Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

REGULATORY TEXT MINING SYSTEM FOR PHARMACEUTICAL QUALITY RISK DETECTION FROM GUIDELINES AND DEVIATION REPORTS

Bruno Martins[1]*, Lucas Pereira[1], Renata Azevedo[2], Pedro Costa[1]

1. ***Department of Computational Pharmacology, Faculty of Pharmacy, University of Minho, Braga, Portugal.***
2. ***Department of Pharmaceutical Intelligence Systems, Faculty of Pharmacy, University of Porto, Porto, Portugal.***

ARTICLE INFO

Received:
08 December 2024
Received in revised form:
27 March 2025
Accepted:
03 April 2025
Available online:
28 April 2025

Keywords: Regulatory text mining, Pharmaceutical quality, Deviation reports, CAPA, Natural language processing, Knowledge graph

ABSTRACT

Pharmaceutical quality relies on the proactive detection of process, product, and compliance risks, yet critical signals are often hidden within unstructured sources such as regulatory guidance, deviation narratives, CAPA records, inspection observations, and other quality documents, which are typically reviewed in a fragmented manner. Traditional quality risk management depends heavily on manual document review and local expertise, making it challenging to identify recurring issues across sites, benchmark internal deviations against external regulatory expectations, or develop a comprehensive view of emerging risks. This article proposes an AI-powered regulatory text mining system that ingests regulatory guidelines, deviation reports, and CAPA records to extract risk entities and their relationships, link them to manufacturing processes, and build a queryable quality-risk knowledge graph. The framework integrates document ingestion, preprocessing, named entity recognition, relation extraction, transformer-based risk classification, knowledge graph construction, and dashboard-based decision support, with human verification to ensure interpretability, auditability, and compliance with regulatory standards. By converting scattered textual information into actionable quality-risk intelligence, the system enables quality teams to anticipate compliance gaps, prioritize CAPA activities, and respond more rapidly to evolving regulatory expectations, shifting pharmaceutical organizations from reactive documentation toward predictive, science-based quality oversight.

To Cite This Article: Martins B, Pereira L, Azevedo R, Costa P. Regulatory Text Mining System for Pharmaceutical Quality Risk Detection from Guidelines and Deviation Reports. Pharmacophore. 2025;16(2):32-42. https://doi.org/10.51847/vwHtDaETbQ

Introduction

Pharmaceutical quality organisations generate a large volume of textual documentation through deviation investigations, CAPA records, batch records, complaints, audit observations, and regulatory updates. Much of this content is data rich but information poor, because the relevant risk signals are embedded in narrative descriptions rather than structured fields. Prior work on manufacturing deviation analysis suggests that machine learning can support root-cause reasoning when quality events are transformed into analysable representations [1]. In pharmaceutical manufacturing contexts, language models have also been proposed as tools that could assist deviation investigations by helping reviewers interpret narrative evidence and organise investigation logic [2].

Manual review remains central to quality risk management, but document-by-document assessment can create silos between regulatory intelligence, quality assurance, production, and technical operations. Regulatory text mining studies in drug labelling have shown that structured extraction from regulated documents can make recurring safety or compliance-relevant information easier to identify and compare [3]. Similar extraction principles have been used to identify adverse events and terminology changes from product labels, demonstrating how regulatory narratives can be converted into standardised signals [4, 5]. In pharmaceutical quality, a comparable approach could help detect repeated process weaknesses that remain hidden when deviation reports are reviewed only as isolated events.

Corresponding Author: Bruno Martins; Department of Computational Pharmacology, Faculty of Pharmacy, University of Minho, Braga, Portugal. E-mail: bruno.martins@gmail.com.

Modern natural language processing could enable continuous mapping of the quality-risk landscape by extracting entities, relations, and document-level risk categories from heterogeneous texts. Domain-specific models such as PharmBERT and RxBERT show how pharmaceutical language can be represented more effectively when models are adapted to drug and regulatory terminology [6, 7]. BERT-based approaches have also been applied to classify drug labelling text for regulatory risk interpretation, supporting the broader case for transformer models in specialised pharmaceutical text understanding [8]. These methods indicate that a regulatory text mining framework could link internal deviation language with external regulatory expectations in a more systematic way.

This article proposes an AI-powered regulatory text mining system for pharmaceutical quality risk detection from guidelines and deviation reports. The system would ingest regulatory and quality documents, extract structured risk entities, connect those entities through a knowledge graph, and surface actionable intelligence through a dashboard. Work on regulatory-document classification and large language model frameworks for transparent regulatory use supports the need for traceability, interpretability, and source-linked outputs in such systems [9, 10]. The proposed AIF is intended to augment, rather than replace, the traditional pharmaceutical quality management framework by making textual evidence easier to search, compare, and act upon. The proposed system should be explicitly aligned with established pharmaceutical quality governance so that text-mined risk signals can be interpreted within recognised lifecycle, validation, and computerized-system controls. ICH Q9(R1) supports the use of systematic quality risk management for better-informed, timely decisions, while ICH Q10 frames these activities within an effective pharmaceutical quality system across the product lifecycle [11]. FDA process validation guidance further reinforces the need to connect process knowledge, continued verification, and manufacturing evidence when assessing recurring deviations. For the digital components of the proposed framework, GAMP 5 provides a risk-based basis for ensuring that computerized systems are fit for intended use and compliant with applicable GxP expectations.

The types of pharmaceutical quality and regulatory documents considered in this study are summarised in **Table 1**.

