ORIGINAL RESEARCH **Open access**

Variational Autoencoder for Unsupervised Detection of Adverse Drug Reaction Signals from Electronic Health Record Clinical Notes and Laboratory Abnormalities

Ahmed Benali[1*], Karim Boudiaf[2], Samir Touati[1]

Abstract

Adverse drug reactions (ADRs) are a major global health issue, contributing to significant morbidity, mortality, and healthcare costs. Many ADRs are detected only after widespread drug use, reflecting limitations of pre-market trials in capturing real-world patient variability. Although electronic health records (EHRs) collected between 2017 and 2023 provide rich data for post-market surveillance, they remain underused for systematic ADR detection. Current pharmacovigilance methods rely heavily on spontaneous reporting systems, which suffer from underreporting and bias, while supervised machine learning approaches require labeled ADR data that are often unavailable for rare or novel events. This paper proposes a variational autoencoder (VAE)-based unsupervised framework to detect ADR signals from multimodal EHR data, including clinical notes and laboratory results. The model learns normal patient data distributions and identifies deviations as potential safety signals without requiring labeled ADR examples. A multimodal architecture combines natural language processing of clinical notes with structured laboratory encoders, forming a shared latent space for anomaly detection based on reconstruction error. The framework enables detection of unknown ADRs by flagging abnormal patterns in patient records across large datasets from 2017 to 2023. Its unsupervised nature makes it suitable for identifying rare or previously unrecognized drug safety issues. Overall, this approach offers a scalable, proactive pharmacovigilance strategy that shifts drug safety monitoring from reactive reporting to predictive detection using routine EHR data.

Keywords Electronic health records, Variational autoencoder, Unsupervised anomaly detection, Adverse drug reactions, Clinical notes, Laboratory abnormalities

*Correspondence:
Ahmed Benali
ahmed.benali@gmail.com
[1] Department of AI Healthcare Systems, University of Algiers, Algiers, Algeria
[2] Department of Healthcare Analytics and Informatics, University of Tunis El Manar, Tunis, Tunisia

Introduction

Adverse drug reactions (ADRs) rank among the fourth to sixth leading causes of death in the United States, contributing to substantial patient harm across hospitalized and ambulatory populations. They account for 5-10% of hospital admissions and impose billions of dollars in annual healthcare costs through extended stays, additional treatments, and lost productivity [1, 2]. The burden extends beyond direct clinical impacts to include prolonged hospital stays and increased resource utilization in modern healthcare systems. Timely detection remains a critical challenge given the volume and complexity of real-world patient data.

Current pharmacovigilance relies on spontaneous reporting systems such as FAERS and EudraVigilance, which exhibit reporting rates below 10% due to underreporting and various selection biases. Many ADRs are discovered years after market approval, often only after widespread exposure in diverse patient groups [3, 4]. This delayed recognition underscores the limitations of passive surveillance mechanisms that depend on voluntary clinician or patient submissions. Consequently, proactive strategies are essential to accelerate signal identification from routinely collected data sources.

Supervised machine learning methods for ADR detection require large volumes of labeled ADR examples that are unavailable for unknown or emerging events. Unsupervised methods can detect anomalies without labels by modeling normal data distributions and flagging deviations, offering a viable alternative for rare event identification [5, 6]. These approaches leverage the inherent structure of EHR data without presupposing annotated outcomes. Such flexibility is particularly valuable in pharmacovigilance where ground-truth labels for novel ADRs are inherently scarce.

The thesis of this paper is a conceptual variational autoencoder (VAE) framework for unsupervised ADR signal detection from clinical notes capturing unstructured symptoms and laboratory abnormalities providing structured quantitative signals. The framework processes multimodal EHR data collected between 2017 and 2023 to learn normal patient patterns without any labeled examples. It generates reconstruction-based anomaly scores that highlight potential drug-related deviations for further review. This roadmap outlines the background, architecture, encoding strategies, detection mechanisms, and prioritization workflow for implementation in real-world healthcare systems.

Figure 1 presents the full hierarchical conceptual architecture through which multimodal EHR inputs, modality-specific encoding, shared variational latent modeling, reconstruction-based anomaly scoring, latent-space interpretation, and pharmacovigilance signal prioritization are integrated for unsupervised adverse drug reaction detection.

