eight 95% intervals excluded 0, the widest being Nomic's −0.0686 to −0.0154) and −0.0582 to −0.0189 under OASIS (mean −0.0418), Holm-significant in 8 of 8 retrievers under the composite and 7 of 8 under OASIS, the exception being MPNet (−0.0189, 95% CI −0.0389 to +0.0011; *P*=.06). It recovered 76% to 106% of the full lowest-to-highest difference under the composite and 66% to 99.6% under OASIS, where a straight line implies 67%, so somewhat more of the decline occurs at the first stratum boundary than a linear term implies.

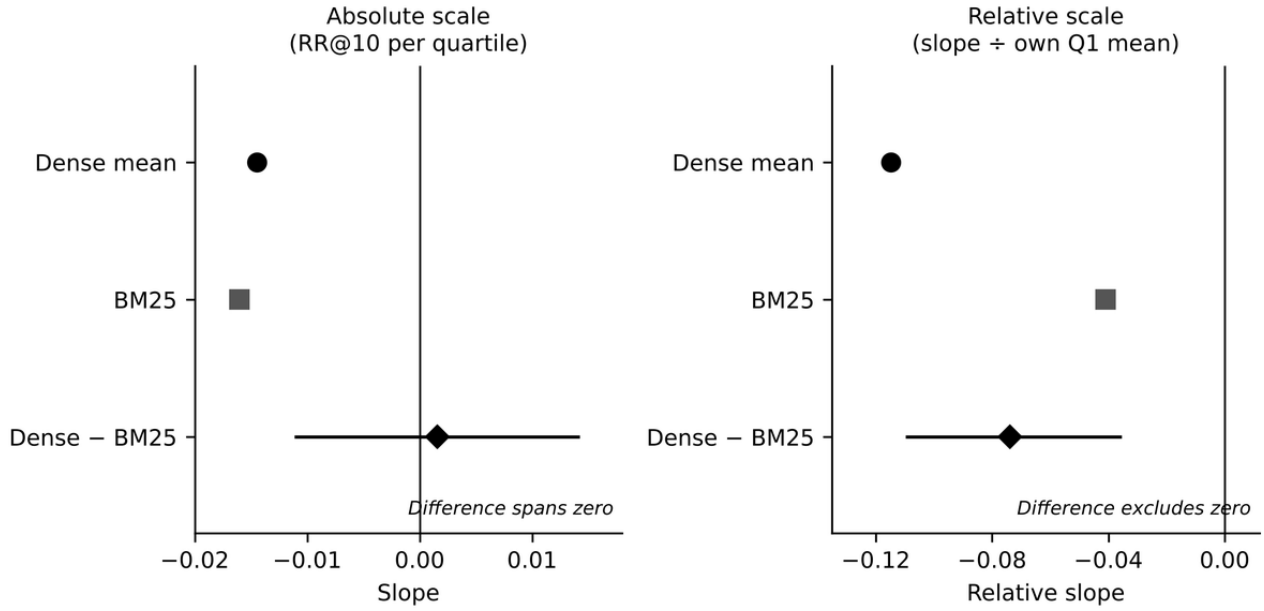


Dense Versus Lexical Retrieval

BM25 also declined monotonically, from RR 0.3908 in Q1 to 0.3384 in Q4 (13.4%; β −0.0160, 95% CI −0.0304 to −0.0017; *P*=.03), so the phenomenon is not unique to neural representations. In absolute RR units, no difference was detected between the mean dense slope (−0.0145 per quartile) and the BM25 slope (difference +0.0016, 95% CI −0.0110 to +0.0141). On the relative scale, the dense family declined more steeply, with −0.115 vs −0.041, a difference of −0.074 (95% CI −0.109 to −0.036); the ratio estimate was 2.80 (95% CI 1.49 to 13.38), but because that interval is wide and right-skewed, the difference in relative slopes is the more stable contrast. BM25 outperformed every dense model in every stratum, and this study does not claim that either should be preferred for deployment (Figure 4).

XSL•FO
RenderX

Figure 4. Dense vs lexical decline on absolute and relative scales. Left, slopes in reciprocal-rank units per quartile; right, the same slopes divided by each method's quartile-1 mean. Dense minus BM25 differences carry 95% CIs from a paired patient-clustered bootstrap. RR@10: reciprocal rank at 10.



Retrieval Deterioration Is Predominantly Target-Side

The target-minus-best-distractor margin fell with the derangement score in all 8 dense retrievers, with CIs excluding 0 in 6, and the number of nontargets scoring above the target increased substantially (Multimedia Appendix 4). The target score fell in all 8 models. The best nontarget rose in 7, and in 6 of those the target's fall exceeded that rise, by ratios from 1.8 to 10.6; the exception was MPNet, whose best nontarget rose by about four times the target's fall. In MedCPT, the best nontarget did not rise. The rank loss is therefore predominantly associated with reduced target-to-query similarity rather than with increased nearest-distractor competition. This is a descriptive decomposition, not a mechanism.