Table 1. Pharmaceutical quality and regulatory document sources relevant to AI-based risk detection

Document source	Typical content	Risk-related information that may be extracted
Deviation reports	Investigation narratives, event descriptions, immediate causes, root-cause analysis, corrective actions	Recurring deviations, repeated failure modes, process weaknesses, unresolved root causes
CAPA records	Corrective and preventive action plans, effectiveness checks, action owners, closure evidence	Repeated corrective actions, delayed CAPA closure, ineffective actions, recurring preventive gaps
Batch records	Manufacturing steps, process parameters, operator entries, in-process checks, exceptions	Process variability, repeated operational errors, parameter excursions, batch-to-batch trends
Complaints	Product quality complaints, customer observations, defect descriptions, investigation outcomes	Repeated product defects, packaging issues, usability concerns, potential patient-impact signals
Audit observations	Internal or external audit findings, compliance gaps, procedural weaknesses, auditor comments	Recurrent GMP deficiencies, documentation gaps, training issues, process-control weaknesses
Regulatory guidance and updates	Regulatory expectations, inspection trends, updated compliance requirements, guideline changes	Emerging compliance risks, changing regulatory expectations, gaps between internal practice and external standards

Background

Pharmaceutical Quality Risk Management and ICH Q9

Pharmaceutical quality risk management depends on science-based decisions that consider process knowledge, product impact, patient safety, and regulatory expectations. Deviation reports, complaint narratives, CAPA records, and inspection observations provide textual evidence that can inform these decisions when they are systematically interpreted. Machine learning approaches to manufacturing quality deviations show that failure descriptions and investigation records can be analysed as signals of recurring operational risk [1]. In pharmaceutical settings, language-model support for deviation investigations could help reviewers structure evidence, compare similar cases, and identify investigation themes that should be escalated [2].

Natural Language Processing for Regulatory Science

Natural language processing has been widely explored in pharmacovigilance, drug labelling, clinical text, and biomedical literature mining, but its use in pharmaceutical manufacturing quality remains less mature. Systems that annotate drug product labels with MedDRA terminology demonstrate how regulated text can be mapped to controlled vocabularies for downstream review [3]. Adverse event extraction from structured product labels and other regulatory sources shows that NLP can support signal identification when textual evidence is converted into standardised data elements [4, 12]. Reviews of NLP in pharmacology and pharmacovigilance further indicate that text mining is becoming an important foundation for regulatory intelligence, even though direct applications to GMP quality documents still require further design [13, 14].

Structure of Deviation Reports and CAPA Documents

Deviation reports and CAPA records usually contain event descriptions, immediate actions, root-cause narratives, impact assessments, corrective actions, preventive actions, and effectiveness checks. These sections often vary across sites, products, and electronic QMS templates, which makes standardised extraction difficult. Manufacturing quality deviation studies highlight the importance of transforming heterogeneous failure narratives into consistent analytical features before automated reasoning can be useful [1]. Pharmaceutical deviation-support language models would therefore need to preserve context, distinguish observed events from suspected causes, and keep extracted findings traceable to source text [2].

Transformer-Based Models for Specialised Text Understanding

Transformer-based language models are well suited to specialised pharmaceutical text because they can represent context-dependent terminology, abbreviations, and regulatory phrasing. PharmBERT was developed as a domain-specific BERT model for drug labels, showing how adaptation to pharmaceutical corpora can improve the suitability of language representations for regulated document analysis [6]. RxBERT similarly focuses on drug labelling text mining and indicates that domain-aware language modelling can support classification and extraction tasks in regulatory documents [7]. BERT-based classification of drug labelling documents for drug-induced liver injury risk further illustrates how transformer models can be adapted for risk-oriented regulatory text interpretation [8].

Knowledge Graphs for Quality Risk Integration

Knowledge graphs provide a structure for linking products, processes, parameters, materials, equipment, failure modes, investigations, CAPA actions, and regulatory expectations. In a pharmaceutical quality system, extracted entities could become graph nodes, while causal, temporal, and compliance relationships could become graph edges. Knowledge graph methods in pharmacovigilance show how drug-event and literature-derived relationships can support safety reasoning across fragmented sources [15, 16]. Related work on pharmacokinetic interaction graphs and knowledge graph guidance also demonstrates how curated biomedical relationships can be organised for queryable, evidence-linked decision support [17, 18].

System Overview

High-Level Architecture

The proposed architecture begins with a document ingestion pipeline that receives periodic batches of regulatory guidelines and real-time streams of internal deviation, CAPA, and inspection-observation text. A preprocessing layer normalises document structure before the NLP engine extracts entities, relationships, and risk categories from each segment. The extracted outputs populate a knowledge graph that connects risk signals to products, unit operations, materials, sites, and regulatory clauses. Dashboard interfaces would then support search, visualisation, and review of risk hot spots, following the broader direction of AI-enabled regulatory document analysis and transparent regulatory language-model frameworks [9, 10].

Figure 1 presents the proposed regulatory text mining architecture, showing how fragmented pharmaceutical quality and regulatory narratives can be transformed into source-linked, human-verified, and actionable quality-risk intelligence.