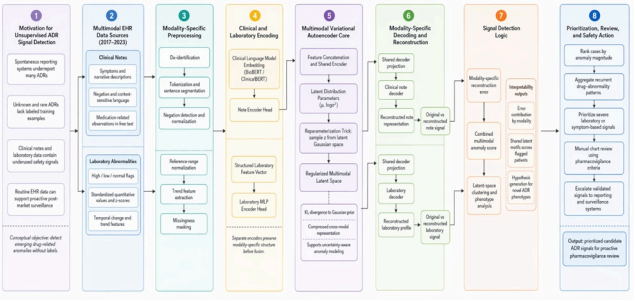

Figure 1. Hierarchical Conceptual Architecture of a Variational Autoencoder Framework for Unsupervised Adverse Drug Reaction Signal Detection from EHR Clinical Notes and Laboratory Abnormalities

Background

Adverse drug reactions and pharmacovigilance

Adverse drug reactions are classified into Type A reactions, which are predictable and dose-dependent extensions of a drug's pharmacological effects, and Type B reactions, which are idiosyncratic and unpredictable based on individual patient factors. Serious ADRs result in death, hospitalization, or permanent disability, while non-serious events may involve milder symptoms that still warrant monitoring in post-market surveillance [1, 7]. Regulatory requirements from agencies such as the FDA and EMA mandate ongoing pharmacovigilance activities to ensure drug safety throughout the product lifecycle. These classifications guide the design of detection systems that must accommodate both predictable and unexpected manifestations in EHR data.

Pharmacovigilance encompasses systematic collection, analysis, and response to drug safety information derived from multiple sources, including clinical practice and regulatory databases. Effective programs integrate data from diverse formats to identify signals that may not emerge during controlled trials [2, 3]. The process supports risk-benefit assessments and potential label updates or market withdrawals when new patterns arise. In the context of EHR systems, pharmacovigilance benefits from the continuous influx of real-world evidence that captures everyday clinical variability.

Limitations of current detection

Spontaneous reporting systems remain the cornerstone of pharmacovigilance yet suffer from pervasive underreporting that can exceed 90% for many ADRs and introduce selection biases toward well-known events. Medical literature reviews also exhibit publication bias, favoring positive or dramatic findings while underrepresenting routine safety signals from routine care [1, 8]. Electronic health records contain detailed clinical notes and laboratory results that are underutilized due to the complexity of unstructured text and the volume of data. These limitations collectively delay the identification of emerging ADR signals in large populations.

The reliance on supervised approaches further constrains detection because labeled datasets are only available for previously recognized ADRs, leaving novel events undetected until manual reporting occurs. Spontaneous systems and literature sources fail to capture the full spectrum of laboratory abnormalities or nuanced symptom descriptions embedded in clinical narratives [4, 7]. As a result, many drug safety signals surface only after years of market exposure and widespread patient exposure. Unsupervised methods address these gaps by modeling baseline distributions directly from unlabeled EHR data without presupposing known outcomes.

Table 1 situates the proposed multimodal variational autoencoder framework within the broader landscape of pharmacovigilance detection paradigms and clarifies why unsupervised multimodal anomaly modeling is conceptually better aligned with the discovery of unknown adverse drug reactions than passive, rule-based, or supervised alternatives.

Table 1. Conceptual Comparison of Pharmacovigilance Detection Paradigms for Identifying Adverse Drug Reaction Signals

Detection paradigm	Primary data source	Label requirement	Ability t detect unknow ADRs
Spontaneous reporting systems	Voluntary clinician, patient, and manufacturer reports	No formal training labels required	Moderat but delaye
Literature-based pharmacovigilance	Published case reports, case series, reviews	No formal training labels required	Low to moderat
Rule-based EHR surveillance	Structured EHR triggers and predefined thresholds	No training labels required, but explicit rules required	Low
Supervised machine learning for ADR detection	Labeled EHR, claims, or curated safety datasets	Yes	Low
Single-modality unsupervised anomaly detection	Either text alone or laboratory data alone	No	Moderat
Proposed multimodal VAE framework	Clinical notes plus laboratory abnormalities from EHRs	No	High

Unsupervised anomaly detection in healthcare

Unsupervised anomaly detection techniques such as standard autoencoders, one-class support vector machines, and isolation forests have demonstrated utility in healthcare by identifying outliers without requiring labeled training data. These methods learn representations of normal physiological or clinical patterns and quantify deviations through reconstruction error or distance metrics [9, 10]. In EHR contexts, they enable screening for rare events such as unexpected laboratory shifts or symptom clusters that may indicate adverse reactions. Their application supports scalable analysis across high-dimensional multimodal datasets.