Robustness to Alternative Explanations and Covariate Adjustment

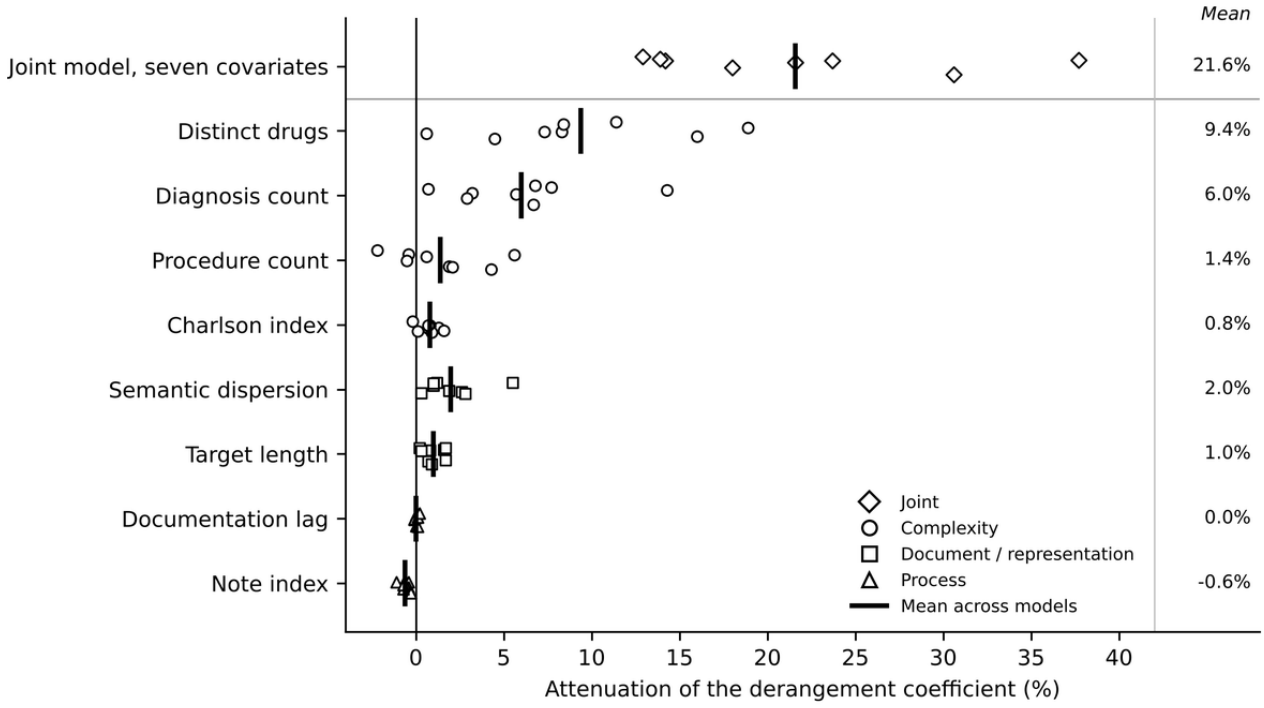
No single measured covariate attenuated the mean dense-model derangement coefficient by more than 9.4% on the complete-case sample: distinct drugs 9.4%, diagnoses 6.0%, procedures 1.4%, target length 1.0%, Charlson 0.8%, documentation lag 0.0%, and note index −0.6%. Joint entry of the 7 covariates attenuated the coefficient by 21.6% on average, ranging from 12.9% for MiniLM to 37.7% for BioLORD, and 6 of 8 retrievers retained nominal $P<.05$; MedCPT ($P=.055$) and BioLORD ($P=.08$) did not. These P values are exploratory and unadjusted for multiplicity, and VIFs were ≤1.9. The attenuation should not be read as a causally explained fraction.

Masking same-patient nontargets from the candidate set did not attenuate the gradient: no model slope moved by more than 0.0002, and the mean relative decline was unchanged. Competition from other notes belonging to the same patient therefore does not materially explain the gradient. That does not exclude every way in which copied narrative might affect a query or its source document.

Mean within-document chunk cosine was flat across strata in 7 of 8 dense models; adjustment for semantic dispersion, chunk count, and length attenuated the derangement coefficient by 2% on average. Alternative query sampling also preserved the finding: five-window eligibility was 94.2%-96.1% of primary-analysis documents in each quartile, and all 40 model-by-draw slopes were negative (Multimedia Appendix 5). Query length was flat across strata (Multimedia Appendix 2) and query position did not vary with the derangement score ($P=.97$). Independently of that score, later-originating passages retrieved less well in 7 of 8 dense models, by a margin comparable to the derangement slopes. These analyses narrow, but do not identify, the pathway. The attenuation profile across candidates is shown in Figure 5.

XSL•FO
RenderX

Figure 5. Attenuation of the derangement coefficient by candidate explanation. Points are the 8 dense retrievers, and vertical bars are the mean across them. Values are from the complete-case sample where applicable; semantic dispersion was estimated on the full sample. Design candidates, controlled by construction, and same-patient competition, assessed by masking, are not shown.



