

ORIGINAL RESEARCH

Open access

Multimodal Machine Learning Model for Predicting High-Cost Hospital Episodes Using Pharmacy Utilization, Procedure Sequences, Length-of-Stay Trends, Intensive Care Transfers, and Administrative Claims Data

Anna Kowalska^{1*}, Piotr Nowak¹, Tomasz Zielinski², Katarzyna Mazur¹

Abstract

A small fraction of hospital episodes accounts for a disproportionate share of inpatient spending. Early recognition of these episodes remains difficult when risk assessment depends mainly on static admission information. More adaptive prediction is needed to support clinical and financial planning during hospitalization. Existing cost prediction models often emphasize claims, diagnoses, or broad utilization histories while underusing the dynamic signals that emerge during the inpatient stay. Pharmacy utilization, procedure sequencing, length-of-stay progression, and intensive care transfers may reveal escalating resource intensity before the final cost is known. Failure to integrate these modalities limits early identification of high-cost episodes. This article proposes a multimodal deep learning framework for predicting whether a hospitalization could become a high-cost outlier. The model is designed to combine pharmacy utilization, procedure sequences, length-of-stay trends, intensive care transfer events, and administrative claims data. The intended use is dynamic risk estimation early and repeatedly during the episode. The conceptual model uses separate modality-specific encoders for static claims features, temporal procedure events, pharmacy utilization patterns, length-of-stay trajectories, and intensive care transfer indicators. These representations are fused into a shared episode-level embedding trained with a cost-sensitive objective. The framework is intended for evaluation in historical and silent prospective deployment settings without assuming immediate clinical intervention effects. Conceptually, the model would output an updated probability that an active hospitalization will exceed a high-cost threshold. This probability would change as new medication orders, procedures, length-of-stay milestones, and intensive care transfers occur. The output could support utilization review, case management, pharmacy stewardship, and financial counseling workflows. A multimodal deep learning model for high-cost hospital episode prediction could help health systems identify emerging cost outliers before discharge. By combining static claims information with dynamic inpatient trajectories, such a model could support earlier resource allocation and more coordinated care planning. Its value should be assessed through careful validation, calibration, workflow integration, and prospective impact evaluation.

Keywords High-cost hospital episodes, Multimodal deep learning, Pharmacy utilization, Administrative claims, Intensive care transfer, Length of stay

*Correspondence:

Anna Kowalska
anna.kowalska@gmail.com

¹ Department of Health Information Analytics, Faculty of Medicine, University of Warsaw, Warsaw, Poland

² Department of Intelligent Healthcare Engineering, Faculty of Engineering, Warsaw University of Technology, Warsaw, Poland

Introduction

High-cost hospital episodes create substantial operational and financial pressure because a relatively small group of patients can account for a large share of inpatient expenditure. Prior work on high-cost patients and high-need high-cost populations shows that prediction is clinically meaningful but challenging when risk is distributed across comorbidities, utilization history, and evolving care intensity [1-3]. Claims-based and population-level studies suggest that early identification could help health systems prioritize care management, utilization review, and financial planning for patients most likely to generate extreme costs [4-6]. A model that anticipates high-cost trajectories during the admission would therefore address both service delivery and resource stewardship needs.

Current predictive cost models remain piecemeal because many rely on admission diagnoses, prior utilization, or static administrative information while giving limited attention to events that unfold during hospitalization. Administrative claims and registry-based models have demonstrated value for risk stratification, yet they often describe baseline risk rather than the accumulating resource burden of an active episode [7-9]. Deep learning approaches using structured electronic health records show that richer longitudinal representations can improve clinical prediction conceptually, but cost forecasting requires direct modeling of resource-use signals rather than diagnosis history alone [10-12]. This gap is especially important for inpatient episodes whose costs may escalate after medication changes, invasive procedures, prolonged stay, or critical care transfer.

Hospitals increasingly collect real-time data streams that could inform dynamic cost prediction, including pharmacy dispenses, procedure logs, daily length-of-stay milestones, and intensive care transfer events. Prior studies on healthcare cost prediction, length-of-stay prediction, and inpatient trajectory modeling indicate that temporal patterns can carry predictive information beyond static covariates [13-16]. Pharmacy-related spending and pharmacotherapy changes can also contribute to expenditure growth, making medication intensity a potentially important signal for episode-level cost forecasting [17, 18]. Multimodal deep learning is well suited to this setting because separate encoders can learn from categorical claims, sequential procedures, pharmacy time series, and event flags before combining them into a unified risk representation [19-21].