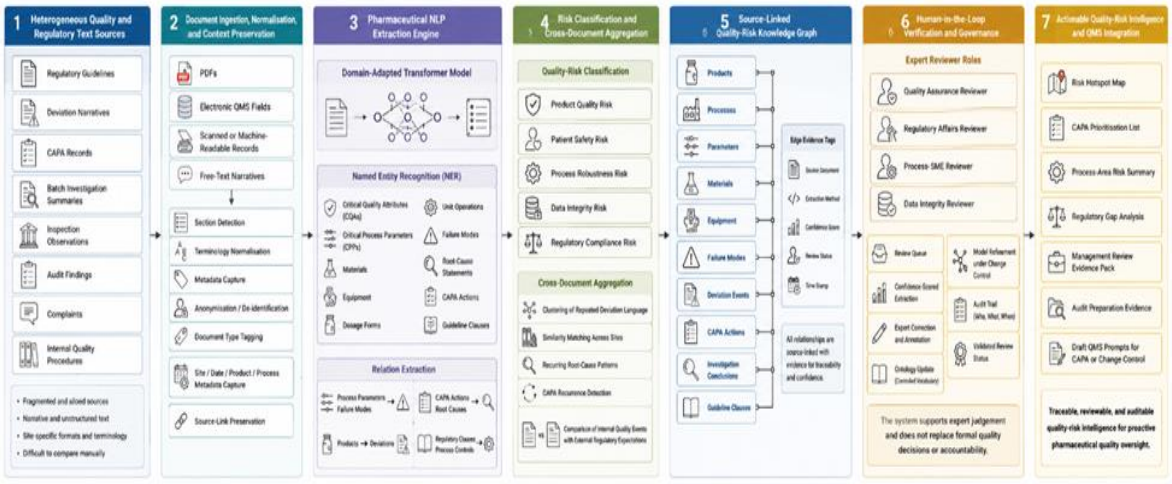

Figure 1. Regulatory Text Mining Architecture for Pharmaceutical Quality Risk Detection from Guidelines, Deviation Reports, and CAPA Records

Core Input Sources and Outputs

The core inputs would include regulatory guideline PDFs, deviation report narratives, CAPA records, batch investigation summaries, inspection observations, and relevant quality procedures. The expected outputs would include source-linked risk entities, risk-process-product triples, draft process-area risk summaries, and prioritised CAPA opportunities for expert review. Drug labelling NLP systems show the importance of retaining source attribution when extracted text is used for regulatory

interpretation [5]. Similar traceability would be essential in quality risk mining, where every extracted risk statement should remain linked to the original deviation or guideline passage.

Design Principles

The system should be domain-adapted for pharmaceutical language, traceable to source documents, compatible with data integrity expectations, and configured for human-in-the-loop validation. Because regulated use requires more than prediction, the architecture should make model outputs inspectable and reviewable rather than treating them as final decisions. Frameworks for using large language models in regulatory environments emphasise transparency and trustworthiness, which are directly relevant to quality risk extraction [10]. Automated comparison tools for FDA labelling changes further show that AI-supported regulatory workflows should preserve evidence, context, and reviewer oversight [19].

Table 2 defines the functional architecture of the proposed regulatory text mining system by linking each technical layer to its pharmaceutical quality-risk contribution and required governance control.

Table 2. Functional Architecture of the Regulatory Text Mining System for Pharmaceutical Quality Risk Detection

System layer	Primary input	Analytical function	Structured output	Quality-risk contribution	Required governance control
Document ingestion layer	Regulatory guidelines, deviation reports, CAPA records, batch investigations, inspection observations, complaints, audit findings, and internal quality procedures	Converts heterogeneous document formats into machine-readable text while preserving document identity and context	Standardised document corpus with document type, source, site, date, product, and process metadata	Creates a unified evidence base across fragmented regulatory and internal quality sources	Controlled access, document versioning, source attribution, and data integrity checks
Preprocessing and normalisation layer	Extracted text, headings, narrative fields, QMS templates, terminology variants, and metadata fields	Normalises pharmaceutical terminology, detects document sections, segments text into meaningful units, and masks unnecessary personal information	Cleaned and segmented text units linked to their original document passages	Reduces noise caused by inconsistent local wording, template variation, and unstructured quality narratives	Anonymisation rules, terminology dictionaries, preprocessing validation, and retention of audit-relevant technical context
Named entity recognition layer	Segmented regulatory and quality text	Identifies critical quality attributes, critical process parameters, materials, equipment, dosage forms, unit operations, failure modes, root causes, CAPA actions, and guideline clauses	Source-linked entity mentions with entity type, text span, confidence score, and document location	Converts free-text quality information into analysable quality-risk elements	Expert-reviewed entity schema, annotation guidelines, confidence thresholds, and periodic error review
Relation extraction layer	Recognised entities and surrounding textual context	Detects relationships among processes, parameters, failures, materials, CAPA actions, root causes, products, and regulatory expectations	Process–failure, CAPA–root cause, product–deviation, material–risk, and clause–control relationships	Makes causal, operational, and compliance links visible across documents	Human adjudication of ambiguous causal claims, source-linked evidence trails, and relation-type validation
Transformer-based risk classification layer	Entity-rich text segments and extracted relationships	Assigns conceptual risk categories such as product quality, patient safety, process robustness, data integrity, and regulatory compliance	Risk-labelled text segments and document-level risk classifications	Prioritises risk signals for expert review without treating model output as a final quality decision	Class-label definitions, reviewer confirmation, model performance monitoring, and documented change control
Cross-document aggregation layer	Risk-labelled segments, repeated terms, similar deviation narratives, CAPA links, and process metadata	Clusters recurring quality themes across documents, sites, products, and time periods	Recurrence clusters, emerging-risk themes, repeated root-cause patterns, and CAPA opportunity lists	Reveals systemic risks that may remain hidden in document-by-document review	Duplicate detection, reviewer assessment of cluster validity, temporal review, and escalation rules
Knowledge graph layer	Extracted entities, relationships, risk labels, source metadata, confidence scores, and review status	Links products, processes, parameters, materials, equipment, failure modes, deviation events, CAPA actions, investigation conclusions, and guideline clauses	Queryable quality-risk knowledge graph with source-linked nodes and edges	Enables traversal from individual events to broader process, product, and regulatory evidence networks	Provenance tracking, graph schema governance, review status labels, and audit trail for graph updates