Exposure and Model-Family Sensitivities

Restratification using the laboratory-only organ-dysfunction score produced 523 documents per quartile (2092 pooled), with median scores of 0, 2, 3, and 6 across strata. The score had near-zero rank correlation with the primary composite quartile (Spearman ρ=−0.010). All 8 dense retrievers had negative quartile slopes, and 7 of 8 had negative slopes on the score value, but only BioLORD survived Holm correction under either parameterization (value −0.0070, 95% CI −0.0109 to −0.0031); BM25 was flat (+0.001, *P*=.90). Retrieval was similar across the lower 3 strata and lower only in the highest (Multimedia Appendix 6).

That analysis differs from the primary one in sample and candidate index as well as in exposure, so both exposures were also evaluated on identical records within the same 2092-document index. Both were negative in all 8 models, and each reached nominal significance in one model, though not the same one, and the per-model difference between the composite and organ-dysfunction slopes included 0 in every model (across-model mean difference −0.0027, reported descriptively because the paired intervals were estimated per model). The composite's own slope was attenuated in this subsample (mean −0.0087, against −0.0145 in the primary analysis), so the contrast between the 2 original analyses cannot be attributed to exposure definition alone. Evidence for an organ-dysfunction association was model-dependent, and specificity to the composite was not established (Multimedia Appendix 7).

Removing age from the exposure and rebuilding the strata reassigned 23.8% of stays and reduced the matched denominator to 696 per quartile. All 8 dense models still declined; the mean Q1-to-Q4 relative decline was 29.7% vs 37.2% in the primary analysis. Seven models remained nominally significant, and 3 remained significant after the Holm adjustment. The relative dense-vs-BM25 slope difference persisted (−0.058, 95% CI −0.099 to −0.016), while the absolute difference again included 0. The gradient is therefore not wholly attributable to age, although age contributes materially to its magnitude and to per-model statistical significance. Full model-level results are reported in Multimedia Appendix 8.

The further sensitivities specified in the Methods section were directionally consistent, with model-level results in Multimedia Appendix 9. Mean relative declines were 31.5% in the unmatched sample (743 documents per quartile; 2972 pooled; 4 of 8 Holm-significant), 27.5% and 24.4% in the earlier separate-index design with and without length matching, and 33.4% with all 3896 usable pairs in a single unmatched pooled index of unequal strata (1247, 1033, 873, and 743 documents), where 8 of 8 survived Holm adjustment. BM25 declined by 11.5%, 11.7%, 8.9%, and 13.7% in those 4 designs, and the absolute dense-vs-BM25 slope difference again included 0 in each. In the all-pairs design, per-model slopes were close to the primary estimates, and the separability pattern was reproduced, with margins falling in all 8 dense models while the best nontarget score moved by at most 0.004. The direction therefore persists without matching or equal-size subsampling. The magnitude differs, but those comparisons change sampling and index composition as well as matching, so the difference cannot be attributed to any one of them.

The mean decline varies across these designs from 24.4% to 37.2%, and the count of Holm-significant models in the pooled designs ranges from 4 to 8. The pooled length-matched design is treated as primary because it holds index size, distractor

XSL•FO
RenderX

composition, and document length constant simultaneously, and it is also the design that yields the largest estimate. All 5 are reported together in Multimedia Appendix 9.

The 2 biomedically specialized dense retrievers, MedCPT and BioLORD, had the lowest absolute performance and among the steepest proportional declines. The 5 masked language model encoders performed only several-fold above random ranking, showed no interpretable gradient, and served only as a floor.

Vital Status, Outcome Adjustment, and Query Length

Because hospital mortality rose across strata, retrieval was also summarized by vital status. Within strata, patients who died were retrieved less well than survivors in most retrievers, but the difference was Holm-significant in 3 of 8 under the composite (BGE, E5, and BioLORD) and in 1 of 8 under OASIS (BioLORD). Adding in-hospital death, discharge disposition, and hospital and ICU length of stay to the adjustment set attenuated the derangement coefficient by a mean of 29.7% under the composite and 22.7% under OASIS, leaving 5 of 8 and 7 of 8 coefficients nominally significant. All 4 added variables are measured after the exposure window and may lie on downstream pathways, so this is a postexposure outcome-adjustment sensitivity analysis: attenuation in it is given no causal interpretation and is not evidence against the unadjusted estimate. Mean RR at 10 by vital status within each stratum, with the conditional death coefficients, is given in Multimedia Appendix 10.

Query length did not vary across strata in the primary query set: 38.7, 38.4, 39.2, and 38.4 words from the lowest to the highest stratum under the composite (−0.01 words per stratum, 95% CI −0.42 to +0.41; *P*=.98) and 39.0, 38.7, 38.6, and 38.5 words under OASIS (−0.16, 95% CI −0.55 to +0.23; *P*=.42). The earlier finding for the five-draw subset therefore also holds for the queries used in the primary analysis.