Reconstruction error serves as a reliable anomaly score because data points that deviate from the learned normal distribution produce higher errors during decoding. This principle has been successfully applied to imaging and structured clinical data streams to flag potential abnormalities [11, 12]. Isolation forests complement these approaches by efficiently partitioning data to isolate anomalies in lower-dimensional subspaces. Together, these unsupervised tools form the foundation for generative frameworks that extend to complex EHR environments involving both text and numerical measurements.

Variational autoencoders

Variational autoencoders extend standard autoencoders by approximating the posterior distribution of latent variables through an encoder network and reconstructing data via a decoder, while regularizing the latent space with Kullback-Leibler divergence to a Gaussian prior. This probabilistic formulation enables generative sampling and uncertainty estimation that are particularly useful for anomaly scoring in healthcare applications [9, 11]. The reparameterization trick allows end-to-end differentiable training despite the stochastic sampling step. Such architectures have shown promise in modeling complex distributions present in clinical time series and imaging data.

Reconstruction probability, rather than simple mean squared error, provides a more robust anomaly metric in VAEs because it accounts for the full posterior distribution in the latent space. This approach has been applied to detect deviations in structured and unstructured healthcare records where traditional error measures may overlook subtle multimodal inconsistencies [8, 13]. β-VAE variants further encourage disentangled representations that improve interpretability of learned factors. These properties position VAEs as an ideal backbone for unsupervised ADR signal detection within multimodal EHR pipelines.

Framework Overview

High-level architecture

The proposed framework ingests paired clinical notes and laboratory abnormalities from EHR records and routes them through modality-specific encoders before concatenation into a shared multimodal VAE encoder. The encoder produces mean and log-variance parameters for the latent Gaussian distribution, from which samples are drawn via the reparameterization trick and passed to separate decoder heads for note and laboratory reconstruction [9, 10]. Reconstruction error is computed across both modalities to generate a combined anomaly score that flags potential ADR signals. This architecture ensures seamless fusion of unstructured text and structured numerical data within a single generative model.

The overall pipeline operates in a fully unsupervised manner on large cohorts of routine patient records collected between 2017 and 2023, producing anomaly rankings that prioritize cases for pharmacovigilance review. Latent representations learned in the bottleneck layer capture joint clinical-laboratory patterns that deviate from normal distributions [11, 12]. Output anomaly scores are thresholded to surface candidate signals without any requirement for labeled training examples. This high-level design supports scalable deployment across institutional or population-level EHR databases.

Core assumptions

The framework assumes the availability of a large EHR dataset comprising patients without documented known ADRs, allowing the VAE to learn a robust representation of normal clinical and laboratory patterns over the 2017–2023 period. Structured laboratory results and unstructured clinical notes are presumed to be routinely captured in sufficient volume to support stable training of the multimodal model [1, 4]. Missing data are handled through explicit masking mechanisms rather than imputation to preserve the integrity of the learned distribution. These assumptions align with the scale of contemporary EHR repositories that generate terabytes of multimodal data annually.

The core premise is that ADRs manifest as deviations from the normal data manifold captured by the VAE, enabling reconstruction-based detection without explicit labeling. This relies on the generative capacity of VAEs to model the underlying probability density of healthy patient trajectories [9, 13]. Temporal consistency within the 2017–2023 window further supports the assumption that training data represent stable baseline distributions prior to signal emergence. Under these conditions, the framework can generalize to unseen patient records while maintaining low false-positive rates for routine variations.

Design principles

The framework adheres to fully unsupervised learning principles, eliminating any dependence on labeled ADR examples and thereby addressing the scarcity of ground-truth data for novel events. Multimodal integration of clinical notes and laboratory abnormalities is prioritized to capture complementary signals that single-modality approaches might miss [2, 10]. Interpretability is achieved through analysis of the latent space and reconstruction contributions from each modality, facilitating clinical review of flagged cases. These principles ensure the system remains practical for real-world pharmacovigilance workflows.