The thesis of this article is that a multimodal deep learning model could ingest both static claims-derived risk factors and dynamic in-hospital trajectories to forecast the probability that an admission will become a high-cost episode. Such a model would not replace clinical judgment, but it could provide a continuously updated risk estimate as the admission evolves. Prior machine learning studies of health expenditure, high-cost patients, and medical insurance cost prediction support the feasibility of algorithmic cost risk stratification, while also highlighting the need for calibration, interpretability, and careful evaluation in local health system contexts. The proposed framework therefore emphasizes conceptual architecture, input design, temporal updating, and operational integration rather than reporting experimental performance or simulated results.

Background

High-cost episodes in hospital care

High-cost hospital episodes are commonly conceptualized as admissions exceeding a percentile-based threshold, such as the top decile or top twentieth of episode costs, or as admissions crossing an absolute institutional cost benchmark. These episodes often reflect a combination of baseline illness burden, complex procedures, long stays, intensive care use, and prior healthcare utilization rather than a single isolated factor [1, 3, 22]. Systematic reviews of high-need high-cost prediction models indicate that definitions vary across settings, but the central policy problem is consistent: identifying patients whose future resource use will be extreme enough to warrant early intervention [5]. In hospital finance, this definition is operational rather than purely statistical, because the high-cost label must align with utilization management capacity, reimbursement context, and service-line priorities [4, 6].

Pharmacy utilization as a cost driver and predictor

Pharmacy utilization can function as both a direct cost component and an indirect marker of clinical complexity. High-cost antimicrobials, chemotherapy, biologics, anticoagulants, parenteral nutrition, sedatives, and multiple concurrent drug classes may indicate severe illness, procedure-related complications, or prolonged care requirements. Studies linking pharmacotherapy to expenditure increases and evidence-weighted cost prediction suggest that medication patterns can add information beyond diagnosis and demographic covariates [17, 18, 23]. In a high-cost episode model, pharmacy data should therefore be represented not only as total drug spending but also as time-stamped drug class exposure, escalation patterns, and high-cost medication flags.

Procedure sequences and length-of-stay trajectories

Procedure sequences provide a temporal record of resource-intensive care, including operations, endoscopy, catheterization, imaging-guided interventions, and repeat procedures. Length-of-stay trajectories complement these events by capturing the accumulating duration of hospital resource use and the deviation from expected discharge timing. Research on hospital length-of-stay prediction and inpatient trajectory modeling supports the idea that temporal clinical patterns can provide dynamic signals of future resource consumption [15, 16, 24, 25]. A model that jointly encodes procedure timing and length-of-stay progression could therefore distinguish a predictable post-procedure recovery from a trajectory marked by repeated interventions and delayed discharge.

Intensive care transfers and administrative claims

Intensive care transfer is a major clinical and financial inflection point because it signals escalation in monitoring intensity, staffing requirements, medication complexity, and procedure risk. Administrative claims data, in contrast, provide a stable baseline profile through diagnoses, diagnosis-related groups, payer type, comorbidity burden, and prior utilization history. Prior high-cost patient and healthcare expenditure studies demonstrate that claims-derived variables are useful for baseline stratification, while dynamic hospital events are needed to refine risk during the admission [2, 7, 9, 26]. Combining ICU transfer indicators with claims-based risk vectors would allow the model to distinguish patients who enter the hospital with high expected cost from patients whose episode becomes costly because of in-hospital deterioration.

Multimodal learning in healthcare predictive analytics

Multimodal learning in healthcare aims to combine heterogeneous data streams such as structured codes, longitudinal events, time-series measurements, and static administrative variables. Deep learning models for medical records, diagnosis prediction, and structured electronic health records show how embeddings, recurrent networks, convolutional models, attention mechanisms, and transformers can represent complex clinical histories [10-12, 19-21]. Cost prediction studies further suggest that fine-grained temporal patterns and multivariate patient trajectories can be adapted to expenditure forecasting when resource-use outcomes are the target [13, 14, 27, 28]. For high-cost hospital episode prediction, this multimodal approach is attractive because no single data type fully captures baseline risk, treatment intensity, clinical escalation, and accumulating length of stay.

Model Development Overview

High-level predictive framework

At admission, the proposed framework would generate an initial high-cost probability from static claims-derived variables such as comorbidities, diagnosis group, payer type, prior utilization, and prior cost history. As the hospitalization progresses, the model would update this probability when pharmacy orders, procedure events, length-of-stay milestones, or intensive care transfers are recorded. This dynamic design follows the broader logic of longitudinal cost prediction and inpatient trajectory modeling, where risk is not fixed at admission but evolves as new evidence accumulates [13, 14, 16]. The model would therefore function as an episode surveillance tool rather than a one-time admission score.