Human verification layer	Confidence-scored extractions, graph links, unresolved ambiguities, and high-priority risk signals	Allows QA, regulatory affairs, process SMEs, and data integrity reviewers to confirm, correct, reject, or escalate model outputs	Verified entities, validated relationships, corrected ontology terms, and reviewer-approved risk summaries	Ensures that automated extraction supports expert judgement rather than replacing regulated decision making	Role-based review, electronic signature where applicable, documented reviewer rationale, and controlled model refinement
QMS integration and decision-support layer	Verified risk graph outputs, CAPA patterns, guideline links, and process-area summaries	Provides quality dashboards, regulatory gap analysis, CAPA prioritisation, management review evidence, and draft QMS prompts	Risk hotspot summaries, CAPA prioritisation lists, audit evidence packs, and change-control prompts	Translates text mining results into operational quality oversight and proactive risk management	Final quality-unit approval, workflow integration controls, periodic effectiveness review, and avoidance of autonomous quality decisions

Document Ingestion and Preprocessing

Handling Heterogeneous Document Formats and Structures

The ingestion module would extract text from PDFs, structured electronic forms, scanned-quality records when available in machine-readable form, and free-text narrative fields from QMS platforms. It would normalise chemical names, dosage-form terms, equipment identifiers, units, and process terminology so that downstream NLP receives consistent input. Drug label text mining studies show that regulated documents contain semi-structured sections whose headings and terminology must be recognised before classification or extraction can be reliable [3, 9]. In pharmaceutical quality records, a comparable preprocessing step would help align deviation descriptions, CAPA text, and regulatory clauses within a common analytical structure.

Segmentation and Context Preservation

Segmentation would divide documents into semantically meaningful units, such as a single deviation event, a root-cause statement, a CAPA action, or a regulatory-guideline paragraph. At the same time, the system should preserve metadata such as document type, site, date, product, process step, batch reference, and authoring workflow stage. Natural language processing systems for drug labels and adverse event extraction show that extracted terms are most useful when they remain connected to their original document section and interpretive context [4, 5]. For quality risk detection, this means that a phrase such as "mixing time exceeded limit" should remain associated with the relevant product, process, and investigation stage rather than being treated as an isolated text fragment.

Anonymisation and Confidentiality Safeguards

Before NLP processing, the system should remove or mask personally identifiable information relating to deviation authors, reviewers, operators, complainants, or patients when such details are present. This safeguard would allow the system to focus on process and product risk while reducing unnecessary exposure of personal data. Work on extracting demographic fields from adverse event reporting systems illustrates that NLP can identify sensitive personal attributes in regulatory safety text, which reinforces the need for controlled preprocessing in regulated environments [12]. In a pharmaceutical quality setting, anonymisation should be designed so that confidentiality is protected while audit-relevant technical context is retained.

Text Mining and Risk Extraction Engine

Named Entity Recognition for Quality Attributes, Processes, and Failures

The named entity recognition module would identify mentions of critical quality attributes, critical process parameters, materials, equipment, dosage forms, unit operations, failure modes, suspected causes, and CAPA actions. Domain-adapted transformer models would be appropriate because pharmaceutical quality language includes specialised terms, abbreviations, and context-dependent expressions that general models may misinterpret. PharmBERT and RxBERT demonstrate how pharmaceutical-domain language models can support more suitable representations for drug and regulatory text [6, 7]. In the proposed system, similar adaptation would help the model recognise quality-specific entities such as dissolution trend, blend uniformity, granulation moisture, feeder speed, and compression force.

Relation Extraction and Risk Classification

The relation extraction module would identify links between extracted entities, such as a process parameter associated with a failure mode, a material attribute linked to a batch deviation, or a CAPA action connected to a recurring root cause. A risk classification layer would then assign conceptual labels such as product quality, patient safety, process robustness, data integrity, or regulatory compliance, while leaving final interpretation to qualified personnel. Clinical and regulatory adverse-event extraction systems show that NLP pipelines can combine entity recognition with event-level interpretation when relationships between terms are explicitly modelled [20, 21]. Multi-domain adverse drug event extraction benchmarks further indicate that relation-focused extraction should be evaluated across varied text styles before it is relied on for regulated decision support [22].

Systemic Risk Aggregation across Documents

The aggregation layer would cluster similar risk descriptions across deviation reports, CAPA records, audit observations, and relevant regulatory guidance. By comparing repeated phrases, shared entities, and linked process contexts, the system could flag patterns that would not be apparent from any single report. Systematic reviews of adverse-event extraction and pharmacovigilance text mining show that NLP can be used to synthesise recurring signals across heterogeneous document collections when extraction outputs are normalised [23, 24]. Quality 4.0 literature also supports the idea that analytics can transform dispersed quality data into decision-support knowledge, provided that human review and operational context remain central [25].

As shown in **Figure 2**, the aggregation layer brings together risk-related information from deviation reports, CAPA records, audit observations, and regulatory guidance. By clustering similar descriptions and comparing repeated phrases, shared entities, and linked process contexts, the system can identify systemic risk patterns that may remain hidden within individual documents.