Outcome Specification, Pooled Estimation, and Representation

Because RR at 10 is 0 for most queries, the untruncated rank was analyzed directly. On the log scale, the geometric mean rank of the correct document rose by a factor of 1.17 per stratum under the composite (1.12-1.20 across retrievers) and 1.24 under OASIS (1.17-1.29), positive and Holm-significant in 8 of 8 retrievers under both (*P*≤.002); averaged across retrievers, the geometric mean rank moved from 71 to 123 under the composite and from 75 to 140 under OASIS. Entered categorically, the highest stratum sat 1.48 to 1.89 times further down the ranked list than the lowest under the composite and 1.57 to 2.09 times under OASIS. The median rank approximately doubled, with a mean ratio of 1.96 under the composite and 2.08 under OASIS, and a Mantel-Cox trend test on the same ranks was Holm-significant in 7 of 8 retrievers under the composite, except E5, and 8 of 8 under OASIS (z −3.43 to −6.14). Hit@10 fell correspondingly, with odds ratios per stratum of 0.81 to 0.92 under the composite and 0.79 to 0.89 under OASIS, Holm-significant in 7 of 8 and 8 of 8, respectively. The association is therefore not an artifact of truncating rank at 10, nor of applying a linear model to a bounded outcome.

A single pooled model across all 8 retrievers, with retriever fixed effects and patient-clustered SEs, estimated a common slope of −0.0145 per stratum under the composite (95% CI −0.0207 to −0.0082; *P*<.001; 23,648 observations, 2751 patients) and −0.0168 under OASIS (95% CI −0.0226 to −0.0110; *P*<.001; 27,584 observations, 3164 patients). The quartile-by-retriever interaction was not significant under the composite (Wald χ^2 =7.3; *P*=.40) but was significant under OASIS (χ^2 =18.7; *P*=.009), and the rank order of the eight per-model slopes was only moderately preserved between exposures (Spearman ρ=0.48), with E5 moving from second shallowest under the composite to second steepest under OASIS. Effect magnitude therefore varies between retrievers, and the pooled estimate should be read alongside the per-model range rather than in place of it.

Scoring BM25 on windowed documents that match the dense scoring geometry did not significantly change its slope. Within the representation implementation, whose whole-document BM25 slope was −0.0145 under the composite against −0.0160 in the primary analysis, the windowed slope was −0.0184 (paired difference +0.0039, 95% CI −0.0066 to +0.0131; *P*=.40); under OASIS the implementation gave −0.0287 against −0.0295 in the primary analysis, and the windowed slope was −0.0289 (+0.0002, 95% CI −0.0088 to +0.0092; *P*=.95). None of 6 contrasts, across 2 exposures, 2 inverse document frequency conventions, and 2 window sizes, reached significance on either the slope or the relative decline, and whole-document BM25 under this implementation reproduced the published per-stratum values to within 0.007. The dense-to-lexical contrast is therefore not an artifact of BM25 scoring whole documents, while the dense retrievers score windows.

The 3 retrievers with differing documented query and document instructions were rescored as specified. The asymmetric configuration reproduced the published per-stratum values to within 6×10 , and every slope remained negative and nominally significant under every setting: BGE −0.0192 to −0.0160, E5 −0.0117 to −0.0102, and Nomic −0.0177 to −0.0124. Absolute level moved in both directions, with E5 improving under symmetric settings from 0.114 to 0.128 in the lowest stratum, consistent with a task closer to symmetric passage matching than to asymmetric question answering.

Discussion

Principal Findings

Across 8 independently trained dense retrievers, known-item retrieval of ICU discharge summaries deteriorated as an independently measured acute physiological derangement composite increased. It was observed in a common pooled index with document-length matching, and its direction persisted under every sensitivity analysis reported here. BM25 also deteriorated, so the phenomenon is not peculiar to embeddings, although dense retrieval lost a larger proportion of a much lower baseline. Under a laboratory-only organ-dysfunction stratification, the slopes were directionally similar but less consistently supported, and a matched comparison on identical records could not distinguish the 2 exposures, so which dimension of derangement the association tracks remains open.

XSL•FO
RenderX

The gradient was reproduced under a validated severity score, OASIS, at essentially the same magnitude and with all 8 retrievers Holm-significant under both exposures. Because OASIS shares 6 of its 10 input variables with the composite, including age and mechanical ventilation, that agreement is convergent validation under an externally defined exposure rather than a component-isolation analysis. The age-free reconstruction addresses the component question, and it preserved direction in all 8 retrievers while reducing magnitude and per-model significance. The evidence therefore supports a gradient associated with early physiological derangement as measured by the composite, in which age is an independently scored component that contributes materially to the magnitude.

The geometric analysis narrows where the ranking failure occurs. The target document becomes less similar to a passage drawn from itself as the derangement score increases, whereas the best competing document changes much less, which argues against simple crowding of high-derangement records as the dominant explanation. It does not identify why self-retrievability falls. The measured alternatives either varied little with exposure or only modestly attenuated the coefficient, leaving section of origin, clinical register, authorial heterogeneity, and higher-order properties of narrative synthesis as plausible pathways.