Core input modalities

The core input modalities would include a static administrative claims vector and several dynamic event streams aligned by hospital day. Claims features would summarize diagnosis categories, diagnosis-related group information, payer class, comorbidity burden, prior encounters, and historical expenditure, while pharmacy features would track daily medication charges, therapeutic classes, and high-cost drug indicators [4, 9, 17]. Procedure features would represent coded interventions with timestamps, and length-of-stay features would encode daily accumulation and deviation from expected stay patterns [15, 25]. Intensive care transfer would be represented as a time-varying event indicator, allowing the model to register escalation in clinical severity and resource intensity.

Design principles

The model should be real-time updatable, robust to skewed cost distributions, sensitive to emerging cost drivers, and interpretable for utilization review staff. Because high-cost episodes are rare and cost outcomes are right-skewed, the framework should support cost-sensitive learning and calibration rather than optimize only average prediction error [9, 26, 29]. Because the model is intended for operational use, it should also provide explanations that link predictions to recognizable drivers such as high-cost medications, repeated procedures, prolonged stay, and ICU escalation [7, 30]. These design principles align the architecture with both machine learning validity and hospital workflow usefulness.

Data Sources and Feature Engineering

Claims-based baseline features

Claims-based baseline features would be extracted from inpatient and prior outpatient records to represent the patient's pre-admission risk profile. Relevant variables could include diagnosis categories, admission diagnosis-related group, payer type, age, prior hospitalizations, prior emergency visits, prior outpatient utilization, and prior year total costs. Studies using administrative claims and national insurance data show that these variables can support high-cost patient prediction, healthcare expenditure modeling, and persistent high-utilizer identification [2-4, 6, 7]. In the proposed model, claims features would serve as the static risk anchor against which dynamic in-hospital cost signals are interpreted.

Pharmacy and procedure sequence integration

Pharmacy transactions would be represented as daily, time-stamped events containing medication name, therapeutic class, dose category, route, charge or cost category, and a high-cost medication flag. Procedure events would be represented as coded temporal sequences that preserve order, timing, and clustering across the admission. Prior work on pharmacotherapy and healthcare cost prediction supports the inclusion of medication information, while temporal deep learning studies show that event sequences can be encoded into meaningful predictive representations [14, 17, 19, 21, 23]. Aligning pharmacy and procedure streams by hospital day would allow the model to detect whether cost risk is rising through medication escalation, procedural complexity, or the interaction between both.

Table 1 maps each input modality to its cost-risk mechanism, computational representation, temporal role, interpretability value, and implementation vulnerability.

Table 1. Analytical Mapping of Input Modalities to Cost-Risk Mechanisms, Model Representations, and Operational Interpretability

Input modality	Core cost-risk mechanism captured	Suggested model representation	Temporal role in prediction	Interpretable explanation for stakeholders	Main implementation risk

Administrative claims	Baseline illness burden, payer context, prior utilization, historical cost exposure	Static claims risk vector encoded through dense layers	Establishes admission-level prior risk	"Patient entered admission with elevated expected resource burden"	Delayed or incomplete claims availability
Pharmacy utilization	Medication intensity, high-cost drug exposure, treatment escalation, complication proxy	Drug class embeddings, daily medication-cost profiles, temporal attention	Updates risk as pharmacotherapy becomes more complex	"Risk increased after initiation of high-cost or intensive medication classes"	Medication charge variation and formulary differences
Procedure sequences	Procedural complexity, repeated interventions, clustered resource use	Transformer, LSTM, or temporal convolution over ordered procedure events	Captures escalation after interventions occur	"Risk reflects repeated or resource-intensive procedures"	Coding lag and inconsistent procedure timestamping
Length-of-stay trends	Accumulating resource use, delayed discharge, deviation from expected stay	Rolling LOS features and deviation-from-expected trajectory encoder	Converts ongoing hospitalization duration into evolving risk	"Risk increased because the stay is exceeding expected duration"	Social or placement delays may be poorly encoded
ICU transfer events	Clinical escalation, high staffing intensity, monitoring burden, critical care cost	Time-varying event flags and escalation timing features	Marks major inflection point in cost trajectory	"Risk increased after transfer to intensive care"	ICU thresholds vary across hospitals
Multimodal fusion	Interaction between baseline risk and dynamic inpatient escalation	Late, mixed, or attention-guided embedding fusion	Produces unified episode-level cost phenotype	"Prediction reflects combined claims, medication, procedure, LOS, and ICU signals"	Fusion may obscure modality-specific accountability