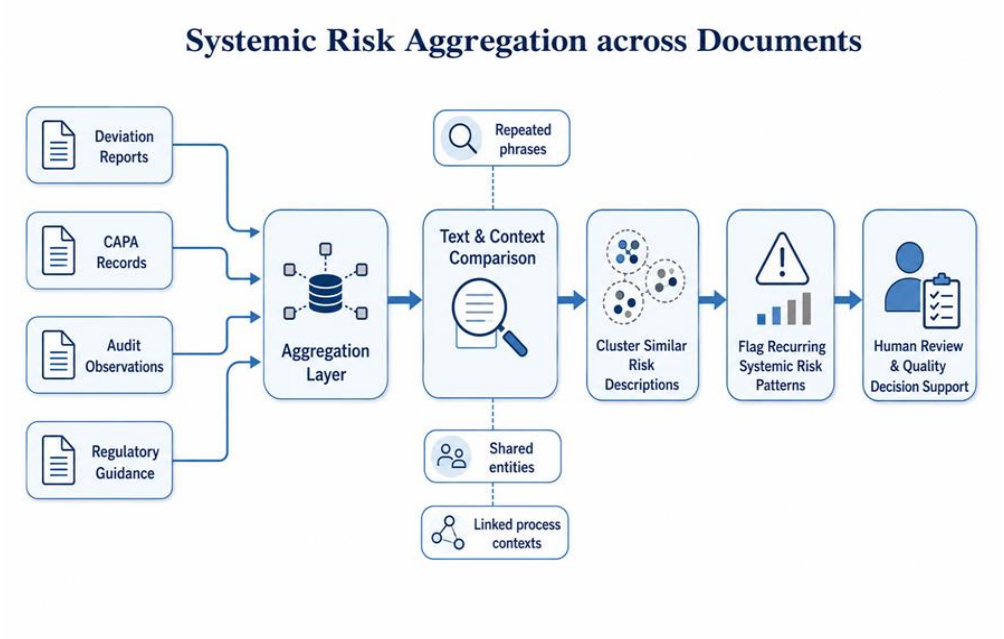

Figure 2. Systemic Risk Aggregation across Documents.

The aggregation layer consolidates heterogeneous quality and regulatory documents, normalises extracted risk information, and clusters similar risk descriptions across sources. This enables the detection of recurring systemic patterns that may not be visible in a single report, while preserving the role of human review and operational context in final quality decision-making.

Knowledge Graph Construction and Risk Linking

Graph Schema and Population

The knowledge graph schema would represent products, processes, parameters, materials, equipment, failure modes, guideline clauses, deviation events, CAPA actions, and investigation conclusions as interconnected nodes. Relations extracted by the NLP engine would become edges, such as "parameter associated with failure mode," "CAPA addresses root cause," or "guideline clause relevant to process control." Literature-derived biomedical knowledge graphs show how noisy text-mined relationships can be structured into reusable decision-support resources when provenance and validation are retained [26]. In the proposed system, each graph edge should therefore preserve its source document, extraction method, review status, and applicable quality context.

Querying and Traversing the Risk Graph

Quality personnel could query the graph to identify paths across documents, such as all CAPA records related to compression failures that are linked to a specific regulatory expectation. A user could also traverse from a process step to all related deviations, associated materials, suspected causes, and relevant guideline passages. Knowledge graph approaches in pharmacovigilance demonstrate how connected representations can support exploration of drug-event relationships across fragmented evidence sources [15, 16]. A similar graph design for quality risk would allow reviewers to move from a single deviation narrative to a broader evidence network rather than relying only on keyword search.

Temporal Analysis and Trend Detection

The graph would include temporal attributes so that risk relationships could be reviewed before and after CAPA implementation, procedure changes, supplier changes, or equipment modifications. Temporal edges would help the system

distinguish recurring, resolved, emerging, and potentially escalating issues in a process area. Knowledge graph guidance in pharmacovigilance emphasises that connected evidence should remain interpretable and traceable as it is updated over time [18]. In pharmaceutical quality, this would allow management review teams to examine whether a repeated deviation theme appears to persist despite preventive action, while avoiding unsupported claims of automated causality.

Human-in-the-Loop Verification and Model Refinement

Expert-Curated Feedback Loop

Quality assurance specialists, process subject matter experts, and regulatory affairs reviewers would periodically examine extracted entities, relations, and graph links. Their corrections would be logged as curated feedback that could guide model refinement, ontology updates, and changes to extraction rules. Human oversight is particularly important because regulatory NLP systems that classify or summarise drug labelling text must maintain reviewer trust and source-level transparency [10, 27]. In this framework, model refinement would be treated as a controlled quality activity rather than an informal technical adjustment.

Confidence-Scored Extraction and Escalation

Each extraction should carry a confidence score and a review status so that uncertain outputs are escalated for human assessment instead of being automatically promoted as validated risk intelligence. Low-confidence entity mentions, ambiguous causal statements, and incomplete CAPA links would be held in a review queue until a qualified reviewer confirms, corrects, or rejects them. AI tools for identifying adverse event changes in FDA labelling illustrate the importance of validation workflows when automated outputs may influence regulatory interpretation [19]. Likewise, AskFDALabel demonstrates how language-model-supported regulatory analysis should be framed as assisted review rather than autonomous decision making [28].

Integration Into Quality Management Systems

Embedding Extracted Risks into CAPA and Change Control Workflows

When the system identifies a recurring high-concern pattern, it could generate a draft CAPA or change-control prompt within the electronic QMS, pre-populated with source excerpts, affected process steps, linked products, and proposed risk categories. Such a draft should remain subject to quality-unit review and should not replace formal investigation or approval workflows. Deep learning applications for adverse drug event detection from structured and unstructured regulatory data show that AI outputs can support regulated review when they are integrated with appropriate human decision points [29]. In a QMS setting, the value of integration would come from accelerating evidence gathering and cross-document comparison rather than automating final quality decisions.