The exposure is built from measurements obtained early in the ICU stay for almost all patients, whereas the outcome is a property of the discharge narrative, so the finding is neither a circular comparison of text-defined subgroups nor a claim that biology acts directly on retrieval. A more appropriate chain is underlying physiological derangement→measurement→recordedclinicalstate→interpretation and action→documentation→information geometry. This study relates an early measured derangement signal to a later property of the discharge-summary representation, without identifying the intervening pathway.

Adjustment for diagnoses, medications, procedures, comorbidity, note length, documentation lag, and note index attenuated the mean coefficient by 21.6%, more than any single covariate. Because several of those variables may lie downstream of the exposure, the attenuation is not a causal decomposition, and most of the unadjusted magnitude remained after adjustment.

Several analyses converge on one methodological point. Apparent differences between retrievers, exposures, and document representations were consistently larger on a relative than on an absolute scale, and in each case traceable to a difference in baseline rather than in degradation. The dense-to-lexical contrast is the clearest instance: absolute slopes were statistically indistinguishable while proportional declines differed by a factor of 2.8, because BM25 operates from a baseline roughly three times higher. Absolute differences are therefore reported first throughout, with proportional declines as secondary.

A single corpus mean can conceal large, patient-associated differences in retrieval performance: dense-model Q1-to-Q4 declines ranged from about 29% to 44% here.

Two deployment patterns are in scope for these results, and one is not. Cohort search and similar-patient lookup retrieve across patients from a shared index, which is the geometry evaluated here, and the subgroup-stratified evaluation we recommend applies to them. Retrieval-augmented generation that searches within a single patient's record is not in scope: its candidate population is that patient's own documents, there is no cross-patient competition, and nothing in this design speaks to it.

Comparison With Prior Work

Prior EHR retrieval evaluation has varied models, query types, pooling strategies, retrieval settings, and documentation context [3-8], and a distinct question is whether model rankings transfer between institutions. Those axes concern which retriever to choose and whether that choice travels. The present study holds the retriever fixed and varies the patient, which, to our knowledge, has not previously been tested against a derangement measure derived independently of the text.

Biomedical specialization did not confer greater robustness in the 2 specialized retrievers evaluated here, consistent with prior clinical-retrieval benchmarking [3,6]. MedCPT is trained on PubMed search logs and BioLORD on ontology-derived concept definitions, and neither training distribution resembles narrative discharge documentation. Whether a retriever trained directly on EHR documentation behaves differently remains open [4]; we did not evaluate DR.EHR (dense retrieval for electronic health record) because it is trained on MIMIC-IV discharge summaries, which are also our evaluation corpus, so without documented nonoverlap, its performance could not be interpreted as an independent benchmark.

Several of the authors' related analyses use the same resource, so the overlap should be explicit. The multicorpus embedding benchmark cited above [8] draws on MIMIC-IV-Note and four of the present encoders, but its exposure is corpus type and query format rather than the patient, and its documents, queries, and cohort differ. A companion analysis of whether retriever rankings transfer between institutions shares the encoder panel and the known-item assay but varies the institution rather than physiological derangement and uses a different candidate index. The largest overlap is with the companion time-series analysis for which the cohort and the composite were built: it shares this cohort and the exposure itself, but its outcome is a physiological time series rather than retrieval, and it uses no document text, query, or embedding. No query, target document, embedding cache, or fitted model is shared with any of the three; what is shared is the underlying resource, the cohort and exposure with the time-series analysis, and the encoder panel and assay with the retrieval studies.

Limitations

Cohort retention differed modestly by exposure: the proportion of ICU stays contributing a linked single-ICU-stay discharge summary rose from 66.5% in the lowest to 73.9% in the highest derangement quartile. Excluding admissions with more than one ICU stay removes some complex trajectories by design, and the effect of these exclusions on the retrieval gradient is not established.

The exposure is a study-specific composite rather than a validated severity score. Its mortality discrimination is modest,

XSL•FO
RenderX

as is that of OASIS on the same records, and it is essentially orthogonal both to the two-component platelet-and-creatinine SOFA subscore (Spearman ρ=0.004) and to the three-component laboratory-only organ-dysfunction score (Spearman ρ=−0.010), which cover different organ axes. The primary analysis was repeated under a validated first-day severity score, and the gradient was reproduced, but age contributes to both scores, and the age-free reconstruction preserved direction in all 8 retrievers while reducing the mean decline to 29.7% and the number of Holm-significant models to three.

The composite's Glasgow Coma Scale term was taken as the lowest recorded value, without the convention of substituting a normal value when the verbal component cannot be assessed because of intubation. Ventilated stays therefore receive both a ventilation point and a depressed Glasgow Coma Scale score, correlating at Spearman ρ=0.53, so early ventilation enters the composite through 2 channels where the design intended one; 78.2% (2234/2857) of ventilated stays fell in the worst band. Ventilation prevalence rises across the strata, from 35.7% (702/1967) in the lowest to 68.4% (718/1049) in the highest, so the double count is not confined to one stratum. Its magnitude is nonetheless bounded: ventilated and nonventilated stays differed by 1.41 points on an 18-point scale, and ventilation status accounted for about 6% of the rank variance of the composite total (Spearman ρ=0.24). The OASIS replication applies the opposite convention and produced the same gradient.