Length-of-stay and ICU transfer features

Length-of-stay features would include the rolling day count, the rate at which the stay exceeds expected duration, and markers of delayed discharge or extended post-procedure recovery. Intensive care transfer features would include a binary transfer indicator, elapsed time to transfer, and whether the transfer occurs before or after major procedures or high-cost medication exposure. Literature on length-of-stay prediction, unstructured inpatient stay modeling, and inpatient trajectories supports the conceptual value of temporal hospital course features in predicting downstream utilization [15,

16, 24, 25]. When combined with claims and pharmacy information, LOS and ICU features would help distinguish routine high-acuity admissions from episodes that are becoming unexpectedly resource intensive.

Multimodal Deep Learning Architecture

Input branches and modality-specific encoders

The proposed architecture would include a static claims branch, a temporal procedure-sequence branch, a pharmacy utilization branch, and a length-of-stay plus intensive care branch. The claims branch could use fully connected layers over administrative risk features, while the procedure branch could use an LSTM, temporal convolution, or transformer encoder to represent ordered intervention events [11, 20, 21]. The pharmacy branch could use one-dimensional convolution or temporal attention over daily medication cost profiles and drug class embeddings, drawing on the logic of fine-grained temporal cost prediction [13, 14]. The LOS and ICU branch would encode time-varying regression features and event flags so that prolonged stay and escalation to critical care influence the shared episode representation.

Multimodal fusion and joint representation

Multimodal fusion could occur through late fusion, mixed fusion, or attention-guided fusion of embeddings from each modality-specific branch. A shared dense representation would learn interactions between baseline claims risk, pharmacy intensity, procedural complexity, length-of-stay deviation, and intensive care escalation. Prior deep learning models for electronic health records and healthcare cost prediction suggest that learned representations can integrate heterogeneous structured and temporal inputs more flexibly than manually specified regression interactions [10, 12-14]. In this framework, the fused embedding would represent the evolving cost phenotype of the hospitalization rather than a static patient profile.

Figure 1 presents the proposed multimodal deep learning architecture for dynamically predicting high-cost hospital episodes from static claims features and evolving inpatient resource-use signals.