Real-Time Regulatory Intelligence and Gap Analysis

New guideline releases, regulatory updates, and relevant labelling changes could be processed as they become available, with extracted requirements compared against internal procedures, deviation themes, and process controls. The system could highlight potential gaps where internal practice appears weakly aligned with new or clarified regulatory expectations. Automated regulatory-document comparison methods show how AI can support the identification of meaningful text changes when outputs remain source-linked and reviewable [19]. Text summarisation methods for drug labelling also suggest that language models could help reviewers interpret lengthy regulatory updates, provided that summaries are treated as aids rather than substitutes for expert assessment [27].

Evaluation Strategy

Information Extraction Performance

The extraction component should be evaluated against a manually annotated reference set created by qualified reviewers, with attention to entity boundaries, entity types, relation types, and source attribution. Although standard information extraction measures such as precision, recall, and F-score may be used, the article does not propose or report numerical performance outcomes. Shared tasks on medication, indication, and adverse event extraction show that gold-standard annotation is useful for evaluating whether NLP systems capture clinically and regulatorily meaningful information [20, 30]. For quality risk mining, similar evaluation should assess whether the system reliably captures process parameters, failure modes, CAPA links, and guideline references in a traceable manner.

Risk Detection and Prioritisation Utility

The risk prioritisation layer should be evaluated through expert adjudication rather than by assuming that automatically ranked risks are inherently correct. Reviewers could examine whether surfaced patterns are plausible, actionable, and supported by source documents, and whether they represent issues that manual periodic review might otherwise overlook. Systematic reviews of pharmacovigilance machine learning highlight the need to evaluate not only model outputs but also their practical contribution to expert workflows [13]. In pharmaceutical quality, the most relevant assessment would be whether the system supports better risk conversations, clearer CAPA prioritisation, and stronger evidence preparation for audits.

User Acceptance and Operational Efficiency

User acceptance should be assessed by quality professionals, process owners, regulatory affairs staff, and auditors who interact with the dashboard and graph interface. Evaluation should consider usability, trustworthiness, explanation quality, traceability, and perceived value during audit preparation, management review, and CAPA planning. Quality 4.0 literature emphasises that analytics systems must be embedded into operational routines and decision cultures to create sustained value [25]. The proposed framework should therefore be assessed as a sociotechnical system, where model behaviour, reviewer confidence, workflow fit, and governance controls all influence adoption.

Table 3 consolidates the traceability, validation, and governance safeguards required for using regulatory text mining outputs in pharmaceutical quality-risk management without replacing expert judgement.

Table 3. Traceability, Validation, and Governance Framework for Regulated Use of Pharmaceutical Quality Text Mining

Governance dimension	Why it matters in pharmaceutical quality text mining	Practical implementation mechanism	Validation or review evidence	Risk if omitted
Source-level traceability	Extracted risk statements must be auditable and defensible in regulated quality environments	Link every entity, relationship, risk label, and graph edge to the original document, passage, section, and version	Source passage review logs, document identifiers, version records, and reviewer confirmation	Risk signals may become detached from evidence, reducing audit defensibility and reviewer trust
Context preservation	Quality meaning depends on product, site, batch, process step, investigation stage, and regulatory context	Preserve metadata for document type, product, batch, site, process area, date, authoring workflow stage, and investigation section	Metadata completeness checks and reviewer sampling of context-linked extractions	The system may misclassify isolated phrases or overgeneralise local deviations across inappropriate contexts
Controlled terminology and ontology management	Site-specific wording and inconsistent abbreviations can fragment risk interpretation	Maintain a controlled vocabulary for CQAs, CPPs, unit operations, failure modes, materials, equipment, root causes, CAPA actions, and guideline clauses	Ontology change logs, synonym mapping records, and periodic SME review	Similar risks may be split across different labels, preventing reliable aggregation and trend detection
Human-in-the-loop verification	Model outputs may influence CAPA prioritisation, audit preparation, and regulatory gap analysis	Route low-confidence or high-impact extractions to QA, regulatory affairs, process SMEs, or data integrity reviewers	Review queue records, accepted/rejected extraction logs, reviewer comments, and escalation decisions	Automated outputs may be mistaken for validated conclusions, creating compliance and quality decision risk
Confidence scoring and escalation	Not all extractions have equal reliability or decision relevance	Attach confidence scores and review status labels to entities, relations, classifications, and graph links	Threshold justification, review-status reports, and exception handling records	Ambiguous or weakly supported outputs may enter management review or CAPA workflows prematurely
Model performance evaluation	Extraction and classification components must be shown to capture meaningful quality information	Evaluate entity boundaries, entity types, relation types, risk labels, source attribution, and reviewer agreement against annotated reference sets	Annotation guidelines, gold-standard datasets, precision/recall/F-score summaries where applicable, and error analyses	The system may perform acceptably on general text but fail on pharmaceutical deviation and CAPA language
Data integrity and audit trail	Regulated quality systems require evidence that records and decisions are complete, accurate, and attributable	Log document ingestion, preprocessing, extraction, reviewer actions, model updates, graph revisions, and dashboard outputs	Audit trail exports, access logs, change records, and system validation documentation	Users may be unable to reconstruct how a risk signal was generated, reviewed, or modified
Change control for model and ontology updates	Regulatory language, product portfolios, QMS templates, and manufacturing processes evolve over time	Manage model retraining, prompt changes, ontology updates, preprocessing rules, and graph schema changes through formal change control	Change requests, impact assessments, validation results, approval records, and release notes	Uncontrolled changes may alter system behaviour without documented justification or revalidation
Confidentiality and anonymisation	Quality records may include personal identifiers for operators, reviewers, complainants, or patients	Mask unnecessary personal information while retaining process, product, batch, and investigation context	Anonymisation test records, privacy review, and verification that technical meaning is preserved	Sensitive personal data may be unnecessarily exposed during text mining or dashboard review
Workflow integration boundaries	The system should support, not replace, formal quality decisions	Integrate outputs as draft evidence, review prompts, risk summaries, or CAPA prioritisation aids requiring quality-unit approval	SOPs, user training records, approval workflows, and management review documentation	AI-generated suggestions may be treated as final decisions, weakening accountability and regulatory defensibility

Periodic effectiveness review	System value depends on sustained usefulness in CAPA planning, management review, audit readiness, and regulatory intelligence	Conduct scheduled review of surfaced risks, missed signals, false positives, user adoption, and CAPA relevance	Periodic quality review reports, user feedback, CAPA linkage analysis, and dashboard usage metrics	The system may become technically functional but operationally irrelevant or poorly trusted
Bias and uneven documentation quality monitoring	Sites, products, or teams with richer documentation may appear to carry more risk than poorly documented areas	Compare risk outputs against documentation completeness, reporting practices, site templates, and reviewer feedback	Documentation-quality metrics, site-level calibration review, and SME adjudication of apparent hotspots	The system may mistake reporting intensity for true risk intensity, distorting prioritisation

Limitations

Language Variability and Report Quality

Deviation reports can be brief, inconsistent, locally worded, or incomplete, which may limit the system's ability to extract reliable risk entities and relationships. Informal abbreviations, site-specific terminology, and vague root-cause narratives may also make it difficult to distinguish confirmed causes from hypotheses. Reviews of adverse drug event extraction from clinical notes show that heterogeneous free text can challenge NLP systems even when the target concepts are well defined [24]. Similar issues would be expected in pharmaceutical quality documents, where report quality strongly affects downstream extraction reliability.

Domain Adaptation and Maintenance

The proposed system would require adaptation to the organisation's product types, manufacturing technologies, terminology, QMS templates, and regulatory scope. Model monitoring, ontology maintenance, and periodic retraining would be necessary because process language, product portfolios, and regulatory expectations change over time. Domain-specific models such as PharmBERT and RxBERT illustrate the value of specialised language adaptation, but they also imply that generic models may not be sufficient for all regulated pharmaceutical contexts [6, 7]. Maintenance should therefore be governed through change control, documented validation, and periodic expert review.

As shown in **Table 4**, effective deployment of the proposed system would require both initial domain adaptation and ongoing maintenance. The system must be aligned with organisation-specific products, technologies, terminology, QMS templates, and regulatory scope, while model monitoring, ontology updates, retraining, change control, validation, and expert review would help ensure continued reliability over time.

Table 4. effective deployment of the proposed system

Area	Adaptation / Maintenance Need	Purpose	Governance Requirement
Product types	Adapt the system to the organisation's specific product portfolio	Ensure risk extraction reflects relevant product characteristics and quality attributes	Document scope and assumptions during validation
Manufacturing technologies	Align models with site-specific processes, equipment, and technologies	Improve recognition of process-related risks and deviations	Review changes through change control
Terminology	Incorporate organisation-specific language, abbreviations, and technical terms	Reduce misclassification caused by generic or unfamiliar wording	Maintain controlled vocabularies or ontologies
QMS templates	Configure extraction logic for local deviation, CAPA, audit, and investigation templates	Improve consistency when processing structured and semi-structured records	Validate template-specific extraction performance
Regulatory scope	Adapt outputs to applicable markets, regulations, and inspection expectations	Support compliance-relevant interpretation of risks	Periodically review against updated regulatory expectations
Model monitoring	Track model performance, errors, drift, and unusual outputs over time	Detect degradation as language, products, or processes change	Define monitoring metrics and escalation criteria
Ontology maintenance	Update risk categories, entities, process terms, and relationships	Keep knowledge structures aligned with current operations	Maintain version control and expert approval
Periodic retraining	Retrain or fine-tune models using updated domain data when needed	Improve performance in changing pharmaceutical contexts	Perform documented validation before deployment
Expert review	Include quality, regulatory, and subject-matter experts in review cycles	Preserve operational context and prevent over-reliance on automated outputs	Schedule periodic review and approval
Change control	Govern major model, ontology, data-source, or workflow changes	Ensure modifications are assessed before implementation	Document impact assessment, testing, and approval

Conclusion

The proposed regulatory text mining system would transform unstructured pharmaceutical quality texts into a structured, queryable, and source-linked quality risk knowledge graph. By combining document ingestion, NLP extraction, risk classification, graph construction, and dashboard-based review, the system could help quality teams interpret dispersed textual evidence more systematically.

Its main strength lies in converting fragmented narratives into connected risk intelligence. Automated extraction could help reveal recurring deviation themes, cross-document risk patterns, audit-relevant evidence, and potential CAPA opportunities while keeping qualified experts responsible for interpretation and approval.

Important challenges would remain. These include variable report quality, site-specific language, the need for domain-specific model adaptation, the governance burden of maintaining validated AI components, and the cultural shift required for quality organisations to trust data-driven risk intelligence.

Collaborative pilots between regulatory affairs, quality assurance, manufacturing science, data integrity teams, and AI specialists would be needed to refine the system in operational settings. Such pilots could establish practical benchmarks, clarify governance expectations, and define how pharmaceutical text mining should support proactive quality oversight without replacing expert judgement.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Lokrantz A, Gustavsson E, Jirstrand M. Root cause analysis of failures and quality deviations in manufacturing using machine learning. Procedia Cirp. 2018;72:1057–62.
2. Salami H, Smith-Goettler B, Yadav V. How can language models assist with pharmaceutical manufacturing deviations and investigations? Int J Pharm. 2025;669:125100.
3. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of natural language processing systems to annotate drug product labeling with MedDRA terminology. J Biomed Inform. 2018;83:73–86.
4. Pandey A, Kreimeyer K, Foster M, Dang O, Ly T, Wang W, et al. Adverse event extraction from structured product labels using the event-based text-mining of health electronic records (ETHER) system. Health Informatics J. 2019;25(4):1232–43.
5. Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, et al. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug Saf. 2021;44(1):83–94.
6. ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Brief Bioinform. 2023;24(4):bbad226.
7. Wu L, Gray M, Dang O, Xu J, Fang H, Tong W. RxBERT: enhancing drug labeling text mining and analysis with AI language modeling. Exp Biol Med. 2023;248(21):1937–43.
8. Wu Y, Liu Z, Wu L, Chen M, Tong W. BERT-based natural language processing of drug labeling documents: a case study for classifying drug-induced liver injury risk. Front Artif Intell. 2021;4:729834.
9. Gray M, Xu J, Tong W, Wu L. Classifying free texts into predefined sections using AI in regulatory documents: a case study with drug labeling documents. Chem Res Toxicol. 2023;36(8):1290–9.
10. Wu L, Xu J, Thakkar S, Gray M, Qu Y, Li D, Tong W. A framework enabling large language models into regulatory environment for transparency and trustworthiness and its application to drug labeling documents. Regul Toxicol Pharmacol. 2024;149:105613.
11. International Council for Harmonisation. ICH Q9(R1): Quality Risk Management. Geneva: ICH; 2023.
12. Dang V, Wu E, Kortepeter CM, Phan M, Zhang R, Ma Y, et al. Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US FDA adverse event reporting system. Front Drug Saf Regul. 2022;2:1020943.
13. Pilipiec P, Liwicki M, Bota A. Using machine learning for pharmacovigilance: a systematic review. Pharmaceutics. 2022;14(2):266.
14. Trajanov D, Trajkovski V, Dimitrieva M, Dobreva J, Jovanovik M, Klemen M, et al. Review of natural language processing in pharmacology. Pharmacol Rev. 2023;75(4):714–38.
15. Joshi P, Masilamani V, Mukherjee A. A knowledge graph embedding based approach to predict adverse drug reactions using a deep neural network. J Biomed Inform. 2022;132:104122.
16. Hauben M, Rafi M, Abdelaziz I, Hassanzadeh O. Knowledge graphs in pharmacovigilance: a scoping review. Clin Ther. 2024;46(7):544–54.

17. Taneja SB, Callahan TJ, Paine MF, Kane-Gill SL, Kilicoglu H, Joachimiak MP, et al. Developing a knowledge graph for pharmacokinetic natural product–drug interactions. J Biomed Inform. 2023;140:104341.
18. Hauben M, Rafi M. Knowledge graphs in pharmacovigilance: a step-by-step guide. Clin Ther. 2024;46(7):538–43.
19. Neyarapally GA, Wu L, Xu J, Zhou EH, Dang O, Lee J, et al. Description and validation of a novel AI tool, LabelComp, for the identification of adverse event changes in FDA labeling. Drug Saf. 2024;47(12):1265–74.
20. 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.
21. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using an NLP system combining knowledge base and deep learning. J Am Med Inform Assoc. 2020;27(1):56–64.
22. Dai X, Karimi S, Sarker A, Hachey B, Paris C. MultiADE: a multi-domain benchmark for adverse drug event extraction. J Biomed Inform. 2024;160:104744.
23. Gonzalez-Hernandez G, Krallinger M, Muñoz M, Rodriguez-Esteban R, Uzuner Ö, Hirschman L. Challenges and opportunities for mining adverse drug reactions: perspectives from pharma, regulatory agencies, healthcare providers and consumers. Database. 2022;2022:baac071.
24. Modi S, Kasmiran KA, Sharef NM, Sharum MY. Extracting adverse drug events from clinical notes: a systematic review of approaches used. J Biomed Inform. 2024;151:104603.
25. Bousdekis A, Lepenioti K, Apostolou D, Mentzas G. Data analytics in quality 4.0: literature review and future research directions. Int J Comput Integr Manuf. 2023;36(5):678–701.
26. Dasgupta S, Jayagopal A, Hong AL, Mariappan R, Rajan V. Adverse drug event prediction using noisy literature-derived knowledge graphs: algorithm development and validation. JMIR Med Inform. 2021;9(10):e32730.
27. Ying L, Liu Z, Fang H, Kusko R, Wu L, Harris S, et al. Text summarization with ChatGPT for drug labeling documents. Drug Discov Today. 2024;29(6):104018.
28. Wu L, Fang H, Qu Y, Xu J, Tong W. Leveraging FDA labeling documents and large language models to enhance annotation, profiling, and classification of drug adverse events with AskFDALabel. Drug Saf. 2025;48(6):655–65.
29. Knisely BM, Hatim Q, Vaughn-Cooke M. Utilizing deep learning for detecting adverse drug events in structured and unstructured regulatory drug datasets. Pharm Med. 2022;36(5):307–17.
30. 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):99–116.