The validated-score replication does not address the age component. OASIS assigns up to 9 of its points to age and 9 to mechanical ventilation, so agreement between OASIS and the composite is expected partly for that reason. The age-free reconstruction remains the relevant analysis for whether a gradient persists without age, and it reduced both magnitude and per-model significance.

The shape of the association is not resolved. Departure from linearity was rejected in only one of 8 retrievers under each exposure, but the lowest-stratum contrast recovered more of the range than a straight line implies, so some front-loading is likely, and the data are not powerful enough to characterize it retriever by retriever.

Known-item retrieval is a controlled assay rather than a deployed clinical search task. It measures self-retrievability of a known target and should not be read as a failure rate for retrieval-augmented generation or clinical decision support, nor does the gradient reported here need to propagate proportionally to factual accuracy in a generation system.

The sentence-eligibility rule excluded any sentence appearing in more than three documents and therefore favored distinctive narrative passages, so the gradient may not generalize to common, templated, or repetitive clinical text; the alternative-query analysis varied which eligible passage was used, not the eligibility rule itself. The query is also a verbatim passage rather than a clinician-authored question, so vocabulary overlap and query-document symmetry are higher than in prospective clinical retrieval.

Query-side and document-side contributions are not fully separable because both query passages and targets originate from the same derangement-associated notes. Alternative sampling, measured query features, and the query-instruction sensitivity reduce this concern, but section, register, and synthesis properties remain incompletely characterized.

The structured complexity and documentation covariates varied only modestly across strata: median distinct diagnoses 21 to 24, procedures 4 to 5, Charlson index 3 in all 4 strata (Spearman ρ=0.00), and documentation lag 22.2 to 24.3 hours. Their attenuation therefore describes these measures within this cohort's range rather than bounding the broader constructs.

The alternative-explanation analyses include exploratory, post hoc components. Their purpose is to narrow plausible explanations rather than to establish a mediator or causal pathway.

The dense-vs-lexical difference is scale-dependent. No difference in absolute RR slopes was detected; the stronger dense effect is proportional to its lower baseline. A representation sensitivity, in which BM25 was scored on windowed documents matching the dense geometry, did not change this conclusion.

The magnitude of the gradient is design-conditional, ranging from a 24.4% to a 37.2% mean decline across the 5 designs reported, although the direction is negative in all 8 models under every design. A laboratory-only organ-dysfunction stratification also produced negative slopes in all 8 models, but only one survived multiplicity correction, and a matched comparison on identical records could not distinguish it from the composite, so which dimension of derangement the association tracks is not resolved.

The source cohort comprised adult medical ICU stays of at least 72 hours whose earliest qualifying stay began in a medical ICU. Shorter stays, transfers in from another unit, and surgical, cardiac, and other ICU populations were not studied, so the findings do not represent all ICU admissions.

Component cut points and strata are defined relative to this cohort rather than by fixed clinical thresholds, so replication should distinguish applying the present cut points (Multimedia Appendix 1) from rederiving them locally. This is a single-site, single-genre study from one academic medical center in one deidentification regime; generalization to other institutions, note types, and deidentification pipelines is untested.

The panel comprises open or open-weight encoders hosted locally. Proprietary API embedding models were not evaluated, so the findings apply strictly to the tested models and may not extend to proprietary end points, which many enterprise deployments rely on. The 5 masked language model comparators were standardized to CLS pooling and function as a performance floor. The separability analysis identifies where score geometry changes, not why target self-similarity falls.

Conclusions

Retrievability of an ICU discharge summary depends in part on the patient it describes. Across 8 contrastively trained dense retrievers, records in the worst quartile of an acute physiological derangement composite were retrieved less well than those in the best, by 0.047 reciprocal-rank units. The association held

XSL•FO
RenderX

under an externally defined and validated severity-of-illness score as well as the study-specific composite and in lexical as well as dense retrieval, and most of its magnitude remained after joint adjustment for measured clinical and documentary characteristics. Corpus-level retrieval metrics are therefore insufficient on their own, and performance should be reported by patient subgroup.

Data Availability

MIMIC-IV (version 3.1) and MIMIC-IV-Note (version 2.2) are available to credentialed users through PhysioNet under a data use agreement [10,11]. The cohort construction, query extraction, retrieval scoring, and all sensitivity and covariate analyses are available in the public GitHub repository [36] and archived as a frozen Zenodo release [37]. Aggregate results are reported in the manuscript and its multimedia appendices. MIMIC-derived per-query outputs and identifiers are not redistributed; they can be regenerated by credentialed users from the released code. No clinical note text or other patient-level data are redistributed. The repository is version-controlled, and release 1.3.0, archived at the cited digital object identifier, is the code state used for the results reported here, including the scripts for the analyses added at revision.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Generative AI Use

The author declares that generative AI was used in the creation of this manuscript. Generative AI (Claude) was used for proofreading, editing, or language polishing; code generation or optimization; and reference formatting. The author also used Claude (Anthropic) for code development and review, figure and table generation, and manuscript structural editing, and Grammarly was used for spell check, grammar, and text clarity. The author verified all outputs of the tool(s) used in the preparation of this manuscript and takes full responsibility for the final content of the manuscript.