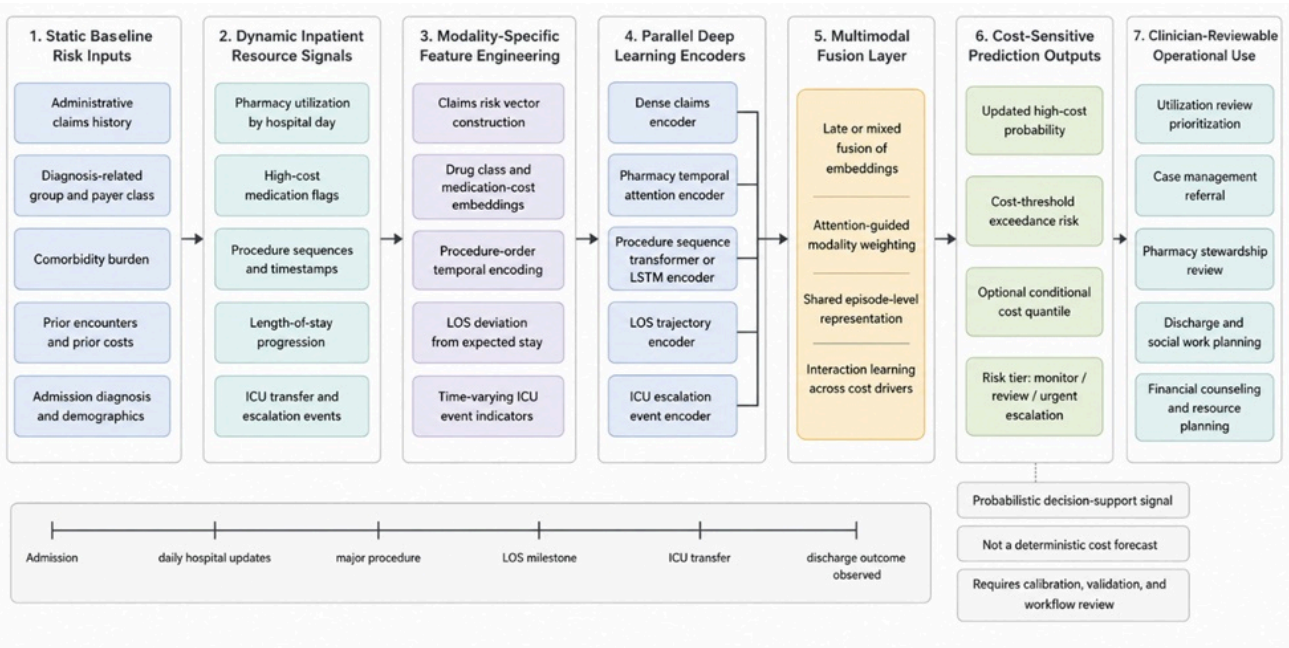


Figure 1. Multimodal Deep Learning Architecture for Dynamic Prediction of High-Cost Hospital Episodes

Output layer and cost-sensitive prediction

The primary output layer would estimate the probability that an active hospitalization will exceed a high-cost threshold, such as an institutionally defined top-cost category. An alternative output head could estimate a conditional cost quantile, allowing the model to characterize potential financial exposure without relying on a single point estimate. Studies of high-cost patient prediction, expenditure modeling, and skewed healthcare cost estimation indicate that models should address rare high-cost outcomes and heavy-tailed spending patterns through cost-sensitive objectives, calibration, and careful validation [9, 26, 28, 29]. The output should be interpreted as a decision-support signal for review and planning, not as a deterministic forecast of final charges.

Handling Imbalance, Skewness, and Temporal Dynamics

Cost distribution modelling and loss functions

Hospital episode costs are typically right-skewed, with a small number of admissions generating very high expenditure relative to the median episode. The proposed model should therefore support loss functions that reduce instability from extreme values while preserving sensitivity to clinically meaningful high-cost trajectories. Log-transformed cost targets, Huber-style losses, tweedie objectives, or quantile-oriented prediction heads could be considered conceptually for continuous cost estimation [9, 26, 29]. Neural in-hospital cost estimation and machine learning studies of expenditure forecasting indicate that the outcome scale and objective function are central design choices when modelling healthcare spending [30, 31].

Class imbalance in high-cost classification

When high-cost status is defined by an upper percentile or institutional threshold, positive cases will represent a minority of hospital episodes. A classification version of the proposed model should therefore account for imbalance through class-weighted loss, focal loss, threshold calibration, or latent-space augmentation strategies rather than treating all misclassifications as equally consequential. High-cost patient prediction studies using claims, registry, and insurance data show that rare-event identification requires attention to discrimination, calibration, and operational usefulness rather than simple overall accuracy [2, 5, 6, 8]. In this setting, the model should be optimized to support review of patients most likely to become costly while avoiding excessive false alerts that would overwhelm case management teams.

Temporal validity and real-time updating

Temporal validity is essential because high-cost prediction should simulate what would have been known at each point during the admission. Training and validation should therefore respect calendar time and episode chronology, avoiding leakage from post-discharge claims, final cost totals, or procedures that occur after the prediction time. The model could be invoked at admission, after early hospitalization milestones, after major procedures, and after intensive care transfer to produce updated risk assessments as new information arrives [13-15]. This design is consistent with longitudinal healthcare prediction frameworks in which patient status changes over time and static baseline models may become insufficient once dynamic clinical events accumulate [11, 12, 16].

Model Interpretability and Cost Prediction Rationale

Explaining high-cost predictions for case management

For clinical and operational adoption, the model should explain why an episode is being flagged as high cost. SHAP-style attribution, modality-level importance scores, or attention summaries could identify whether risk is being driven by high-cost medications, repeated procedures, length-of-stay deviation, prior utilization, or intensive care transfer. Prior work on explainable cost prediction and interpretable healthcare analytics suggests that prediction alone is insufficient when stakeholders must decide whether to initiate pharmacy stewardship, discharge planning, social work review, or financial

counseling [7, 29, 30]. Explanations should therefore translate learned representations into cost drivers that utilization review staff can recognize and act upon.

Transparency for financial and clinical stakeholders

The model output should be presented through a transparent dashboard that displays the predicted cost category, recent risk change, and main contributing modalities. Financial stakeholders may need to understand whether risk is driven by payer structure, prior utilization, pharmacy escalation, or expected prolonged stay, while clinicians may need to know whether the alert reflects deterioration, procedural complexity, or care coordination barriers. Studies of high-utilizer identification, diabetes registry prediction, and healthcare cost analytics emphasize that models become more useful when their outputs are aligned with practical care management decisions [7, 30, 32]. Transparency should therefore be treated as part of the model design rather than as an optional reporting layer.

Operational Integration and Risk Stratification

Embedding in utilization review and case management workflow

Operationally, the model could run nightly and on demand after major in-hospital events such as a new high-cost medication, invasive procedure, prolonged stay milestone, or intensive care transfer. Newly flagged patients would be routed to utilization review, pharmacy stewardship, case management, social work, or financial counseling depending on the dominant drivers of risk. Previous studies of high-cost patients, persistent high utilizers, and high-cost classification using institutional or claims data suggest that prediction is most valuable when linked to a defined intervention pathway [1, 3, 6, 7]. The model should therefore be integrated into existing hospital workflows rather than deployed as a standalone score without accountability for follow-up.

Tiered risk stratification and proactive interventions

A tiered risk approach would allow the model to distinguish moderate concern from urgent high-cost escalation. Lower tiers could trigger monitoring or discharge planning review, while higher tiers could prompt pharmacy review, palliative care consultation, care coordination escalation, or financial counseling when clinically appropriate. Machine learning studies of high-cost patients and condition-specific hospitalization cost prediction support the idea that resource-use risk is heterogeneous and should be interpreted in clinical context rather than as a single uniform category [22, 28, 33]. In the proposed framework, tiered stratification would help align the intensity of intervention with the likely drivers and timing of cost accumulation.

Evaluation Strategy

Predictive performance metrics

Evaluation should include discrimination, calibration, and decision-oriented measures appropriate for high-cost classification. Area under the receiver operating characteristic curve and precision-recall analysis could assess ranking performance, while calibration plots and threshold-specific measures could assess whether predicted probabilities are meaningful for operational decisions. For continuous or quantile cost prediction, evaluation could include error metrics and calibration around high-cost thresholds without presenting unsupported performance values [4, 9, 26]. Because high-cost alerts are intended to guide action, model assessment should also consider whether the risk categories are understandable and stable enough for utilization review workflows [29, 30].

Temporal and prospective validation

The model should be validated with a temporal split so that training precedes validation in calendar time and prediction windows mimic real operational use. A silent prospective deployment could then compare model flags generated during active admissions with later observed high-cost status, while withholding alerts from clinical teams until calibration, alert burden, and workflow fit are understood. Prior studies using nationwide claims, registry data, and longitudinal electronic health records show the importance of external or temporally separated validation when models are intended for real-world healthcare prediction [3, 5, 12, 32]. Regular recalibration would be necessary because prescribing patterns, procedure mix, payer contracts, and cost structures may shift over time.

Table 2 provides a deployment-readiness framework for evaluating whether the proposed model is temporally valid, calibrated, interpretable, and operationally actionable before active clinical use.

Table 2. Validation and Deployment Readiness Framework for Multimodal High-Cost Episode Prediction

Evaluation domain	Key question	Recommended assessment approach	Minimum evidence before active deployment	Risk if ignored
Temporal validity	Does the model use only information available at the prediction time?	Admission-day, daily-update, and event-triggered prediction windows with strict chronology	No leakage from discharge codes, final costs, or future procedures	Overestimated performance and unsafe operational trust
Discrimination	Can the model rank likely high-cost episodes above lower-risk admissions?	AUROC and precision-recall analysis across cost thresholds	Stable ranking performance across service lines and time periods	Poor prioritization of utilization review workload
Calibration	Are predicted probabilities meaningful for decision-making?	Calibration plots, calibration slope/intercept, threshold-specific calibration	Probabilities align with observed high-cost rates in local data	Misleading risk tiers and inappropriate alert thresholds
Imbalance handling	Does the model remain sensitive to rare high-cost episodes?	Class-weighted loss, focal loss, threshold tuning, precision-recall evaluation	Acceptable recall without excessive false alerts	Missed high-cost episodes or alert fatigue
Interpretability	Can staff understand why an episode was flagged?	Modality-level attribution, driver summaries, SHAP-style explanations	Explanations identify actionable drivers such as pharmacy, LOS, ICU, or procedures	Low trust and poor workflow adoption
Silent prospective validation	Does the model perform under real operational data latency?	Run model prospectively without showing alerts to clinical teams	Stable calibration, acceptable alert volume, and no major data-feed failures	Premature deployment before workflow feasibility is known
Workflow impact	Do alerts trigger useful review rather than passive notification?	Track referral timing, review completion, case management action,	Defined ownership and intervention pathway for each risk tier	Prediction becomes a standalone score without operational value

		pharmacy review, discharge planning response		
Local generalizability	Does performance hold across units, payers, and service lines?	Stratified validation by unit, diagnosis group, payer, ICU use, and procedure category	No major subgroup failure or uncalibrated service-line bias	Unequal or unreliable risk identification

Operational and financial impact assessment

After predictive validation, the model should be assessed for operational and financial impact through carefully designed implementation studies. Relevant outcomes could include whether flagged episodes receive earlier case management review, whether pharmacy or discharge planning interventions occur sooner, and whether the model supports more appropriate resource allocation. Studies of readmission charges, hospitalization cost prediction, and high-cost inpatient modelling show that cost prediction should ultimately be connected to budget impact, workflow change, and institutional decision-making rather than treated as a purely technical task [27, 28, 33]. Any impact assessment should avoid assuming savings from prediction alone and should instead evaluate whether prediction-guided interventions improve care coordination and financial stewardship.

Limitations

Data timeliness and completeness

A major limitation is that key data streams may not be available at the exact time when a prediction is needed. Pharmacy orders, medication administration records, procedure coding, and intensive care transfer documentation can lag behind clinical events, while claims data may be delayed until after billing processes are complete. Length-of-stay predictors may also be affected by unpredictable complications, social barriers, or discharge placement constraints that are not fully captured in structured data [15, 16, 24]. These limitations mean that the model should be designed for probabilistic updating rather than deterministic early certainty.

Generalizability across hospitals and cost structures

The proposed model would likely be sensitive to local practice patterns, coding conventions, chargemaster rates, payer contracts, pharmacy pricing, and service-line mix. A medication or procedure that signals high cost in one institution may have a different cost implication elsewhere, particularly across hospitals with different reimbursement arrangements or intensive care capacity. Prior healthcare cost prediction and high-cost patient studies highlight the need for recalibration, local validation, and careful interpretation when models are transferred across settings [4-6, 30]. Generalizability should therefore be evaluated as an empirical implementation question rather than assumed from model architecture alone.

Conclusion

A multimodal deep learning model for predicting high-cost hospital episodes could help health systems identify admissions whose resource use is likely to escalate before final charges are known. By combining claims-derived baseline risk with pharmacy utilization, procedure sequences, length-of-stay trends, and intensive care transfer indicators, the model would represent both pre-admission vulnerability and evolving inpatient complexity. Its purpose would be to support earlier planning rather than to replace clinical or financial judgment.

The central strength of the proposed framework is its fusion of static administrative information with dynamic in-hospital trajectories. Static claims features would establish baseline risk, while pharmacy, procedure, length-of-stay, and intensive

care signals would update the forecast as the episode unfolds. Interpretable modality-level explanations could make the prediction more useful for utilization review, pharmacy stewardship, case management, and financial counseling.

Important challenges remain before such a model could be used confidently in practice. Data latency, incomplete coding, skewed cost distributions, local cost structures, and changing payer contracts could all affect prediction quality. Prospective studies would be needed to determine whether prediction-guided interventions improve care coordination, reduce avoidable resource use, or support better financial planning.

Health systems operating under value-based payment or constrained inpatient capacity may be especially motivated to pilot this type of model. The most appropriate pilots would begin with silent validation, calibration monitoring, and workflow assessment before active alerts are introduced. If thoughtfully implemented, multimodal high-cost episode prediction could become a practical tool for focusing limited resources on patients whose hospital course is becoming clinically and financially complex.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jan 2024

Revised: 08 Mar 2024

Accepted: 06 Apr 2024

Published online: 20 July 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Wang Y, Iyengar V, Hu J, Kho D, Falconer E, et al. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. *Front Psychiatry*. 2017;8:114.

Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed Eng Online*. 2018;17(Suppl 1):131.

Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *NPJ Digit Med*. 2020;3(1):148.

Vimont A, Leleu H, Durand-Zaleski I. Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France. *Eur J Health Econ*. 2022;23(2):211-23.

de Ruijter UW, Kaplan ZR, Bramer WM, Eijkenaar F, Nieboer D, et al. Prediction models for future high-need high-cost healthcare use: a systematic review. *J Gen Intern Med*. 2022;37(7):1763-70.

Langenberger B, Schulte T, Groene O. The application of machine learning to predict high-cost patients: a performance-comparison of different models using healthcare claims data. *PLoS One*. 2023;18(1):e0279540.

Howson SN, McShea MJ, Ramachandran R, Burkom HS, Chang HY, et al. Improving the prediction of persistent high health care utilizers: retrospective analysis using ensemble methodology. *JMIR Med Inform*. 2022;10(3):e33212.

Choi Y, An J, Ryu S, Kim J. Development and evaluation of machine learning-based high-cost prediction model using health check-up data by the national health insurance service of Korea. *Int J Environ Res Public Health*. 2022;19(20):13672.

Wu Z, Berkowitz SA, Heagerty PJ, Benkeser D. A two-stage super learner for healthcare expenditures. *Health Serv Outcomes Res Methodol*. 2022;22(4):435-53.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1(1):18.

Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10(1):7155.

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. 2021;4(1):86.

Morid MA, Sheng OR, Kawamoto K, Ault T, Dorius J, et al. Healthcare cost prediction: leveraging fine-grain temporal patterns. *J Biomed Inform*. 2019;91:103113.

Morid MA, Sheng OR, Kawamoto K, Abdelrahman S. Learning hidden patterns from patient multivariate time series data using convolutional neural networks: a case study of healthcare cost prediction. *J Biomed Inform*. 2020;111:103565.

Lequertier V, Wang T, Fondrevelle J, Augusto V, Duclos A. Hospital length of stay prediction methods: a systematic review. *Med Care*. 2021;59(10):929-38.

Herrero-Zazo M, Fitzgerald T, Taylor V, Street H, Chaudhry AN, et al. Using machine learning to model older adult inpatient trajectories from electronic health records data. *iScience*. 2023;26(1):105987.

Jödicke AM, Zellweger U, Tomka IT, Neuer T, Curkovic I, et al. Prediction of health care expenditure increase: how does pharmacotherapy contribute? *BMC Health Serv Res*. 2019;19(1):953.

Panay B, Baloian N, Pino JA, Peñafiel S, Sanson H, et al. Feature selection for health care costs prediction using weighted evidential regression. *Sensors (Basel)*. 2020;20(16):4392.

Wickramasinghe N. A convolutional net for medical records. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Piscataway (NJ): IEEE; 2019. p. 4563-6.

Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2017. p. 787-95.

Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2017. p. 1903-11.

Luo L, Li J, Lian S, Zeng X, Sun L, Li C, et al. Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. *Health Informatics J.* 2020;26(3):1577-98.
<https://doi.org/10.1177/1460458219881335>.

Panay B, Baloian N, Pino JA, Peñafiel S, Sanson H, Bersano-Méndez NI. Predicting health care costs using evidence regression. *Proceedings.* 2019;31(1):74.

Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. *BMC Med Inform Decis Mak.* 2021;21(1):351.
<https://doi.org/10.1186/s12911-021-01722-4>.

Jain R, Singh M, Rao AR, Garg R. Predicting hospital length of stay using machine learning on a large open health dataset. *BMC Health Serv Res.* 2024;24(1):860.

Drewe-Boss P, Enders D, Walker J, Ohler U. Deep learning for prediction of population health costs. *BMC Med Inform Decis Mak.* 2022;22(1):32.

Gopukumar D, Ghoshal A, Zhao H. Predicting readmission charges billed by hospitals: machine learning approach. *JMIR Med Inform.* 2022;10(8):e37578.

Yang P, Qiu H, Wang L, Zhou L. Early prediction of high-cost inpatients with ischemic heart disease using network analytics and machine learning. *Expert Syst Appl.* 2022;210:118541.

Orji U, Ukwandu E. Machine learning for an explainable cost prediction of medical insurance. *Mach Learn Appl.* 2024;15:100516.

Rao AR, Jain R, Singh M, Garg R. Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data. *Healthc Anal.* 2024;6:100351.

Wu C, Wu F, Huang Y, Xie X. NICE: neural in-hospital cost estimation from medical records. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM; 2019. p. 2409-12.

Tan JK, Quan L, Salim NN, Tan JH, Goh SY, Thumboo J, et al. Machine learning-based prediction for high health care utilizers by using a multi-institutional diabetes registry: model training and evaluation. *JMIR AI.* 2024;3(1):e58463.

Fan S, Abulizi A, You Y, Huang C, Yimit Y, Li Q, et al. Predicting hospitalization costs for pulmonary tuberculosis patients based on machine learning. *BMC Infect Dis.* 2024;24(1):875.
<https://doi.org/10.1186/s12879-024-09771-6>.