Scalability and generalizability across institutions guide the design, with emphasis on standard preprocessing steps that accommodate variations in EHR vendor formats and documentation styles. The generative nature of the VAE supports synthetic data augmentation during training if needed, while reconstruction error provides an inherently explainable anomaly metric [11, 12]. No supervised fine-tuning is required post-training, preserving the unsupervised character of the entire pipeline. This design philosophy positions the framework as a foundational tool for proactive, label-free drug safety surveillance.

Table 2 operationalizes the proposed framework by mapping each architectural stage to its detection function, interpretive contribution, and downstream pharmacovigilance role, thereby clarifying how the conceptual model translates from multimodal representation learning into actionable adverse drug reaction surveillance.

Table 2. Functional Mapping of the Proposed Multimodal VAE Pipeline to Detection, Interpretation, and Pharmacovigilance Action

Framework stage	Main inputs	Principal operation	Imme
Clinical note preprocessing	Raw narrative EHR notes	De-identification, tokenization, normalization, negation-aware cleaning	Sta cli se
Laboratory abnormality preprocessing	Raw laboratory values and timestamps	Reference-range normalization, abnormality flagging, trend extraction, missingness masking	St labora
Note embedding and note encoder	Standardized note sequences	Domain-specific embedding and note-specific encoding	Fixed- repr
Laboratory encoder	Structured laboratory feature matrix	MLP-based nonlinear transformation with masking awareness	Fixed- la repr
Shared multimodal VAE encoder and latent space	Encoded note and laboratory representations	Fusion, posterior parameterization, stochastic sampling, KL-regularized latent modeling	Joint repr

Dual decoders and reconstruction	Latent samples	Modality-specific reconstruction of text and laboratory signals	Rec n labora
Composite anomaly scoring	Original and reconstructed multimodal records	Combination of normalized modality-specific reconstruction losses into a unified score	Rank valu pati
Latent-space analysis and clustering	Latent vectors of flagged cases	Clustering, subgroup discovery, and cross-case pattern analysis	Canc phen recurr
Signal prioritization and expert review	Ranked flagged cases plus drug exposure context	Error-based ranking, recurrence aggregation, severity weighting, chart review	Prior sig pharm

Vae Architecture

Encoder network

The encoder network employs a multi-head design with separate branches for clinical note embeddings and laboratory features before concatenation into a unified MLP that outputs parameters for the latent Gaussian distribution. Each modality is processed independently to preserve its intrinsic characteristics prior to joint fusion, allowing the model to weigh contributions dynamically during training [9, 10]. The resulting mean and log-variance vectors parameterize the approximate posterior from which latent samples are drawn. This architecture ensures that both textual and numerical inputs inform the latent representation equally.

Modality-specific preprocessing feeds directly into the encoder heads, with the note branch leveraging transformer-derived embeddings and the laboratory branch utilizing fully connected layers on normalized features. The concatenated vector undergoes batch normalization and dropout to enhance training stability across the 2017–2023 EHR cohorts [8, 11]. The encoder thus produces a compact yet expressive latent encoding that captures cross-modal dependencies relevant to ADR patterns. This design choice supports end-to-end differentiability while respecting the heterogeneous nature of EHR data.

Latent space

The latent space is configured with a dimensionality typically between 32 and 128 dimensions, providing sufficient capacity to encode complex multimodal patterns without excessive parameterization that could lead to overfitting on normal data. A standard Gaussian prior is imposed via the KL divergence term to regularize the posterior approximation and enable meaningful generative sampling [9, 13]. The reparameterization trick ensures that gradients flow through the stochastic sampling step during backpropagation. This probabilistic latent representation allows the model to quantify uncertainty in reconstructions for anomaly scoring.

Dimensionality selection balances expressiveness with computational efficiency, with lower dimensions favoring disentanglement and higher dimensions accommodating finer-grained clinical variations observed in laboratory time series. The Gaussian assumption facilitates closed-form computation of the KL term and supports downstream clustering of latent vectors for phenotype discovery [11, 12]. During inference, the latent space serves as a compressed manifold against which new patient records are projected to compute reconstruction probabilities. Such properties make the latent space a powerful tool for unsupervised ADR signal exploration.

Decoder network

The decoder network receives sampled latent vectors and routes them through a shared MLP before branching into

separate heads for clinical note reconstruction and laboratory value reconstruction. Each head is tailored to the output modality, with the note decoder generating token probabilities or embeddings and the laboratory decoder producing normalized continuous values or abnormality flags [9, 10]. Skip connections or residual blocks may be incorporated to improve gradient flow and reconstruction fidelity across modalities. This symmetric design mirrors the encoder to ensure consistent information flow.

The dual-head structure allows independent optimization of reconstruction objectives for text and structured data while sharing the latent representation, promoting joint learning of cross-modal relationships. During training, the decoder learns to regenerate normal patterns observed in the 2017–2023 EHR data, making deviations in new records indicative of potential anomalies [8, 11]. Output layers are chosen to match input distributions, such as softmax for tokenized notes or linear activations for laboratory measurements. The resulting reconstructions form the basis for computing modality-specific and combined error scores.

Loss function

The loss function is defined as the evidence lower bound (ELBO), which combines reconstruction losses for clinical notes (typically cross-entropy or binary cross-entropy on token predictions) and laboratory abnormalities (mean squared error on normalized values) with a weighted KL divergence term. The β coefficient in β-VAE variants controls the trade-off between reconstruction fidelity and latent regularization to encourage disentangled representations suitable for anomaly interpretation [9, 11]. This formulation maximizes the marginal likelihood of the observed data while preventing posterior collapse. Optimization proceeds via stochastic gradient descent on minibatches drawn from the unlabeled EHR corpus.

Separate weighting of note and laboratory reconstruction terms allows emphasis on the modality most relevant to specific ADR signals, such as laboratory trends for certain drug classes. The negative ELBO is minimized during training on normal patient records, resulting in a model that assigns low reconstruction probability to anomalous inputs [8, 13]. Hyperparameter tuning of β focuses on balancing generative capability with discriminative power for outlier detection. This loss structure underpins the unsupervised nature of the entire framework by relying solely on reconstruction quality as a proxy for normality.

Clinical Note Encoding

Text preprocessing

Clinical note preprocessing begins with de-identification to remove protected health information through rule-based and machine-learning PHI detectors, ensuring compliance with privacy standards while preserving clinical content. Tokenization follows using domain-specific tools that handle medical abbreviations, acronyms, and sentence boundaries commonly found in EHR narratives [2, 5]. Sentence segmentation and negation detection are applied to accurately capture symptom descriptions and their contextual qualifiers. These steps produce clean, standardized input sequences suitable for downstream embedding models.

Additional normalization addresses spelling variations, numeric expressions, and temporal references within notes to reduce vocabulary sparsity across the 2017–2023 dataset. Stop-word removal is applied selectively to retain clinically meaningful terms while discarding boilerplate phrases common in templated documentation [3, 4]. The resulting preprocessed text serves as input to embedding layers without loss of semantic richness. This rigorous preprocessing pipeline ensures consistency and quality for the subsequent note encoder component of the VAE framework.

Note embedding

Note embeddings are generated using pre-trained clinical language models such as BioBERT or ClinicalBERT, either fine-tuned on the target EHR corpus or used in frozen form to extract contextual representations. Mean pooling or extraction of the CLS token produces fixed-length vectors that summarize the entire note while preserving semantic nuances related to symptoms and clinical context [2, 6]. Dimensionality reduction via principal component analysis or a lightweight projection layer aligns the embedding size with laboratory features prior to multimodal fusion. This process transforms variable-length clinical text into compact vectors suitable for VAE encoding.

The embedding strategy emphasizes domain adaptation to capture medical terminology and implicit relationships that general-purpose models might miss, enhancing sensitivity to subtle ADR indicators. During VAE training, these embeddings are fed directly into the note-specific encoder head, contributing to the joint latent space that models normal documentation patterns [10, 14]. Gradient flow

through the embedding layer is optionally enabled to allow task-specific adaptation without full retraining. The resulting note representations integrate seamlessly with laboratory features to support comprehensive anomaly detection in the overall framework.

Laboratory Abnormality Encoding

Lab data representation

Laboratory data are encoded by classifying values as high, low, or normal relative to demographic-specific ranges and transforming them into z-scores or categorical indicators. Temporal features (e.g., change from prior values, short-term trends) capture evolving abnormalities that single measurements may miss [15, 16]. Missing values are explicitly marked rather than imputed to avoid bias during VAE training.

This representation preserves quantitative detail while enabling multimodal fusion with clinical text. It captures both current abnormalities and longitudinal patterns (e.g., gradual creatinine rise), supporting detection of subtle ADR-related signals. Cohort-wide normalization (2017–2023) ensures consistency across sources [7, 8]. The resulting feature vector provides a compact, objective complement to unstructured clinical notes for unsupervised anomaly detection.

Lab encoder

The lab encoder is a lightweight multilayer perceptron that transforms structured features into a fixed-dimensional embedding aligned with clinical notes. It uses ReLU activations, dropout, and batch normalization for stable training across heterogeneous EHR data [10, 17]. Missingness is handled באמצעות indicator masking, allowing the model to down-weight absent values without bias.

The encoder output is concatenated with note embeddings before the shared VAE bottleneck, enabling joint learning of textual and numerical patterns. Its shallow design limits overfitting while capturing nonlinear relationships among analytes [18, 19]. Fully differentiable training supports end-to-end optimization, ensuring balanced integration of modalities without requiring labeled ADR data.

Unsupervised ADR Signal Detection

Reconstruction error scoring

Reconstruction error scoring combines normalized note cross-entropy and laboratory mean squared error into a single anomaly metric, weighted according to the ELBO. Patients exceeding a dynamic threshold (e.g., 99th percentile of training data) are flagged as potential ADR signals without labeled data [9, 11]. This assumes most 2017–2023 EHR records reflect normal patterns, enabling scalable outlier detection.

Using VAE generative properties, atypical inputs receive higher reconstruction error, highlighting drug-related anomalies in text or labs [12, 13]. Modality-specific contributions can be examined post hoc for clinical interpretation. As a fully unsupervised approach, it avoids bias from labeled training data [8, 14]. The output is a ranked list of anomalous patient records for signal prioritization.

Latent space analysis

Latent space analysis clusters encoded representations (e.g., via GMM or DBSCAN) of flagged patients to identify shared patterns indicative of potential ADR phenotypes. Attribution methods link latent features back to clinical notes or lab deviations, improving interpretability [11, 20].

New patient records can be projected into this space to detect emerging subgroups and compare against historical patterns [21, 22]. This supports discovery of novel drug–ADR associations through recurring latent structures across patients. As clustering is unsupervised, patterns reflect intrinsic data structure rather than predefined labels [23-25]. Latent analysis complements reconstruction scoring by providing an additional perspective on safety signals.

Signal Prioritization

Ranking signals

Ranking of ADR signals occurs by sorting flagged patient records first according to the magnitude of their combined reconstruction error, which quantifies overall deviation from normal patterns learned by the VAE. Secondary sorting incorporates the frequency with which the same drug-ADR pair appears across multiple high-error cases within the

2017–2023 window, elevating signals that demonstrate recurrence rather than isolated outliers [1, 26]. Clinical severity is further integrated by weighting laboratory abnormality contributions, prioritizing cases with life-threatening deviations such as severe neutropenia or hepatotoxicity [7, 16]. This multi-criteria ranking produces a prioritized list that balances statistical anomaly strength with practical pharmacovigilance relevance.

The prioritization algorithm operates entirely post-training and requires no additional labeled data, preserving the unsupervised character of the entire pipeline. Drug exposure information extracted via rule-based or lightweight NLP modules from clinical notes is used solely for grouping, not for model training [2, 14]. By focusing on error magnitude and pattern frequency, the framework naturally surfaces both rare idiosyncratic reactions and more widespread dose-dependent effects. The ranked output is designed for direct integration into existing pharmacovigilance dashboards without manual re-labeling.

Clinical review workflow

The clinical review workflow assigns top-ranked signals to pharmacovigilance teams for manual chart review, where clinicians examine the original clinical notes and laboratory trends to confirm or refute potential drug causality using established algorithms such as Naranjo or WHO-UMC criteria. Validated signals are then prepared for submission to regulatory agencies through standardized reporting formats while retaining the original reconstruction error scores as quantitative support [3, 27]. Feedback from review outcomes can optionally inform future model iterations through lightweight fine-tuning of hyperparameters without violating the unsupervised training paradigm. The workflow is deliberately lightweight to minimize operational burden.

Integration with existing spontaneous reporting infrastructure allows flagged cases to be cross-referenced against FAERS or EudraVigilance entries, highlighting signals that might otherwise remain unreported [4, 17]. Automated alerts are generated for cases exceeding predefined severity thresholds derived from laboratory flags, ensuring timely escalation. Documentation of review decisions is logged to support audit trails and regulatory compliance. This human-in-the-loop process transforms the VAE-generated anomaly list into actionable pharmacovigilance intelligence while preserving the core unsupervised detection capability.

Evaluation Strategy

Detection metrics

Evaluation uses retrospective ADR labels from external sources only at test time, preserving the unsupervised framework. Metrics include recall and precision for identifying confirmed signals within top-ranked cases, and AUROC based on reconstruction error as a continuous anomaly score [5, 6, 9, 11]. Results are reported separately for note-based, lab-based, and combined signals to assess multimodal contributions.

Because no labels are used in training, these metrics provide an unbiased estimate of real-world performance. Comparisons with baseline methods (e.g., autoencoders, isolation forests) highlight gains from the VAE and multimodal design [10, 12]. Stratification by drug class and therapeutic area further identifies strengths and limitations.

Validation protocols

Validation uses temporal splits, training on earlier (2017–2023) data and testing on later periods to simulate prospective detection and prevent information leakage [8, 15]. Performance is also compared with spontaneous reporting systems to assess timeliness and sensitivity [3, 4], with expert chart review providing qualitative validation.

Cross-institutional testing projects new data into the learned latent space without retraining, demonstrating generalizability [17, 18]. Sensitivity analyses (e.g., latent size, β values) assess robustness. The protocol remains fully unsupervised, supporting realistic evaluation and future deployment readiness.

Conclusion

The variational autoencoder framework presented here provides a comprehensive conceptual approach for unsupervised ADR signal detection by jointly modeling clinical notes and laboratory abnormalities extracted from routine EHR records spanning 2017 to 2023. The multimodal architecture learns a normal data distribution in an entirely label-free manner and flags deviations through reconstruction error and latent space analysis. This design integrates seamlessly with existing healthcare data pipelines without requiring any annotated ADR examples. The resulting system outputs prioritized signals ready for pharmacovigilance review.

Key advantages of the framework include its complete independence from labeled ADR data, enabling detection of previously unknown reactions that supervised methods cannot address. Multimodal fusion captures complementary signals from both unstructured symptom descriptions and objective laboratory trends, improving sensitivity over single-modality baselines. The generative properties of the VAE further support interpretability through latent space clustering and attribution techniques. These strengths position the approach as a scalable solution for proactive post-market drug safety surveillance.

Limitations include the requirement for large volumes of unlabeled training data to accurately model normal distributions and the challenge of distinguishing true ADRs from other clinical anomalies such as disease progression or documentation artifacts. Validation remains dependent on retrospective comparison with spontaneous reporting systems, which themselves suffer from underreporting. Future refinements may incorporate additional modalities or temporal modeling extensions while preserving the core unsupervised principle. These constraints are inherent to any generative synthetic data framework operating in complex real-world healthcare environments.

Implementation on large-scale EHR databases such as CPRD, Optum, or MIMIC is strongly encouraged to realize the framework's potential for enhancing pharmacovigilance. Direct comparison against spontaneous reporting systems in prospective settings will further quantify incremental value in early signal detection. Adoption of this VAE-based methodology could accelerate the identification of drug safety issues and reduce the public health burden of unrecognized ADRs. Overall, the conceptual framework advances unsupervised anomaly detection as a foundational tool for next-generation pharmacovigilance using generative and synthetic data principles.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Mar 2023 Revised: 01 Jun 2023 Accepted: 24 Jul 2023
Published online: 20 January 2024

References

Lee S, Choi J, Kim HS, Kim GJ, Lee KH, Park CH, et al. Standard-based comprehensive detection of adverse drug reaction signals from nursing statements and laboratory results in electronic health records. J Am Med Inform Assoc. 2017;24(4):697-708.

Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf. 2019;42(1):99-111. https://doi.org/10.1007/s40264-018-0762-y.

Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020;27(1):3-12.

Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication

safety: a narrative review of recent advances and challenges. Pharmacotherapy. 2018;38(8):822-41. https://doi.org/10.1002/phar.2152.

Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018;25(3):331-6.

Yang X, Bian J, Wu Y. Detecting medications and adverse drug events in clinical notes using recurrent neural networks. In: International Workshop on Medication and Adverse Drug Event Detection; 2018 May 16; New York. PMLR; 2018. p. 1-6.

Longoni C, Cian L, Kyung EJ. Algorithmic transference: people overgeneralize failures of AI in the government. J Mark Res. 2023;60(1):170-88. https://doi.org/10.1177/00222437221109743.

Liao W, Derijks HJ, Blencke AA, De Vries E, Van Seyen M, Van Marum RJ. Dual autoencoders modeling of electronic health records for adverse drug event preventability prediction. Intell Based Med. 2022;6:100077. https://doi.org/10.1016/j.ibmed.2022.100077.

Sandfort V, Yan K, Graffy PM, Pickhardt PJ, Summers RM. Use of variational autoencoders with unsupervised learning to detect incorrect organ segmentations at CT. Radiol Artif Intell. 2021;3(4):e200218. https://doi.org/10.1148/ryai.2021200218.

Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22(5):1589-604. https://doi.org/10.1109/JBHI.2017.2767063.

Akrami H, Joshi A, Aydore S, Leahy R. Quantile regression for uncertainty estimation in VAEs with applications to brain lesion detection. In: Int Conf Inf Process Med Imaging; 2021 Jun 14; Cham. Springer; 2021. p. 689-700. https://doi.org/10.1007/978-3-030-78191-0_53.

Mansour RF, Escorcia-Gutierrez J, Gamarra M, Gupta D, Castillo O, Kumar S. Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. Pattern Recognit Lett. 2021;151:267-74. https://doi.org/10.1016/j.patrec.2021.08.021.

Pinaya WHL, Tudosiu PD, Gray R, Rees G, Nachev P, Ourselin S, et al. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Med Image Anal. 2022;79:102475. https://doi.org/10.1016/j.media.2022.102475.

Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Saf. 2019;42(1):123-33. https://doi.org/10.1007/s40264-018-0763-x.

Mesbah S, Yang J, Sips RJ, Torre MV, Lofi C, Bozzon A, et al. Training data augmentation for detecting adverse drug reactions in user-generated content. In: Proc Conf Empir Methods Nat Lang Process Int Jt Conf Nat Lang Process (EMNLP-IJCNLP); 2019 Nov; Stroudsburg. ACL; 2019. p. 2349-59.

Lee S, Lee JH, Kim GJ, Kim JY, Shin H, Ko I, et al. A data-driven reference standard for adverse drug reaction (RS-ADR) signal assessment: development and validation. J Med Internet Res. 2022;24(10):e35464. https://doi.org/10.2196/35464.

Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: a scoping review of supervised learning methods. PLoS One. 2023;18(1):e0279842. https://doi.org/10.1371/journal.pone.0279842.

Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. AMIA Summits Transl Sci Proc. 2021;2021:420-9.

Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7(2):e12239. https://doi.org/10.2196/12239.

Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration. Inf Sci. 2017;418-419:546-60. https://doi.org/10.1016/j.ins.2017.08.045.

Wu C, Wu F, Wu S, Yuan Z, Liu J, Huang Y. Semi-supervised dimensional sentiment analysis with variational autoencoder. Knowl Based Syst. 2019;165:30-9. https://doi.org/10.1016/j.knosys.2018.11.016.

Ghosh SS, Dhar R, Marcus DS, Sotiras A. Siam-VAE: a hybrid deep learning based anomaly detection framework for automated quality control of head CT scans. In: Medical Imaging 2023: Computer-Aided Diagnosis; 2023 Apr 7; Bellingham. SPIE; 2023. p. 184-90. https://doi.org/10.1117/12.2653920.

Ehrhardt J, Wilms M. Autoencoders and variational autoencoders in medical image analysis. In: Biomedical image synthesis and simulation. London: Academic Press; 2022. p.

129-62.
https://doi.org/10.1016/B978-0-12-821273-4.00011-2.

Röchner P, Rothlauf F. Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. BMC Med Res Methodol. 2023;23(1):125. https://doi.org/10.1186/s12874-023-01935-z.

Luo Q, Chen J, Zi Y, Chang Y, Feng Y. Multi-mode non-Gaussian variational autoencoder network with missing sources for anomaly detection of complex electromechanical equipment. ISA Trans. 2023;134:144-58. https://doi.org/10.1016/j.isatra.2022.08.030.

Mower J, Bernstam E, Xu H, Myneni S, Subramanian D, Cohen T. Improving pharmacovigilance signal detection from clinical notes with locality sensitive neural concept embeddings. AMIA Summits Transl Sci Proc. 2022;2022:349-58.

Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In: Text Analysis Conference (TAC); 2017; Gaithersburg. NIST; 2017.