Authors' Contributions

YM was solely responsible for the conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, project administration, writing of the original draft, and reviewing and editing of the manuscript. The author read and approved the final paper.

Conflicts of Interest

The author (YM) is separately employed as Chief Medical Officer of AlgiPharma AS, a clinical-stage biopharmaceutical company; this employment is unrelated to the present work. AlgiPharma AS had no role in the design, conduct, analysis, interpretation, or reporting of this study or in the decision to submit it for publication. The author declares no other relationships or activities that could appear to have influenced the submitted work.

Multimedia Appendix 1

Source cohort and specification of the acute physiological derangement composite: cohort criteria, components, aggregation, plausibility bounds, cut points, missing-data handling, and stratum boundaries; the sentence-normalization and line-exclusion rules used for query eligibility; per-model instruction prefixes, pooling conventions, and pinned revisions; and the two reference-definition properties replicated in the Oxford Acute Severity of Illness Score computation.
[DOCX File , 24 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Query characteristics by derangement quartile in the 5-draw analysis subset, with 5-window eligibility among the primary-analysis documents.
[DOCX File , 15 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Alternative explanations and sensitivity results: candidate, assessment, result, and mean attenuation of the dense-model derangement coefficient.
[DOCX File , 17 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Retrieval separability across derangement quartiles: target and best-non-target maximum query-to-chunk cosine similarity, the number of nontargets scoring above the target, and the margin slope per quartile.

XSL•FO
RenderX

[DOCX File , 16 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Stability of the derangement slope across 5 random query draws.
[DOCX File , 15 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Restratification by a laboratory-only organ-dysfunction score: per-model slopes on score quartile and on score value.
[DOCX File , 16 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Matched comparison of the 2 exposures on identical records within a single candidate index, with paired bootstrap intervals for the slope difference.
[DOCX File , 15 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Age-free exposure sensitivity: pooled index, 2784 documents, with original values for comparison.
[DOCX File , 15 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Sensitivity analyses: unmatched sample, separate-index design, and all usable query-target pairs.
[DOCX File , 15 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Mean reciprocal rank at 10 by vital status within each stratum, under both exposures, with the death coefficient conditional on stratum.
[DOCX File , 18 KB-Multimedia Appendix 10]

References

1. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-615. [FREE Full text] [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
2. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health. 2025;4(6):e0000877. [FREE Full text] [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
3. Zhao Z, Yuan H, Liu J, Chen H, Ying H, Zhou S. CliniQ: a multi-faceted benchmark for electronic health record retrieval with semantic match assessment. arXiv. Preprint posted online on April 8, 2025. 2025. [doi: 10.48550/arXiv.2502.06252]
4. Zhao Z, Ying H, Yu S. Dense retrieval for electronic health record with knowledge injection and synthetic data. IEEE J Biomed Health Inform. 2026:1-10. [doi: 10.1109/JBHI.2026.3692546] [Medline: 42118632]
5. Ou J, Huang T, Zhao Y, Yu Z, Lu P, Shen Y. Experience retrieval-augmentation with electronic health records enables accurate discharge QA. Association for Computational Linguistics; 2026. Presented at: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2-7, 2026:23406-23430; San Diego, CA. [doi: 10.18653/v1/2026.acl-long.1073]
6. Myers S, Miller TA, Gao Y, Churpek MM, Mayampurath A, Dligach D, et al. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. J Am Med Inform Assoc. 2025;32(2):357-364. [doi: 10.1093/jamia/ocae308] [Medline: 39703187]
7. Arzideh K, Schäfer H, Idrissi-Yaghir A, Schmidt CS, Eryilmaz B, Bahn M, et al. Improving retrieval augmented generation for health care by fine-tuning clinical embedding models: development and evaluation study. J Med Internet Res. 2026;28:e82997. [FREE Full text] [doi: 10.2196/82997] [Medline: 41880603]
8. Mikkelsen Y. Clinical context variables collectively rival model choice in embedding-based retrieval: multi-corpus benchmark study. JMIR Med Inform. 2026;14:e94241. [FREE Full text] [doi: 10.2196/94241] [Medline: 42097608]
9. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Found Trends Inform Retr. 2009;4(1-2):1-174. [doi: 10.1561/1500000019]
10. Johnson A, Pollard T, Horng S, Celi L, Mark R. MIMIC-IV-note: deidentified free-text clinical notes (version 2.2). PhysioNet. 2023. [doi: 10.13026/1n74-ne17]

XSL•FO
RenderX

11. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S. MIMIC-IV (version 3.1). PhysioNet. 2024. [doi: 10.13026/kpb9-mt58]
12. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
13. Pollard T, Moody BE, Lehman LH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1(8):792-795. [doi: 10.1038/s44360-026-00096-z] [Medline: 42841059]
14. Johnson AEW, Kramer AA, Clifford GD. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Crit Care Med. 2013;41(7):1711-1718. [doi: 10.1097/CCM.0b013e31828a24fe] [Medline: 23660729]
15. MIT-LCP/mimic-code. GitHub. URL: https://github.com/MIT-LCP/mimic-code/blob/c7e07560dc847e32cbb0b2890213e8e7cbd8bc7e/mimic-iv/concepts/score/oasis.sql [accessed 2026-10-12]
16. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D. C-Pack: packed resources for general Chinese embeddings. arXiv. Preprint posted online on September 24, 2024. 2024. [doi: 10.48550/arXiv.2309.07597]
17. BAAI/bge-base-en-v1.5. Hugging Face. URL: https://huggingface.co/BAAI/bge-base-en-v1.5 [accessed 2026-08-21]
18. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. arXiv. Preprint posted online on August 7, 2023. 2023. [doi: 10.48550/arXiv.2308.03281]
19. Wang L, Yang N, Huang X, Yang L, Jiao B, Jiang D. Text embeddings by weakly-supervised contrastive pre-training. arXiv. Preprint posted February 22, 2024. 2024. [doi: 10.48550/arXiv.2212.03533]
20. Nussbaum Z, Morris J, Duderstadt B, Mulyar A. Nomic Embed: training a reproducible long context text embedder. arXiv. Preprint posted online on February 3, 2025. 2025. [doi: 10.48550/arXiv.2402.01613]
21. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 3-7, 2019:3982-3992; Hong Kong, China. [doi: 10.18653/v1/d19-1410]
22. sentence-transformers/all-mpnet-base-v2. Hugging Face. URL: https://huggingface.co/sentence-transformers/all-mpnet-base-v2 [accessed 2026-10-13]
23. sentence-transformers/all-MiniLM-L6-v2. Hugging Face. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 [accessed 2026-10-13]
24. Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics. 2023;39(11):btad651. [FREE Full text] [doi: 10.1093/bioinformatics/btad651] [Medline: 37930897]
25. Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. J Am Med Inform Assoc. 2024;31(9):1844-1855. [FREE Full text] [doi: 10.1093/jamia/ocae029] [Medline: 38412333]
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1; June 2-7, 2019:4171-4186; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
27. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-1240. [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
28. medicalai/ClinicalBERT. Hugging Face. URL: https://huggingface.co/medicalai/ClinicalBERT [accessed 2026-10-13]
29. Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nat Med. 2023;29(10):2633-2642. [FREE Full text] [doi: 10.1038/s41591-023-02552-9] [Medline: 37710000]
30. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2021;3(1):1-23. [doi: 10.1145/3458754]
31. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:3615-3620; Hong Kong, China. URL: https://aclanthology.org/D19-1371/ [doi: 10.18653/v1/D19-1371]
32. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. J Hum Resour. 2015;50(2):317-372. [doi: 10.3368/jhr.50.2.317]
33. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70. [FREE Full text]
34. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130-1139. [doi: 10.1097/01.mlr.0000182534.19832.83] [Medline: 16224307]
35. MIT-LCP/mimic-code. GitHub. URL: https://github.com/MIT-LCP/mimic-code/blob/25f6c47ca99bb3b904b704aa4f072850cec6681b/mimic-iv/concepts/comorbidity/charlson.sql [accessed 2026-08-21]

XSL•FO
RenderX

36. yngvemikkelsen/derangement-retrieval-mimic. GitHub. URL: https://github.com/yngvemikkelsen/derangement-retrieval-mimic [accessed 2026-10-13]
37. Mikkelsen Y. Analysis code for: clinical free-text retrieval across an acute physiological derangement composite in intensive care unit discharge summaries (version 1.3.0) [computer software]. Zenodo. 2026. [doi: 10.5281/zenodo.23333326]

Abbreviations

API: application programming interface
AUROC: area under the receiver operating characteristic curve
CLS: classification token
DR.EHR: dense retrieval for electronic health record
EHR: electronic health record
hit@10: hit at 10
ICD: International Classification of Diseases
ICD-9: International Statistical Classification of Diseases, Ninth Revision
ICD-9-CM: International Statistical Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICU: intensive care unit
MIMIC-IV: Medical Information Mart for Intensive Care IV
MIT-LCP: MIT Laboratory for Computational Physiology
OASIS: Oxford Acute Severity of Illness Score
RR: reciprocal rank
RR@10: reciprocal rank at 10
SOFA: Sequential Organ Failure Assessment
VIF: variance inflation factor

Edited by G Nicora; submitted 21.Sep.2026; peer-reviewed by V Ayyagari, N Khairnar, A Kury Rivas; comments to author 12.Oct.2026; revised version received 09.Nov.2026; accepted 11.Nov.2026; published 23.Nov.2026

Please cite as:
Mikkelsen Y
Clinical Free-Text Retrieval Across an Acute Physiological Derangement Composite in Intensive Care Unit Discharge Summaries: Stratified Evaluation Study
JMIR AI 2026;5:e112610
URL: https://ai.jmir.org/2026/1/e112610
doi: 10.2196/112610
PMID:

©Yngve Mikkelsen. Originally published in JMIR AI (https://ai.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX