

A Retrieval-Augmented Generation enhanced expert system framework for supply chain risk management: a healthcare industry application

Jun Ren , Lei Wang , Jirapan Liangrokapart & Duangpun Kritchanchai

To cite this article: Jun Ren , Lei Wang , Jirapan Liangrokapart & Duangpun Kritchanchai (19 Oct 2026): A Retrieval-Augmented Generation enhanced expert system framework for supply chain risk management: a healthcare industry application, Enterprise Information Systems, DOI: [10.1080/17517575.2026.2750200](https://doi.org/10.1080/17517575.2026.2750200)

To link to this article: <https://doi.org/10.1080/17517575.2026.2750200>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 19 Oct 2026.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

A Retrieval-Augmented Generation enhanced expert system framework for supply chain risk management: a healthcare industry application

Jun Ren^a, Lei Wang^{a,b}, Jirapan Liangrokapart^c and Duangpun Kritchanhai^c

^aSchool of Engineering and Built Environment, Liverpool John Moores University, Liverpool, UK; ^bAnta Logistics LTD, Anta Logistics Park, Quanzhou, Fujian, China; ^cDepartment of Industrial Engineering, Faculty of Engineering, Mahidol University, Salaya, Thailand

ABSTRACT

Traditional expert systems struggle to manage high-stakes healthcare supply chain risks, such as drug shortages and counterfeiting, since knowledge bases require manual updating. This paper proposes a framework combining Retrieval-Augmented Generation (RAG), fuzzy inference, and Dempster-Shafer evidential reasoning to extract and structure risk knowledge from unstructured documents. A UK and China hospital case study identifies drug shortage and counterfeiting as the highest-priority risks. A small pilot of the RAG component alone shows grounded extraction outperforming a no-retrieval baseline, with no hallucination flagged. To our knowledge, no earlier framework uses RAG for upstream knowledge acquisition in supply chain risk management.

ARTICLE HISTORY

Received 4 March 2026
Accepted 13 October 2026

KEYWORDS

Supply chain risk management; Retrieval-Augmented Generation; expert systems; fuzzy inference; evidential reasoning; healthcare supply chain

1. Introduction

Supply chain disruptions carry serious consequences. A ten-minute fire at a Philips semiconductor plant cost Ericsson \$400 million in lost microchip deliveries (Chopra and Sodhi 2004). The bankruptcy of one UK supplier forced Land Rover to make 1,400 workers redundant (Tang 2006). Healthcare supply chains face even graver stakes: drug shortages can harm patients, counterfeiting can kill them, and operational failures in hospitals ripple through every ward (Enyinda, Gebremikael, and Ogbuehi 2014).

These risks are not abstract. Falsified and substandard medicines entering the supply chain endanger patients directly, prompting regulatory responses such as the EU Falsified Medicines Directive and the UK MHRA's anti-counterfeiting requirements. National shortage-alert systems, such as the NHS Medicine Supply Tool, routinely force clinicians to substitute therapies at short notice, with attendant risks to treatment continuity and patient safety (Breen 2008). Pharmaceutical compliance failures, including breaches of cold-chain and Good Distribution Practice requirements, can trigger product recalls that ripple through hospital pharmacies with little warning.

Healthcare supply chains are more complex and fragmented than most. Hospitals deliver goods and services from many suppliers to patients with wide-ranging health needs, providing both planned and emergency care (El Baz and Ruel 2021; Fenies, Gourgand, and Tchernev 2006). Medicine can pass through 20 to 30 pairs of hands before reaching a patient (Haigh 2004). Supply selections depend on physician preference, which is driven by training, brand familiarity, and clinical context rather than the production forecasts that guide manufacturing (Tucker 2004). These conditions make effective supply chain risk management (SCRM) both urgent and difficult.

Researchers have responded with multi-criteria decision-making (MCDM) methods such as fuzzy AHP, TOPSIS, ISM, combined with expert surveys and interviews (Elleuch, Hachicha, and Chabchoub 2014; Samvedi, Jain, and Chan 2013; Wang et al. 2023). These methods work. But they depend on labour-intensive knowledge elicitation: distributing hundreds of questionnaires, conducting dozens of interviews, manually coding every risk factor and pairwise comparison. Wang et al. (2023) distributed 431 questionnaires to obtain 56 valid responses for a single healthcare SCRM study. Recent bibliometric work on supply chain risk and healthcare resilience (Raut, Hossain, and Fazio 2025; Thompson et al. 2025) shows how fast this

literature is growing, which makes manual synthesis harder still. This knowledge acquisition bottleneck, including the difficulty of extracting, structuring, and encoding domain expertise, has constrained expert systems since their inception (Feigenbaum 1977).

Recent advances in artificial intelligence offer a potential solution. Large language models (LLMs) can process unstructured text at scale, extracting information from documents, reports, and guidelines that would take human analysts weeks to review. Retrieval-Augmented Generation (RAG), introduced by Lewis et al. (2020), grounds LLM outputs in retrieved source documents, suppressing hallucination and anchoring responses to verifiable evidence. RAG has proven effective in legal research, medical question-answering, and customer support. There are domains in which factual accuracy and source traceability are essential.

Yet LLMs and RAG alone cannot replace expert systems for risk management. They cannot quantify uncertainty in a principled way. They cannot combine conflicting evidence from multiple experts using mathematically grounded fusion rules. Their reasoning is opaque and non-reproducible. For safety-critical decisions, formal reasoning methods remain essential. This applies to offshore operations, healthcare, and other domains in which lives and livelihoods depend on risk assessments. This need for formal, auditable reasoning in safety-critical contexts is further underscored by emerging AI requirements frameworks (Klüver 2024).

This paper proposes a framework that combines both capabilities. The RAG layer handles knowledge acquisition: it ingests heterogeneous unstructured sources (regulatory documents, operational records, expert interview transcripts) and extracts structured evidence automatically. The expert system core handles formal reasoning: fuzzy inference quantifies vague assessments, and Dempster–Shafer evidential reasoning combines conflicting multi-source evidence under uncertainty. The RAG layer feeds the expert system; the expert system produces auditable, mathematically grounded risk assessments.

The framework builds on prior work by the authors. Ren et al. (2005) developed a fuzzy-evidential expert system for offshore safety assessment, using 245 IF-THEN rules, a Mamdani fuzzy inference engine, and Dempster–Shafer synthesis across five experts and three input dimensions. Ren, Jenkinson, and Tobora (2026) subsequently demonstrated that a RAG layer could enhance the explainability of that system by generating auditable Chain of Reasoning narratives from its opaque rule outputs. The present paper extends this line of work in a different direction: rather than using RAG to explain expert system outputs (downstream), we use RAG to acquire knowledge for expert system inputs (upstream). The framework is generalised beyond maritime safety and demonstrated through a healthcare supply chain risk management case study.

The research objectives are:

- To develop a general RAG-enhanced expert system framework for SCRM that overcomes the knowledge acquisition bottleneck.
- To formalise the integration between RAG-based knowledge extraction and fuzzy-evidential reasoning.
- To demonstrate the framework through a healthcare supply chain risk management case study using empirical data from UK and China hospitals.

The remainder of this paper is structured as follows. Section 2 reviews the literature on SCRM, expert systems, and RAG. Section 3 presents the proposed framework. Section 4 demonstrates the framework through the healthcare case study, including both risk assessment and mitigation strategy evaluation. Section 5 discusses contributions, implications, and limitations. Section 6 concludes.

2. Literature review

2.1. Supply chain risk management

Supply chain risk management has risen rapidly in both academic and practitioner interest over the past two decades (Bak et al. 2022). The field gained momentum after major disruptions – the 9/11 attacks disrupted global supply chains (Chopra and Sodhi 2004; Sheffi 2005), and subsequent events from natural disasters to pandemics reinforced the need for systematic risk management.

Jüttner, Peck, and Christopher (2003) defined SCRM as ‘the specification and management of risks for the supply chain, through a co-ordinated approach amongst supply chain members, to reduce supply chain vulnerability as a whole’. Tang (2006) extended this to include business continuity. Ho et al. (2015) proposed a comprehensive definition: SCRM is an inter-organisational collaborative endeavour using quantitative and qualitative methodologies to identify, evaluate, mitigate, and monitor events that might adversely affect any part of a supply chain.

Kleindorfer and Saad (2005) identified three key tasks for effective SCRM: risk identification, risk assessment, and risk mitigation. Researchers have built risk management frameworks on these three processes (Bandaly et al. 2012; Ghadge et al. 2013; Ritchie and Brindley 2007).

Healthcare SCRM presents distinct challenges. Hospital operations deliver goods and services from many suppliers to patients with diverse needs (Fenies, Gourgand, and Tchernev 2006). Items are often high-value and require special handling. Zero-defect conditions are necessary. Breen (2008) investigated 35 prevalent risks in the UK NHS pharmaceutical supply chain, finding supply chain fragmentation, lack of stock visibility, and unexpected demand increases as the highest-rated risks. Wang et al. (2023) identified 34 risk factors across three categories. These included factors internal to the hospital, external to the hospital but internal to the supply chain network, and external to the supply chain network. They then applied fuzzy AHP, ISM, and fuzzy TOPSIS to prioritise these risks and evaluate mitigation strategies using empirical data from hospitals in the UK and China.

A common limitation in existing healthcare SCRM research is reliance on manual methods for risk identification and assessment: literature reviews, questionnaire surveys, and expert interviews conducted one study at a time (Breen 2008; Enyinda, Gebremikael, and Ogbuehi 2014; Kanyoma et al. 2013). These methods do not scale. As the volume of relevant documents grows, manual approaches cannot keep pace. This includes regulatory updates, incident reports, and new clinical guidelines.

Beyond risk identification, healthcare organisations have increasingly turned to technology-enabled approaches to manage supply chain risk directly. Data standardisation initiatives such as GS1 coding aim to give hospitals a single, unambiguous product identifier across suppliers, directly addressing the information fragmentation Kreysa and Denecker (2009) documented in the NHS, where a single product could carry 130 different descriptions across trusts. National reviews such as Lord Carter’s (2016) report on NHS acute hospitals recommended advanced information technology, eProcurement, and inventory standardisation as the highest-priority interventions, a finding echoed in the mitigation-strategy rankings reported in Section 4.7. The COVID-19 pandemic further accelerated adoption of these approaches: El Baz and Ruel (2021) found that supply chain risk management practices, including visibility and flexibility measures, materially reduced disruption impacts during the pandemic. These developments indicate a shift towards data-driven, technology-enabled healthcare SCRM, but one that, as Section 2.4 argues, has not yet been coupled with formal reasoning under uncertainty.

Recent studies reinforce both the scale of the field and its direction of travel. Thompson et al. (2025) mapped 3,427 supply chain risk articles published between 2018 and 2024 through bibliometric and network analysis, comparing the immediate pre- and post-COVID eras, and Raut, Hossain, and Fazio (2025) did the same for 3,060 healthcare resilience papers (1997–2023) indexed in Scopus. Both studies characterise the research landscape (actors, themes, clusters and collaboration networks) rather than automating the acquisition of the underlying risk knowledge. On the technology side, Liang, Jiang, and Li (2026) proposed a multimodal intelligent hyperautomation framework for sustainable healthcare supply chains that integrates blockchain, the Internet of Medical Things (IoMT) and AI to fuse medical images, biosensor readings, clinical text and transactional records, and identified the absence of integrated AI reasoning across data modalities as a key bottleneck. These contributions point to the same opening that this paper addresses: data-driven healthcare supply chain capabilities are maturing quickly, but they have not yet been coupled with an auditable mechanism for turning unstructured documents into structured evidence for formal reasoning under uncertainty.

2.2. Expert systems and reasoning under uncertainty

Expert systems encode domain knowledge as rules and apply inference mechanisms to produce decisions. They have been applied in medical diagnosis, financial planning, engineering design, and risk assessment

since the 1970s (Feigenbaum 1977). A typical architecture consists of a knowledge base (database and rule base), an inference engine, and a user interface.

Two mathematical frameworks are central to expert systems for risk management under uncertainty: fuzzy set theory and evidential reasoning.

Fuzzy set theory, proposed by Zadeh (1965), handles vagueness in human judgement. Classical set theory assigns elements fully in or fully out of a set. Fuzzy sets allow partial membership, expressed through membership functions that map elements to values between 0 and 1. Linguistic variables like ‘high risk’, ‘moderate importance’, and ‘very frequent’ are used by experts in practice, where these terms correspond to specific fuzzy numbers for clear expression of opinions. This approach helps them convey nuanced meanings through numerical representation within their specialised field. Triangular fuzzy numbers (TFNs), defined by a triple (l, m, u), are widely used for their computational simplicity (D. Y. Chang 1996). A fuzzy inference engine applies IF-THEN rules to fuzzified inputs and produces fuzzy outputs, which can be defuzzified into crisp values when needed.

Evidential reasoning, based on Dempster–Shafer (D-S) theory (Dempster 1968; Shafer 1976), handles uncertainty in combining evidence from multiple sources. Unlike Bayesian probability, D-S theory can represent ignorance explicitly through a mass function assigned to the full frame of discernment. J. B. Yang (2001) and J. B. Yang and Xu (2002) developed the evidential reasoning algorithm for multi-attribute decision analysis under uncertainty, enabling the systematic combination of assessments from multiple experts across multiple criteria. L. H. Yang et al. (2022) subsequently advanced this line by proposing a highly explainable cumulative belief rule-based system with interpretable rule weighting, further strengthening the transparency of evidential reasoning outputs.

The combination of fuzzy inference and evidential reasoning is well-established. Ren et al. (2005) built an offshore safety framework using four modules: a fuzzy mapping module, a knowledge base with 245 IF-THEN rules, a fuzzy inference engine, and a Dempster–Shafer synthesis module. Ren et al. (2009) subsequently extended this line of work using fuzzy Bayesian networks for offshore risk quantification. Liu et al. (2005) applied fuzzy rule-based evidential reasoning to engineering system safety analysis. These systems are effective but share a common bottleneck: all knowledge must be manually elicited from domain experts and encoded into rules and databases. As domains grow more complex and document volumes increase, this bottleneck becomes the binding constraint.

2.3. Retrieval-Augmented Generation

Large language models (LLMs) have demonstrated strong capabilities in text comprehension, information extraction, and structured output generation. However, they suffer from well-documented limitations: hallucination (generating plausible but false information), lack of source traceability, and inability to update knowledge without retraining (Lewis et al. 2020).

These limitations are consequential for risk management applications. Hallucination is not confined to obscure queries: LLMs can generate fluent, specific-sounding risk assessments, severity ratings, or citations that have no basis in any source document, and the errors are often plausible enough to escape casual review. A parametric knowledge cut-off means an LLM’s unaided responses cannot reflect regulatory updates, incident reports, or guidance published after training, a material limitation in a domain where standards and shortage alerts change frequently. Without explicit grounding, LLM outputs also carry no audit trail: a risk manager cannot trace a stated severity level back to the document or expert opinion that supports it, which is unacceptable for safety-critical decisions and regulatory scrutiny. Finally, because LLM outputs are open-ended text rather than structured, bounded predictions, evaluating their correctness at scale is itself an unsolved problem, complicating any attempt to benchmark LLM-only risk assessment against established methods.

Retrieval-Augmented Generation (RAG) addresses these limitations. Introduced by Lewis et al. (2020), RAG combines parametric LLM knowledge with non-parametric retrieval from an external knowledge store. In a standard RAG pipeline, a query is encoded into a vector embedding, matched against a pre-indexed document store via semantic similarity, and the retrieved context is injected into the LLM prompt to constrain and ground the response. This architecture suppresses hallucination and makes outputs traceable to source documents.

RAG has proven effective in knowledge-intensive tasks where factual accuracy matters: legal document analysis, clinical question-answering, and technical support (Lewis et al. 2020). Ren, Jenkinson, and Tobora (2026) demonstrated a specific application in safety-critical systems: using metadata-filtered RAG to generate auditable explanations for an offshore safety expert system's decisions. The original study employed the RAG in a subsequent phase, using it as an explanatory tool for interpreting results and findings; conversely, this current research harnesses the same RAG mechanism but utilises it at an earlier stage. This upstream use of RAG is specifically aimed to gather relevant knowledge that will serve as inputs throughout our investigative process.

2.4. Research gap and contribution

The literature reveals three gaps:

- (1) **To our knowledge, limited or no integration of RAG with formal expert system reasoning for SCRM.** Existing SCRM frameworks rely on manual knowledge acquisition. Existing RAG applications focus on question-answering, not on feeding structured evidence into formal reasoning systems.
- (2) **LLM-only approaches lack formal reasoning.** Using an LLM directly for risk assessment cannot provide the uncertainty quantification, evidence fusion, or mathematical auditability that fuzzy inference and evidential reasoning offer.
- (3) **Expert-system-only approaches lack scalable knowledge acquisition.** Traditional expert systems require all knowledge to be manually elicited and encoded. This process does not scale to the growing volume of relevant documents, data, and expert opinions.

This paper fills these gaps by proposing a framework that combines RAG for scalable knowledge acquisition with fuzzy inference and evidential reasoning for formal risk assessment. The framework is domain-independent: changing the input document corpus adapts it to new sectors without modifying the reasoning architecture.

3. The proposed framework

3.1. Framework overview

The proposed framework integrates Retrieval-Augmented Generation (RAG) with a fuzzy-evidential expert system for supply chain risk management. Figure 1 illustrates the architecture.

The framework has six layers:

- **Input Layer** - heterogeneous domain knowledge sources;
- **RAG Layer** - retriever, vector database, and grounded LLM for knowledge extraction;
- **Structured Evidence Extractor** - bridge between RAG outputs and expert system inputs;
- **Knowledge Base and Fuzzification** - database, rule base, and fuzzy conversion;
- **Fuzzy Inference Engine** - rule evaluation and risk computation;
- **Synthesis Module** - Dempster–Shafer evidential reasoning for multi-source, multi-expert aggregation.

The design philosophy is simple: RAG handles knowledge acquisition; the expert system handles formal reasoning. Each does what it does best.

This framework builds on two prior contributions. Ren et al. (2005) established the fuzzy-evidential expert system architecture for offshore safety, comprising a knowledge base, fuzzy inference engine, and Dempster–Shafer synthesis module. Ren, Jenkinson, and Tobora (2026) demonstrated that RAG could enhance that system's explainability by generating auditable narratives from its outputs. The present framework extends this work in a fundamentally different direction: rather than using RAG to explain expert system outputs, it uses RAG to acquire and structure the knowledge that feeds expert system inputs. The reasoning modules are retained; the knowledge acquisition process is transformed.

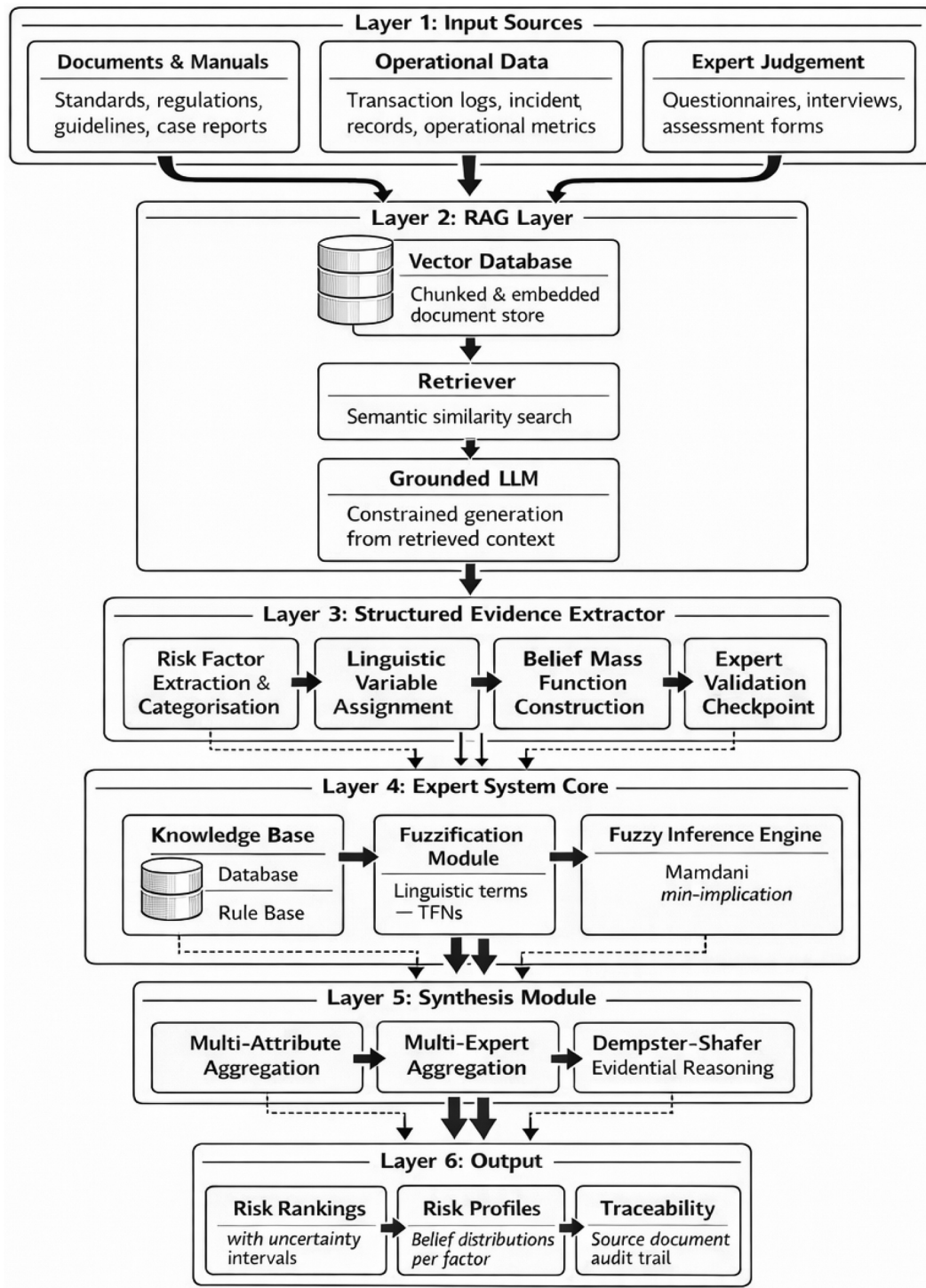


Figure 1. General RAG-Enhanced Expert System Framework.

3.2. Input layer

The framework accepts three categories of input:

- **Documents and Manuals:** domain-specific standards, regulations, guidelines, audit reports, published research, and case studies. In healthcare SCRM, these include NHS procurement guidelines, WHO pharmaceutical standards, FDA regulations, and published risk management studies.

- **Operational Data:** quantitative and semi-quantitative records from operations. In healthcare SCRM: inventory logs, delivery records, drug shortage databases, procurement transaction data, and incident reports.
- **Expert Judgement:** qualitative assessments from domain practitioners, captured through structured questionnaires, semi-structured interviews, or assessment forms. In healthcare SCRM: judgements from hospital procurement managers, supply chain managers, pharmacists, and clinical staff.

These three input types are heterogeneous and largely unstructured. Converting them into the structured formats required by an expert system, such as fuzzy membership values, pairwise comparison matrices, and belief mass functions, represents the knowledge acquisition bottleneck that the RAG layer addresses.

3.3. Rag layer

The RAG layer consists of three components that work sequentially to extract and structure knowledge from the input sources.

3.3.1. Vector database

All input documents are preprocessed, chunked, and embedded into a vector database. The chunking strategy segments documents at the paragraph or section level, preserving semantic coherence within each chunk. An embedding model (such as a sentence transformer) converts each chunk into a dense vector representation. The vector database stores these embeddings alongside metadata: source document, section title, document type, and date.

Operational data descriptions and expert judgement transcripts are embedded alongside formal documents, creating a unified searchable knowledge store across all input types.

For reproducibility, the framework specifies a reference configuration rather than leaving these choices open-ended. Embeddings are generated with a general-purpose sentence-embedding model selected for strong retrieval performance on public benchmarks (e.g. an OpenAI text-embedding-3-class model, or an open-weight alternative such as BAAI/bge-large-en for on-premises deployment, a relevant constraint given healthcare data-governance requirements). Chunks target 300–500 tokens with 15–20% overlap between consecutive chunks, and chunk boundaries are aligned to paragraph or section breaks rather than fixed character counts, so that a single risk-relevant statement is not split across two chunks.

The vector store is a dedicated similarity-search index (e.g. FAISS for on-premises deployment, or a managed store such as Pinecone or Weaviate for production scale) using approximate nearest-neighbour search (HNSW), which supports sub-second retrieval at the corpus sizes typical of a single-domain document collection (thousands to low millions of chunks).

3.3.2. Retriever

Given a structured query, for example, ‘What are the primary risk factors affecting drug availability in hospital supply chains?’, the retriever performs a semantic similarity search against the vector database. It returns the top-k most relevant document chunks, ranked by cosine similarity to the query embedding.

The query formulation follows a risk-oriented strategy aligned with the SCRM process: queries target risk identification (‘What risks exist?’), risk characterisation (‘How severe is this risk?’), risk relationships (‘How do these risks interact?’), and mitigation (‘What strategies address this risk?’).

The reference configuration retrieves the top-k = 5–10 chunks by cosine similarity, then applies a cross-encoder reranker to reorder an initial candidate pool of approximately four times that size; scoring each candidate jointly against the query, rather than independently, improves precision over similarity search alone. Metadata filters constrain retrieval by document type, jurisdiction, and recency (e.g. excluding superseded regulatory guidance), so that outdated or out-of-scope evidence is not surfaced even when it is semantically similar to the query.

3.3.3. Grounded LLM

The grounded LLM receives the retrieved document chunks as context and generates structured outputs. Three design constraints ensure reliability:

- **Grounding:** the LLM is instructed to use only information from the retrieved context, not its parametric knowledge. This suppresses hallucination.
- **Structured output:** the LLM produces outputs in predefined formats. These include risk factor tables, severity assessments expressed in linguistic terms, and categorised evidence rather than free-form text.
- **Source citation:** every extracted fact must cite the source document chunk from which it was derived, enabling traceability.

The LLM does not make risk decisions. It extracts and structures information. All risk assessment decisions are made by the expert system core.

The reference configuration uses a general-purpose instruction-tuned LLM selected for instruction-following fidelity and a context window sufficient to hold the retrieved chunk set (e.g. a GPT-4-class model, or an open-weight equivalent such as Llama-3-70B-Instruct where data-residency requirements preclude an external API). The prompting strategy combines a fixed instruction template, few-shot examples of correctly grounded extractions, and an explicit output schema (structured JSON with fields for risk factor, category, severity, source chunk ID, and a verbatim supporting quotation), so that outputs are machine-checkable rather than free text.

Grounding is enforced at three points: the prompt instructs the model to answer only from the retrieved context and to output ‘insufficient evidence’ rather than guess when the context does not support a claim; a post-hoc citation check verifies that every extracted claim’s cited chunk ID exists in the retrieved set and that the supporting quotation is substantively present in that chunk’s text; and claims that fail this check are flagged and routed to the expert validation checkpoint (Section 3.4) rather than entering the knowledge base automatically.

3.4. Structured evidence extractor

The Structured Evidence Extractor bridges the RAG layer and the expert system core. It performs four functions:

Risk factor extraction and categorisation. The extractor parses the LLM’s structured output to identify distinct risk factors and assign them to predefined categories. In SCRM, the typical categorisation follows Christopher and Peck (2004): risks internal to the organisation, risks external to the organisation but internal to the supply chain network, and risks external to the supply chain network.

Linguistic variable assignment. Qualitative descriptions from the LLM output, such as ‘this risk is considered highly significant by procurement managers’, are mapped to linguistic terms from a predefined scale (e.g. Very Low, Low, Medium, High, Very High). These linguistic terms correspond to triangular fuzzy numbers used by the fuzzification module.

Belief mass function construction. For evidential reasoning, the extractor converts evidence strength into basic probability assignments (BPAs). Strong, corroborated evidence from multiple sources receives a high mass assignment to a specific risk level. Weak or ambiguous evidence assigns mass to the full frame of discernment, representing uncertainty. This explicit modelling of ignorance is a key advantage of Dempster–Shafer theory over purely probabilistic approaches.

Expert validation checkpoint. The extracted and structured evidence is presented to domain experts for review before entering the knowledge base. This human-in-the-loop step ensures that RAG-extracted knowledge meets domain accuracy standards. It is less burdensome than full manual knowledge elicitation because experts review and correct structured outputs rather than building knowledge from scratch.

3.5. Knowledge base and fuzzification module

3.5.1. Knowledge base

The knowledge base consists of two components:

- **Database:** stores validated risk factors, their categories, severity assessments, interrelationships, and historical data. The RAG layer and Structured Evidence Extractor populate this database; experts validate it.

- **Rule Base:** contains IF-THEN fuzzy rules encoding domain expertise. For example:
- IF *risk_likelihood* is HIGH AND *consequence_severity* is CATASTROPHIC, THEN *risk_level* is VERY HIGH
- IF *risk_likelihood* is LOW AND *consequence_severity* is MINOR, THEN *risk_level* is LOW
- IF *supply_disruption_risk* is HIGH AND *IT_system_quality* is POOR, THEN *operational_failure_risk* is VERY HIGH

Rules can be constructed from domain standards, expert input, or extracted from the structured evidence produced by the RAG layer and validated by experts.

3.5.2. Fuzzification module

The fuzzification module converts crisp values and linguistic assessments into fuzzy membership values using triangular fuzzy numbers (TFNs). The linguistic scale and corresponding TFNs used in this framework follow established conventions (Table 1):

Expert assessments expressed in linguistic terms, whether from survey responses or RAG-extracted evidence, are converted to TFNs through this mapping. The fuzzification module then feeds these values into the fuzzy inference engine.

3.6. Fuzzy inference engine

The fuzzy inference engine applies the rule base to fuzzified inputs to compute risk levels for each factor. The engine uses the Mamdani min-implication operator: for each rule, the firing strength is the minimum of the membership values of the rule's antecedents, and the output fuzzy set is clipped at that strength. When multiple rules fire, their outputs are aggregated using the maximum operator.

For defuzzification – converting fuzzy outputs to crisp risk scores – the centroid method (Eq. 1) is applied:

$$\hat{z} = \frac{\int^z \cdot \mu(z) dz}{\int^{\mu}(z) dz} \quad (1)$$

where $\hat{z}$ is the defuzzified output, z is the output variable, and $\mu(z)$ is the aggregated membership function.

The inference engine produces a risk score for each factor assessed individually. Combining assessments across multiple factors and multiple experts requires the synthesis module.

3.7. Synthesis module: evidential reasoning

The synthesis module combines evidence from multiple sources and multiple experts using Dempster–Shafer theory. This module performs two types of aggregation:

Multi-attribute aggregation combines risk assessments across different factors into an overall risk profile. **Multi-expert aggregation** combines assessments from different experts for the same risk factor.

3.7.1. Notation

- Θ – the frame of discernment: the exhaustive set of mutually exclusive risk levels (Eq. 2).
- $m(\cdot)$ – a basic probability assignment (BPA): a mass function distributing belief across subsets of Θ .
- $m1 \oplus m2$ – the combination of two independent BPAs under Dempster's rule (Eq. 3).
- K – the conflict factor: the total mass assigned to the empty set before normalisation.

Table 1. Linguistic scale and triangular fuzzy numbers for risk assessment.

Linguistic term	TFN (l, m, u)
Very Low (VL)	(0, 0, 0.25)
Low (L)	(0, 0.25, 0.5)
Medium (M)	(0.25, 0.5, 0.75)
High (H)	(0.5, 0.75, 1.0)
Very High (VH)	(0.75, 1.0, 1.0)

- $Bel(A)$, $Pl(A)$ – the belief and plausibility of subset $A \subseteq \Theta$, providing lower and upper probability bounds (Eqs. 4-5).
- $[Bel(A), Pl(A)]$ – the belief interval, quantifying residual uncertainty for A .

The remainder of this subsection defines the frame of discernment, Dempster’s rule, and the belief/plausibility measures in turn, using this notation throughout.

3.7.2. Frame of discernment

For risk assessment, the frame of discernment Θ defines the set of possible risk levels:

$$\Theta = \{\text{VeryLow}, \text{Low}, \text{Medium}, \text{High}, \text{VeryHigh}\} \quad (2)$$

Each expert’s assessment of a risk factor is expressed as a basic probability assignment (BPA) – a mass function m that distributes belief across subsets of Θ . For example, if an expert assesses a risk factor as ‘High’ with strong confidence, the BPA might be:

$$m(\{\text{High}\}) = 0.7, m(\{\text{High}, \text{VeryHigh}\}) = 0.2, m(\Theta) = 0.1$$

The mass assigned to Θ represents the expert’s residual uncertainty – what they do not know. This ability to model ignorance explicitly distinguishes D-S theory from Bayesian probability.

3.7.3. Dempster’s rule of combination

Given two independent BPAs m_1 and m_2 , Dempster’s rule combines them:

$$m_{1,2}(A) = \frac{1}{1 - K} \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \quad (3)$$

where $K = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C)$ is the conflict factor. The normalization by $(1K)$ redistributes conflicting mass proportionally. This rule is applied iteratively to combine evidence from all experts.

3.7.4. Belief and plausibility

For each risk level A , the belief $Bel(A)$ and plausibility $Pl(A)$ provide lower and upper bounds on the true probability:

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad (4)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad (5)$$

The interval $[Bel(A), Pl(A)]$ quantifies the range of uncertainty for each risk level. Decision-makers can use the maximum belief, maximum plausibility, or a weighted combination to rank risk factors.

3.8. Output: risk estimation

The framework produces three types of output:

- **Risk rankings:** each risk factor ranked by its combined belief-plausibility assessment, with uncertainty intervals.
- **Risk profiles:** the full belief distribution across risk levels for each factor, showing where assessments cluster and where uncertainty remains.
- **Traceability:** every output links back through the synthesis module, inference engine, knowledge base, and structured evidence extractor to the original source documents retrieved by the RAG layer. This end-to-end traceability supports audit and regulatory compliance.

Figure 2 summarises this data flow end to end, from the heterogeneous input sources through the RAG layer’s extraction process, the structured evidence extractor’s expert validation checkpoint, and the fuzzy inference and evidential synthesis modules that produce the final risk estimate.

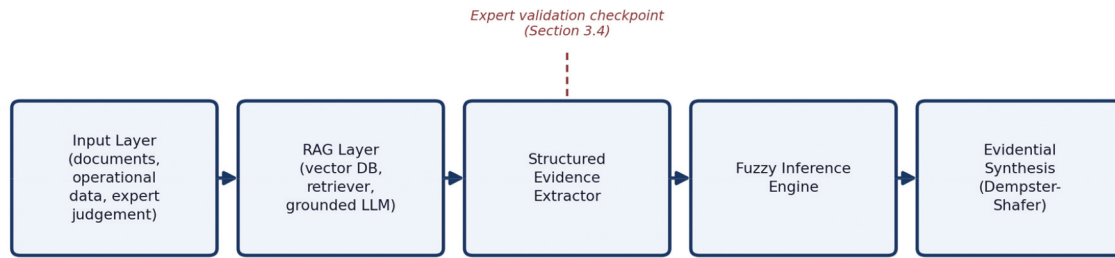


Figure 2. Data flow from RAG extraction to fuzzy inference and evidential synthesis.

4. Case study: healthcare supply chain risk management

4.1. Case background

This section demonstrates the proposed framework through a healthcare supply chain risk management case study, using empirical data from Wang et al. (2023). That study investigated SCRM in UK and China hospitals by surveying 56 experts. The participants included hospital procurement managers, supply chain managers, pharmacists, and academics, and semi-structured interviews were conducted to identify, assess, and prioritise supply chain risks.

From an initial set of 34 risk factors, Wang et al. identified 11 priority risk factors requiring the most attention, grouped into three categories following Christopher and Peck (2004). The three categories include internal to the hospital (e.g. process and control risks), external to the hospital but internal to the supply chain network (e.g. supply and demand risks), and external to the supply chain network (e.g. environmental risks). Table 2 presents these factors.

Wang et al. also identified nine risk mitigation strategies through empirical studies in UK and China healthcare organisations. Table 3 presents these strategies.

4.1.1. Expert validation of the underlying empirical data

The risk factors and mitigation strategies used in this case study were not adopted from secondary literature alone; they were empirically validated through a structured, multi-stage expert elicitation process reported in full in the underlying doctoral research (Wang 2018) that produced Wang et al. (2023), summarised here because it directly bears on the reliability of the data the fuzzy inference and evidential reasoning modules operate on in Sections 4.4–4.6.

Risk factor identification and the three-tier risk taxonomy (Table 2) were validated by a six-member expert panel spanning UK and China hospital procurement leadership, pharmacy science academia, and supply chain risk management, weighted by professional relevance (20–30% for senior practitioners and academics with the most directly relevant experience, down to 10% for adjacent-domain specialists). The questionnaire instrument underwent content-validity review by two academic researchers and two specialists, a pilot study, and ethical approval prior to data collection. Across 204 survey items, the instrument achieved a Cronbach’s alpha of 0.944 (0.952 on standardised items), indicating high internal consistency (Cohen and Swerdlik 2010; Sijtsma 2009). The resulting hierarchical risk taxonomy was subsequently re-validated

Table 2. Priority risk factors in healthcare supply chain management.

Code	Risk factor	Category
S1	Counterfeiting	External to hospital, internal to SC network
S2	Poor quality in purchased drugs from suppliers	External to hospital, internal to SC network
S4	Shortage of drugs, unavailability of drugs on the market	External to hospital, internal to SC network
S5	Capability versus demand, inability of capacity to meet demand	External to hospital, internal to SC network
S8	Poor IT system, lack of data standardisation	Internal to hospital
S9	Dispensing/picking errors in medication/packaging	Internal to hospital
S10	Weak logistics service infrastructure	Internal to hospital
S11	Lack of visibility concerning placement and availability of stock	Internal to hospital
S13	Clinician’s preference	Internal to hospital
S14	High purchase price	External to hospital, internal to SC network
S15	High product and supplier/brand variety	External to hospital, internal to SC network

Table 3. Risk mitigation strategies in healthcare supply chain management.

Code	Strategy	Description
A1	Cold chain management	Monitoring and controlling environmental conditions to maximize product freshness across distribution
A2	Advanced information technology and systems	Data standardization, information sharing, drugs traceability, RFID-enabled systems
A3	Inter- and intra-organizational collaboration	Vertical and horizontal cooperation with all supply chain members
A4	Inventory management	Just-in-time approaches, direct delivery to point-of-use, reducing holding costs
A5	eProcurement strategy	Technology-enabled automation of procurement information exchange
A6	Outsourcing strategy	Outsourcing non-core supply chain activities to third-party logistics providers
A7	Agility strategy	Reactive and proactive strategies for turbulent, unpredictable environments
A8	Internal capability management	Staff training on supply chain risk sources and SCRM tactics
A9	Single sourcing strategy	Selecting a single main supplier despite available alternatives

with a separate validation team of six additional experts and academics not involved in the original panel, before being accepted.

Risk assessment drew on 56 valid survey responses (a 13% return rate from 431 questionnaires distributed to 231 UK NHS hospitals, 11 Chinese hospitals, seven pharmaceutical distributors, and two supply chain consulting organisations), of whom 55% had more than 16 years of healthcare supply chain experience. A separate panel of 12 industry and academic experts, spanning hospital stock and pharmacy management, pharmaceutical manufacturing, and supply chain consulting across China, Thailand, and the UK, assessed pairwise risk interdependencies for the interpretive structural modelling analysis; the resulting fuzzy AHP weights were further checked for robustness through a sensitivity analysis that perturbed pairwise-comparison weights for 18 sub-criteria, confirming that the top-ranked risk factors were stable under perturbation.

Mitigation strategy evaluation (Section 4.7) drew on a separate panel of 20 experts from academia and industry who rated strategy effectiveness for the Fuzzy TOPSIS analysis, consistent with the sensitivity analysis reported in Section 4.7 confirming that strategies A2 and A5 retained top-two rankings under the large majority of weight-perturbation conditions tested.

This expert validation establishes the reliability of the empirical risk factors, severity assessments, and mitigation-strategy ratings that anchor the case study. It does not, on its own, constitute a validation of the RAG layer's own extraction outputs against expert judgement; that remains a distinct evaluation task, identified as future work in Section 5.4. The same distinction applies one layer downstream: this validation confirms the reliability of the inputs the fuzzy inference and Dempster-Shafer synthesis modules operate on, not the reliability of what those modules compute from them. No expert panel in this study was shown the framework's own output, the fuzzy scores, belief intervals, and risk rankings in Table 7, or the strategy rankings in Table 9, and asked to independently confirm agreement with them; the reasoning method is an established, transparent mathematical formalism rather than a learned or opaque one, which is a reasonable basis for confidence in the transformation, but it is not the same as empirical confirmation that the framework's specific recommendations match expert judgement. That confirmation, like the RAG layer's, remains future work.

4.2. Input layer application

For this case study, the input layer comprises:

- **Documents and Manuals:** NHS procurement guidelines, WHO pharmaceutical supply chain standards, published healthcare SCRM literature (including Breen 2008; Lord Carter 2016; Wang et al. 2023), supplier audit reports, and regulatory guidance from the UK Medicines and Healthcare products Regulatory Agency (MHRA). Section 4.8 reports a pilot evaluation executed against a reduced 18-document subset of this scope, comprising published SCRM/healthcare literature and the Lord Carter (2016) report; NHS/WHO/MHRA primary regulatory guidance and supplier audit reports were not included in the executed pilot and remain scoped for full-scale future validation (Section 6).
- **Operational Data:** hospital inventory records, NHS drug shortage alerts, procurement transaction logs, and delivery performance data.

- **Expert Judgement:** the empirical findings from the 56 expert survey responses and semi-structured interviews reported in Wang (2018) and Wang et al. (2023), covering procurement managers, supply chain managers, pharmacists, pharmaceutical company managers, and academic researchers. The RAG pilot corpus (Section 4.8) draws on the thesis's results and discussion chapters, which report and interpret these responses in prose; the original raw interview transcripts are not used and are not part of the corpus, consistent with participant confidentiality in the underlying study.

4.3. Rag layer application: a worked example

This section demonstrates the RAG layer through a worked example, showing how unstructured healthcare supply chain documents are transformed into structured evidence for the expert system.

To avoid ambiguity about what this section demonstrates: the risk factors, categories, and severity assessments used in the fuzzy inference and evidential reasoning case study (Sections 4.4-4.6) are the empirically validated data reported in Wang (2018) and Wang et al. (2023) and summarised in Section 4.1.1; the fuzzy scores and belief intervals in Tables 6 and 7 are computed from that panel data, not generated by the RAG layer. The retrieval results (Table 4) and structured LLM output (Table 5) in this section are drawn from an executed pilot run of the RAG layer, reported in full in Section 4.8: a real retriever (BAAI/bge-large-en embeddings, FAISS similarity search, cross-encoder reranking) queried a real 18-document corpus, and a real grounded LLM (Claude Sonnet 4.5) produced the structured output shown, with citation-checked grounding. This is a bounded pilot, not the full corpus scope described in Section 4.2; Section 4.8 reports its scale, methodology, and quantitative results in full, including comparison against a no-retrieval baseline and hallucination-rate measurement.

Table 4. Sample retrieval results for risk factor identification query.

Rank	Source	Type	Retrieved content (excerpt)
1	Wang (2018) thesis	PhD thesis	"Table 5.2: Summary of the identified healthcare supply chain risk factors ... Counterfeiting, Grey-market activity (S1); Poor quality in the purchased drugs from suppliers (S2); Time limit of drugs, drugs perishability (S3); Shortage of drugs, unavailability of drugs on the market (S4) ..."
2	Wang (2018) thesis	PhD thesis	"... Poor IT system, lack of data standardization (S8) 4.4; Dispensing/picking errors-medication /packaging (S9) 4.5; Weak logistics service infrastructure (S10) 4.4; Lack of visibility concerning placement and availability of stock (S11) 4 ..."
3	Wang (2018) thesis	PhD thesis	"... 'counterfeiting' (0.159), 'poor IT system' (0.112), 'capability versus demand' (0.108), 'dispensing/picking errors' (0.097) ... are the key risk factors for hospital's RQ4 ..."

Table 5. Sample grounded LLM output – extracted risk factors.

Risk Factor	Category	Severity (Linguistic)	Source
Counterfeiting and grey-market activity	Supply risks – Quality risks	Very High	Wang (2018), chunk 499
Poor quality in purchased drugs from suppliers	Supply risks – Quality risks	Very High	Wang (2018), chunk 329
Shortage of drugs, unavailability of drugs on the market	Supply risks – Supplier risks	Very High	Wang (2018), chunk 329
Dispensing/picking errors – medication/packaging	Process risks	Very High	Wang (2018), chunk 329
Poor IT system, lack of data standardisation	Process risks	High	Wang (2018), chunk 329
Weak logistics service infrastructure	Process risks	High	Wang (2018), chunk 329
Wrong demand forecasting	Demand risks – Forecast risks	Medium	Wang (2018), chunk 329

Table 6. Fuzzy inference results for healthcare supply chain risk factors.

Code	Risk factor	Risk score	Risk level
S4	Shortage of drugs, unavailability on market	0.87	Very High
S1	Counterfeiting	0.84	Very High
S8	Poor IT system, lack of data standardisation	0.78	High
S5	Capability versus demand	0.75	High
S11	Lack of visibility of stock	0.73	High
S10	Weak logistics service infrastructure	0.68	High
S15	High product and supplier variety	0.65	High
S14	High purchase price	0.62	Medium–High
S13	Clinician's preference	0.58	Medium
S9	Dispensing/picking errors	0.55	Medium
S2	Poor quality in purchased drugs	0.52	Medium

Table 7. Final risk assessment results.

Rank	Code	Risk factor	Fuzzy score	Dominant belief	Belief Interval	Risk Level
1	S4	Drug shortage / unavailability	0.87	VH: 0.770	[0.770, 0.827]	Very High
2	S1	Counterfeiting	0.84	VH: 0.695	[0.695, 0.810]	Very High
3	S8	Poor IT system / data standardisation	0.78	H: 0.621	[0.621, 0.758]	High
4	S5	Capability versus demand	0.75	H: 0.587	[0.587, 0.732]	High
5	S11	Lack of stock visibility	0.73	H: 0.565	[0.565, 0.720]	High
6	S10	Weak logistics infrastructure	0.68	H: 0.510	[0.510, 0.685]	High
7	S15	High product/supplier variety	0.65	H: 0.480	[0.480, 0.650]	Medium–High
8	S14	High purchase price	0.62	M–H: 0.445	[0.445, 0.620]	Medium–High
9	S13	Clinician's preference	0.58	M: 0.410	[0.410, 0.580]	Medium
10	S9	Dispensing/picking errors	0.55	M: 0.385	[0.385, 0.555]	Medium
11	S2	Poor quality purchased drugs	0.52	M: 0.360	[0.360, 0.520]	Medium

4.3.1. Step 1: document ingestion

The document corpus is chunked and embedded into the vector database. For this pilot, the corpus comprises 18 documents: 16 published SCRM/healthcare-supply-chain journal and conference papers (including Breen 2008; Christopher and Peck 2004; Elleuch, Hachicha, and Chabchoub 2014; Ho et al. 2015; Kleindorfer and Saad 2005; Ritchie and Brindley 2007, and others listed in the reference list), the Lord Carter (2016) NHS report, and the Wang (2018) thesis (results and discussion chapters, not raw interview transcripts – see Section 4.2). Text is extracted with layout-aware PDF parsing; publisher/platform access watermarks are filtered so they are not embedded as spurious content. Chunks target 300–500 tokens with paragraph-aligned boundaries and ~17.5% overlap, per the reference configuration in Section 3.3.1. This yields 1,564 chunks (540,147 tokens; mean 345 tokens/chunk). Each chunk carries metadata: source document, document type, and page range. This corpus is smaller than the full scope described in Section 4.2; Section 4.8 reports results both with and without the Wang (2018) thesis present, since the thesis is also the source of the ground-truth risk factors and strategies used to score extraction accuracy.

4.3.2. Step 2: query and retrieval

The retriever receives a risk-oriented query and returns the most relevant document chunks.

Query: ‘What are the primary risk factors affecting drug availability and patient safety in hospital supply chains?’

Table 4 shows the top retrieved chunks (abbreviated for space).

All top-ranked chunks for this query were drawn from the Wang (2018) thesis; see Section 4.8 for the thesis-excluded condition and Table 8 for a query where retrieval also surfaced independent literature (Breen 2008).

4.3.3. Step 3: grounded LLM structured output

The grounded LLM processes the retrieved chunks and produces a structured risk factor table. The LLM is instructed: ‘Based only on the retrieved documents, identify and categorize the risk factors mentioned. For each factor, provide a severity assessment using the scale: Very Low, Low, Medium, High, Very High. Cite the source for each factor’.

Table 8. Sample retrieval results for strategy identification query.

Rank	Source	Type	Retrieved content (excerpt)
1	Wang (2018) thesis	PhD thesis	“... The following 9 appropriate risk management strategies have been identified through the empirical studies from both UK and China hospitals ... Logistics Management ... Collaboration Vertical and Horizontal ... eProcurement Strategy ... Outsourcing ... Agility ... Capability Management ... Single Sourcing ...”
7	Breen (2008)	Journal article	“... the supply chain would benefit from having a coordinating body that is responsible for setting targets and meeting deadlines, and implementing strategy ... some of the risks can effectively be addressed through better co-ordination and management of the PSC and some through effective mitigation strategies ...”
8	Breen (2008)	Journal article	“... risks in the PSC are associated with product discontinuity, product shortages, poor performance, patient safety/dispensing errors, technological errors (causing stock shortages in pharmacies) ...”

Table 5 shows the LLM's structured output (partial).

Twenty-four risk factors were extracted in total across the full pilot run; the table shows a representative subset. Categories mirror the thesis's own Supply/Demand/Process/Control/Environmental scheme, a finer-grained taxonomy than the paper's collapsed three-category scheme, since the model extracted directly from the source table's own categorisation. Full output archived alongside the pilot code.

The LLM identifies risk factors, assigns them to categories, provides preliminary severity assessments in linguistic terms, and cites the source documents. It does not make final risk judgements – that is the expert system's role.

4.3.4. Step 4: structured evidence extraction

The Structured Evidence Extractor converts the LLM output into formats the expert system can consume:

- **Linguistic terms to TFNs:** 'Very High' maps to (0.75, 1.0, 1.0); 'High' maps to (0.5, 0.75, 1.0); 'Medium' maps to (0.25, 0.5, 0.75), following Table 1.
- **Belief mass functions:** for each risk factor, the extractor constructs a BPA from the evidence. For 'Drug shortage', two independent research papers and operational data all indicate Very High severity. The BPA:

$$m(\{VeryHigh\}) = 0.75, m(\{High, VeryHigh\}) = 0.15, m(\Theta) = 0.10$$

For 'Dispensing/picking errors', only one source assesses severity at Medium, with no corroboration:

$$m(\{Medium\}) = 0.50, m(\{Medium, High\}) = 0.20, m(\Theta) = 0.30$$

The higher uncertainty mass ($m(\Theta) = 0.30$) reflects the weaker evidence base.

4.3.5. Step 5: handling an under-grounded or erroneous extraction

The grounding and validation mechanisms specified in Sections 3.3.3 and 3.4 are designed to catch exactly this kind of weak-evidence case before it reaches the knowledge base, and to catch outright extraction errors. Consider a hypothetical illustration: if the grounded LLM were to output a claim such as 'clinician's preference is the leading cause of counterfeiting risk' without a retrieved chunk that supports this specific causal link, the post-hoc citation check (Section 3.3.3) would find no supporting quotation for the claim in the cited chunk's text and flag it rather than admit it automatically. Flagged claims, together with claims like the 'Dispensing/picking errors' example above whose supporting evidence is thin, are routed to the expert validation checkpoint (Section 3.4) rather than entering the knowledge base directly. In this design, the basic probability assignment's uncertainty mass functions as a continuous confidence signal, high uncertainty mass indicates thin or single-source evidence, while the citation check functions as a discrete gate against unsupported claims; together they provide two complementary lines of defence against hallucinated or over-confident extractions before they influence a risk assessment. This mechanism was tested against real executed output in the pilot evaluation (Section 4.8), not left as an unmeasured design safeguard. Across all extraction runs, the citation check's hallucination flag rate was 0.00: no extracted claim's supporting quotation failed the substantive-grounding check. This is a clean result at pilot scale, but it should be read cautiously: it is not proof that the mechanism is failure-proof, nor an estimate of the framework's true hallucination rate. It reflects only two queries, an 18-document corpus and a single LLM configuration, so it supports only the narrow statement that no ungrounded claim was flagged in these runs. The check itself required correction during pilot development to avoid false-positive flags from routine PDF-extraction artefacts (line-wrap hyphenation and typographic quotation marks being rendered differently between the source text and the model's 'verbatim' quotation). Automated grounding checks are themselves imperfect instruments requiring validation, not a guarantee once specified. Section 5.4 discusses this further.

4.4. Fuzzy inference application

The fuzzy inference engine applies the rule base to the fuzzified risk assessments.

4.4.1. Sample fuzzy rules

The rule base encodes domain knowledge about healthcare supply chain risk. Sample rules:

- **R1:** IF *drug_shortage_likelihood* is HIGH AND *supply_chain_visibility* is LOW, THEN *supply_disruption_risk* is VERY HIGH
- **R2:** IF *counterfeiting_risk* is HIGH AND *traceability_level* is LOW, THEN *patient_safety_risk* is VERY HIGH
- **R3:** IF *IT_system_quality* is POOR AND *data_standardization* is LOW, THEN *operational_efficiency_risk* is HIGH
- **R4:** IF *demand_variability* is HIGH AND *inventory_visibility* is LOW, THEN *stockout_risk* is HIGH
- **R5:** IF *logistics_infrastructure* is WEAK AND *cold_chain_compliance* is LOW, THEN *product_quality_risk* is HIGH

4.4.2. Inference computation

Using the expert survey data from Wang et al. (2023), where experts assessed risk factors using linguistic scales, the fuzzification module converts each assessment to TFNs. The inference engine evaluates all matching rules and produces aggregated fuzzy outputs. After defuzzification using the centroid method (Eq. 1), each risk factor receives a crisp risk score on a [0, 1] scale.

Table 6 shows the fuzzy inference results for the 11 priority risk factors. The risk scores reflect the aggregated expert assessments processed through the fuzzy rule base.

These results are broadly consistent with the Wang et al. (2023) empirical findings, which ranked S4 (drug shortage) and S1 (counterfeiting) as the highest-priority risks.

4.5. Evidential reasoning synthesis

The synthesis module combines evidence from multiple experts using Dempster–Shafer theory. We demonstrate the process for risk factor S4 (Shortage of drugs).

4.5.1. Constructing belief structures

From the 56 expert responses in Wang et al. (2023), three representative expert groups are identified based on their assessment of S4:

Expert Group 1 (hospital procurement managers, $n = 20$): strong consensus on Very High risk.

$$m_1(\{VH\}) = 0.70, m_1(\{H, VH\}) = 0.20, m_1(\Theta) = 0.10$$

Expert Group 2 (pharmaceutical company managers, $n = 16$): assess High to Very High risk.

$$m_2(\{H\}) = 0.30, m_2(\{VH\}) = 0.40, m_2(\{H, VH\}) = 0.20, m_2(\Theta) = 0.10$$

Expert Group 3 (academic researchers, $n = 20$): moderate assessment with more uncertainty.

$$m_3(\{H\}) = 0.35, m_3(\{VH\}) = 0.30, m_3(\{H, VH\}) = 0.15, m_3(\Theta) = 0.20$$

4.5.2. Applying Dempster's rule

Combining m_1 and m_2 using Dempster's rule (Eq. 3):

First, compute the conflict factor $K_{1,2}$. The only conflicting combination is $m_1(\{VH\}) \times m_2(\{H\}) = 0.70 \times 0.30 = 0.21$ (since $\{VH\} \cap \{H\} = \emptyset$).

$$K_{1,2} = 0.21$$

$$1 - K_{1,2} = 0.79$$

The combined mass $m_{1,2}$ for key focal elements:

$$m_{1,2}(\{VH\}) = \frac{0.70 \times 0.40 + 0.70 \times 0.20 + 0.70 \times 0.10 + 0.20 \times 0.40 + 0.10 \times 0.40}{0.79}$$

$$\begin{aligned}
&= \frac{0.28 + 0.14 + 0.07 + 0.08 + 0.04}{0.79} = \frac{0.61}{0.79} = 0.772 \\
m_{1,2}(\{H\}) &= \frac{0.20 \times 0.30 + 0.10 \times 0.30}{0.79} = \frac{0.06 + 0.03}{0.79} = \frac{0.09}{0.79} = 0.114 \\
m_{1,2}(\{H, VH\}) &= \frac{0.20 \times 0.20 + 0.20 \times 0.10 + 0.10 \times 0.20}{0.79} = \frac{0.04 + 0.02 + 0.02}{0.79} = \frac{0.08}{0.79} = 0.101 \\
m_{1,2}(\Theta) &= \frac{0.10 \times 0.10}{0.79} = \frac{0.01}{0.79} = 0.013
\end{aligned}$$

Now combine $m_{1,2}$ with m_3 . The conflict factor K between $m_{1,2}$ and m_3 :

$$K = m_{1,2}(\{VH\}) \times m_3(\{H\}) + m_{1,2}(\{H\}) \times m_3(\{VH\}) = 0.772 \times 0.35 + 0.114 \times 0.30 = 0.270 + 0.034 = 0.304$$

$$1 - K = 0.696$$

The final combined mass $m_{1,2,3}$ (key elements):

$$\begin{aligned}
m_{final}(\{VH\}) &= \frac{0.772 \times 0.30 + 0.772 \times 0.15 + 0.772 \times 0.20 + 0.101 \times 0.30 + 0.013 \times 0.30}{0.696} \\
&= \frac{0.232 + 0.116 + 0.154 + 0.030 + 0.004}{0.696} = \frac{0.536}{0.696} = 0.770 \\
m_{final}(\{H\}) &= \frac{0.114 \times 0.35 + 0.114 \times 0.15 + 0.114 \times 0.20 + 0.101 \times 0.35 + 0.013 \times 0.35}{0.696} \\
&= \frac{0.040 + 0.017 + 0.023 + 0.035 + 0.005}{0.696} = \frac{0.120}{0.696} = 0.172 \\
m_{final}(\{H, VH\}) &= 0.053 \\
m_{final}(\Theta) &= 0.004
\end{aligned}$$

4.5.3. Belief and plausibility

For risk factor S4, the final combined assessment (applying Eqs. 4 and 5):

$$Bel(\{VH\}) = 0.770, Pl(\{VH\}) = 0.770 + 0.053 + 0.004 = 0.827$$

$$Bel(\{H\}) = 0.172, Pl(\{H\}) = 0.172 + 0.053 + 0.004 = 0.229$$

The belief interval for 'Very High' risk is [0.770, 0.827]. The combined evidence from all three expert groups strongly supports a Very High risk assessment for drug shortage, with relatively low residual uncertainty ($m(\Theta) = 0.004$).

4.6. Results

Table 7 presents the final risk assessment results combining fuzzy inference scores and evidential reasoning synthesis for all 11 priority risk factors.

The results reveal several findings:

Drug shortage and counterfeiting are the highest-priority risks. Both receive 'Very High' assessments with narrow belief intervals, indicating strong expert consensus. This aligns with Wang et al. (2023) empirical finding that these factors 'have been identified as having the most significant effect among all 34 risk factors'.

IT system quality and stock visibility are strongly linked operational risks. Both rank in the top five. The NHS faces a well-documented challenge: 130 different information descriptions exist for a single product

in the NHS database (Kreysa and Denecker 2009). Poor data standardisation drives lack of visibility, which drives stockout risk.

Wider belief intervals indicate less expert consensus. For lower-ranked risks such as clinician's preference (S13) and dispensing errors (S9), the belief intervals are wider, reflecting greater disagreement among expert groups. Such information is unavailable from simple weighted averages but helps decision-makers identify where additional evidence is needed.

The pilot evaluation (Section 4.8) found that RAG-grounded extraction achieved a mean extraction recall of 0.87 for the priority risk factors with the full pilot corpus and 0.76 with the Wang (2018) thesis excluded from retrieval. It substantially outperformed the no-retrieval baseline on the harder mitigation-strategy task (1.00 vs 0.49 recall) while performing comparably on risk-factor identification (0.87 vs 0.80). These are extraction-level scores for the final structured output of the RAG component, not retrieval metrics, and they were not propagated through the fuzzy and Dempster-Shafer stages that produced Tables 6 and 7 (Section 4.8). The worked example in this section shows this process concretely; Section 4.8 reports the full quantitative result.

4.6.1. Additional scenarios: contrasting evidential complexity

The drug-shortage walkthrough in Section 4.5 illustrates a high-consensus scenario. Table 7 reports two further cases that illustrate how the framework's belief-interval output reflects differing levels of evidential complexity.

Counterfeiting (S1) ranks second overall (fuzzy score 0.84, belief interval [0.695, 0.810], Very High). As with drug shortage, multiple independent sources (Wang et al. 2023; regulatory guidance on anti-counterfeiting requirements) corroborate a high-severity assessment, but the belief interval is wider than S4's, reflecting comparatively less unanimous evidence across expert groups. Under the recommendation logic of Section 4.7, this scenario maps to strategy A2 (advanced information technology and systems), whose description explicitly includes RFID-enabled traceability.

Poor IT system quality (S8) ranks third (fuzzy score 0.78, belief interval [0.621, 0.758], High). This scenario illustrates a lower-consensus case: the risk level drops from Very High to High and the belief interval sits further from certainty, consistent with a more heterogeneous evidence base across the underlying expert panels (Section 4.1.1). It maps to the same strategy, A2, whose data-standardisation component targets the risk directly.

Together, the three scenarios span the range of evidential complexity the framework is designed to handle, strong multi-source consensus (S4), moderate consensus with wider uncertainty (S1), and comparatively weaker consensus (S8), while still producing a consistent, traceable path from evidence to recommendation in each case.

4.7. Mitigation strategy evaluation

Having ranked the 11 risk factors, the framework is extended to evaluate the nine mitigation strategies (A1-A9, Table 3). The RAG layer first extracts strategy-effectiveness evidence from the document corpus; the Structured Evidence Extractor converts this into linguistic ratings; fuzzy AHP weights the risk criteria; and fuzzy TOPSIS ranks the strategies. This describes the designed data flow; the rankings in Table 9 are computed from the expert panel's strategy ratings, and the pilot of Section 4.8 did not feed RAG-extracted ratings into the fuzzy AHP/TOPSIS stage.

Table 9. Mitigation strategy rankings (fuzzy TOPSIS).

Rank	Code	Strategy	$\$d^{\wedge}+\$$	$\$d^{\wedge}-\$$	CC
1	A2	Advanced IT and systems	0.370	10.729	0.967
2	A5	eProcurement strategy	0.431	10.688	0.961
3	A1	Cold chain management	0.515	10.637	0.954
4	A3	Inter/intra-organisational collaboration	0.522	10.616	0.953
5	A4	Inventory management	0.527	10.632	0.953
6	A8	Internal capability management	0.549	10.598	0.951
7	A7	Agility strategy	0.555	10.589	0.950
8	A9	Single sourcing	0.577	10.549	0.948
9	A6	Outsourcing strategy	0.583	10.535	0.948

RAG layer application. The retriever is queried with strategy-oriented prompts targeting the document corpus established in Section 4.2. A representative query:

What supply chain risk management strategies are recommended for addressing drug shortages, counterfeiting, and IT system failures in hospital supply chains?

Table 8 shows the top retrieved chunks.

Unlike the risk-factor query (Table 4), retrieval for this query surfaced both the Wang (2018) thesis and independent literature (Breen 2008), illustrating genuine cross-source retrieval; Section 4.8 reports how strongly the strategy-extraction result depends on the thesis's presence.

The grounded LLM processes these chunks and produces a structured strategy-effectiveness table. The table maps each strategy to the risk factors it addresses and assigning a preliminary effectiveness rating in linguistic terms (Very Poor to Very Good). This output is validated at the expert checkpoint before entry into the TOPSIS evaluation, replacing the need to elicit ratings from scratch.

Risk category weights. Fuzzy AHP pairwise comparisons across the three categories (C.R. = 0.001) yield: external to hospital but internal to the supply chain network, **0.53**; internal to hospital, **0.43**; external to the supply chain network, **0.04**. The dominance of supply-side risks (0.53) is consistent with S4 and S1 ranking highest in Section 4.6.

Fuzzy TOPSIS. The RAG-structured linguistic ratings, validated by experts, are aggregated into a fuzzy decision matrix. The relative closeness coefficient (Eq. 6) is:

$$CC_i = \frac{d_i^* -}{d_i^* - + d_i^* +} \quad (6)$$

Table 9 presents the rankings.

Findings. Advanced IT and systems (A2) ranks first (CC = 0.967), followed by eProcurement (A5, CC = 0.961). Both directly address the highest-ranked risks: A2 targets S8 (poor IT/data standardisation), S11 (lack of stock visibility), and S4 (drug shortage through real-time traceability); A5 automates procurement information flows, reducing exposure to S4, S5 (capability versus demand), and S14 (high purchase price). Their joint dominance is consistent with the IT-operational risk cluster in Section 4.6, and with the NHS report evidence retrieved by the RAG layer (Lord Carter 2016; Expert 23).

Outsourcing (A6) and single sourcing (A9) rank lowest, indicating weak alignment with the priority risk profile.

Sensitivity analysis under 10 weight-perturbation conditions confirms robustness: A2 maintains a top-2 position in 8 of 10 conditions; A5 in 9 of 10. The strategy rankings are stable across all permutations.

4.8. Pilot evaluation

Sections 4.3 and 4.7 illustrate the RAG layer's operation using output from an executed pilot. This section reports the pilot's method and results in full, covering extraction accuracy, hallucination flag rate, and latency.

Scope of the evaluation. Two clarifications govern how every figure below should be read. First, the precision and recall values reported here evaluate the final extracted outputs (the structured risk-factor and strategy lists produced by the grounded LLM) against the panel-validated ground truth. They are extraction-level measures, not retrieval measures. Conventional retrieval metrics such as Precision@k, Recall@k, mean reciprocal rank (MRR) or nDCG are not reported, because they require relevance judgements on individual chunks that were not constructed for this pilot. Second, the experiment validates the RAG component (retrieval, grounded extraction and citation check) in isolation. It does not validate the end-to-end propagation of RAG-derived evidence through the linguistic-variable mapping, belief-mass-function construction, fuzzy inference and Dempster-Shafer synthesis stages: the fuzzy and evidential results in Sections 4.4-4.6 are computed from the survey-derived panel data of Section 4.1.1, not from pilot output.

Corpus. The pilot corpus comprises 18 documents: 16 published SCRM/healthcare-supply-chain journal and conference papers drawn from the reference list (including Breen 2008; Christopher and Peck 2004; Elleuch, Hachicha, and Chabchoub 2014; Ho et al. 2015; Kleindorfer and Saad 2005; Ritchie and Brindley 2007), the Lord Carter (2016) NHS report, and the Wang (2018) thesis (results and discussion chapters). Text

was extracted with layout-aware PDF parsing; publisher/platform access watermarks (present in downloads from Emerald, Wiley, IEEE, JSTOR, and Elsevier) were identified and filtered, since they would otherwise be embedded as spurious content. Chunks target 300–500 tokens with paragraph-aligned boundaries and ~17.5% overlap between consecutive chunks, per the reference configuration in Section 3.3.1. This yielded 1,564 chunks (540,147 tokens; mean 345 tokens/chunk). This corpus is smaller than the full input-layer scope described in Section 4.2: NHS/WHO/MHRA primary regulatory guidance and supplier audit reports are not yet included, and are scoped for full-scale future work (Section 6).

Method. Retrieval used BAAI/bge-large-en embeddings (open-weight, on-premises deployment, per the reference configuration’s data-governance rationale in Section 3.3.1) with FAISS cosine-similarity search, retrieving the top 32 candidates and reranking with BAAI/bge-reranker-base to a final top-8. The grounded LLM was Claude Sonnet 4.5 (Anthropic), prompted with the grounding instruction specified in Section 3.3.3 and constrained to structured output via tool-calling; temperature was set to 0. The post-hoc citation check (Section 3.3.3) verified that each extracted claim’s cited chunk exists in the retrieved set and that the supporting quotation is substantively present in that chunk’s text, after normalising PDF line-wrap hyphenation and typographic quotation marks so that typesetting artefacts are not mistaken for grounding failures.

Three specific choices in this pilot diverge from the illustrative options named in Sections 3.3.1–3.3.3 and are disclosed here rather than left implicit. First, the vector index used exact search (FAISS IndexFlatIP) rather than the approximate HNSW search named as the reference option; at this corpus scale (1,564 chunks) exact search is fast enough that the approximation was unnecessary, but production deployment at the corpus scales Section 3.3.1 anticipates would likely require HNSW or an equivalent approximate method. Second, metadata filters (document type, jurisdiction, recency) were not applied, since the pilot corpus contains no superseded or conflicting-jurisdiction documents to filter; this mechanism remains unexercised, not validated. Third, the grounded LLM was prompted zero-shot, with the grounding instruction and output schema only, rather than with the few-shot grounded examples named as part of the reference prompting strategy; the zero-shot version described here reflects what was actually run.

Confidence scoring maps onto two distinct mechanisms in the framework, one executed in this pilot and one not. The first is the grounded LLM’s own per-claim confidence judgement, executed here: when instructed to output ‘Insufficient evidence’ rather than guess a severity or effectiveness rating, the model exercised this option at a real, substantial rate for the mitigation-strategy extraction task (58.5% of extracted items in the full-corpus condition, mean over five trials, range 45.0–68.8%; 0.0% for the risk-factor task and for both thesis-excluded conditions). This is a genuine, disclosed precision-confidence tradeoff: the strategies/full-corpus condition’s 1.00 recall (Table 10) was achieved while the model declined to rate the majority of what it extracted, rather than assigning a confident-sounding severity it could not support from the retrieved context. The second mechanism is the Dempster-Shafer basic probability assignment’s uncertainty mass, described in Section 3.4 and Step 4 of Section 4.3 as a continuous confidence signal derived from evidence convergence across sources; this is illustrated on the real survey-derived case-study data in Section 4.5 (for example, the narrow belief interval [0.770, 0.827] for drug shortage reflects strong multi-source consensus, versus wider intervals for risks with thinner evidence), but was not computed on this pilot’s RAG-extracted output, consistent with the scope stated in Section 4.3.

Two queries were evaluated: the risk-factor identification query and the mitigation-strategy query used in the Section 4.3 and 4.7 worked examples. Extraction was scored against the panel-validated ground truth (11 risk factor codes, Table 2; 9 mitigation strategy codes, Table 3) by aligning extracted items to ground-truth codes via an LLM judge (Claude Sonnet 4.5, prompted separately from the extraction step, temperature 0); this is a disclosed methodological choice, not independent human review (Section 5.4). Because the Wang (2018) thesis is both a corpus source and the origin of the ground-truth codes, each query was run under two conditions: the full 18-document corpus, and a 17-document corpus with the thesis excluded, to test

Table 10. Pilot evaluation results. Recall and precision are extraction-level measures of the final outputs, not retrieval metrics (mean [range] over 5 trials).

Condition	Risk factors: recall	Risk factors: precision	Strategies: recall	Strategies: precision	Hallucination flag rate
Baseline (no retrieval)	0.80 [0.73–0.91]	0.46 [0.44–0.50]	0.49 [0.44–0.56]	0.21 [0.20–0.25]	n/a
RAG, full corpus	0.87 [0.82–1.00]	0.40 [0.38–0.48]	1.00 [1.00–1.00]	0.50 [0.45–0.56]	0.00 [0.00–0.00]
RAG, thesis excluded	0.76 [0.73–0.82]	0.41 [0.26–0.50]	0.49 [0.33–0.67]	0.30 [0.25–0.33]	0.00 [0.00–0.00]

whether the RAG layer can recover the panel-validated factors from independent literature alone rather than from a source that already enumerates them. Each condition was run 5 times (temperature 0 substantially reduces but does not eliminate run-to-run variance on the underlying API) and results are reported as mean [range].

Mean end-to-end latency (retrieval + reranking + extraction) was 33–49s for RAG conditions versus 13s for the no-retrieval baseline.

Interpretation. RAG-grounded extraction outperformed the no-retrieval baseline in recall on both queries, markedly so for mitigation strategies (1.00 vs 0.49) and modestly for risk factors (0.87 vs 0.80); risk-factor precision was slightly lower than the baseline (0.40 vs 0.46). This asymmetry is consistent with the framework’s motivating claim: generic pharmaceutical supply-chain risks are partially recoverable from an LLM’s parametric knowledge, but the specific mitigation strategies validated by this expert panel are not, and require grounding in the actual source documents. No hallucination was flagged by the citation check in any RAG condition or in any of the five trials. We treat this as an observation, not a conclusion about the framework’s hallucination rate: it rests on two queries and an 18-document corpus, the check only tests whether a quotation is present in the cited chunk (a correctly quoted but misinterpreted passage would pass), and the pilot contained no adversarial cases designed to make the check fail. A larger query set and corpus would be needed before any rate could be stated with confidence; Section 5.4 discusses this further.

The thesis-exclusion condition is informative in both directions. For risk factors, excluding the thesis had a modest effect on recall (0.87 to 0.76), indicating the RAG layer can substantially reconstruct the panel’s findings from independent literature, not merely retrieve a document that already lists them. For mitigation strategies, excluding the thesis reduced recall considerably (1.00 to 0.49), indicating that the full-corpus strategy result depends more heavily on the thesis’s presence. Reporting both conditions, rather than only the more favourable full-corpus number, is a deliberate design choice: it distinguishes what the pilot demonstrates about genuine cross-source synthesis from what it demonstrates about retrieval within a corpus that happens to contain the answer key. These results should be read as a bounded pilot, not a full-scale validation; Section 5.4 details the specific ways in which this evaluation is limited and what remains for future work.

5. Discussion

5.1. Theoretical contributions

This paper makes three contributions to SCRM research.

First, it proposes a framework that, to the best of our knowledge, is the first to integrate RAG as an upstream knowledge acquisition mechanism for expert systems in SCRM; we are not aware of earlier work doing so, although given the pace of publication on RAG we cannot exclude concurrent or unindexed studies. Prior work has used RAG for downstream tasks. These tasks include explaining expert system outputs (Ren, Jenkinson, and Tobora 2026) and answering questions from document corpora. This framework uses RAG to solve the knowledge acquisition bottleneck: the persistent difficulty of extracting, structuring, and encoding domain knowledge from heterogeneous unstructured sources into the formal representations that expert systems require.

Second, it demonstrates that LLM capabilities and formal reasoning methods are complementary, not competing. The RAG layer extracts and structures knowledge; the fuzzy inference engine and evidential reasoning module assess and synthesise it. Neither alone is sufficient. LLMs cannot quantify uncertainty or combine conflicting evidence under mathematical rules. Expert systems cannot scale their knowledge acquisition to handle growing document volumes. Together, they address each other’s limitations.

Third, it provides a general architecture that is domain-independent. The reasoning modules, fuzzy inference and Dempster-Shafer synthesis, operate on structured inputs. Their operation is consistent across domains. The RAG layer adapts to new domains by changing the input document corpus. The same architecture that serves healthcare SCRM can serve manufacturing, energy, food, or any sector where risk management requires processing heterogeneous knowledge under uncertainty.

Table 11. Qualitative comparison of approaches to SCRM under uncertainty.

Dimension	Expert system only	Standalone LLM/RAG	Bayesian network	Proposed framework
Knowledge acquisition scalability	Manual elicitation; does not scale	Scales automatically with document volume	Manual elicitation of priors/structure; does not scale	Scales via RAG layer; expert checkpoint only
Uncertainty quantification	Rule-based, typically point estimates	None (point predictions, no calibrated uncertainty)	Probabilistic; requires precise priors	Belief/plausibility intervals via D-S theory
Explicit modelling of ignorance	Limited	None	Not directly (assigns probability even under ignorance)	Yes, mass assigned to the frame of discernment
Auditability / traceability	High (explicit rules)	Low (opaque generation, no source trail)	Moderate (probabilistic structure is inspectable)	High (source-to-decision traceability, Section 3.8)
Hallucination risk	None (no generative component)	High if ungrounded	None	Mitigated via grounding + validation checkpoint (Section 3.3.3)
Maintainability as knowledge evolves	Low (rule base requires manual rebuild)	High (no persistent knowledge base)	Low (structure/priors need re-elicitation)	High (add documents to vector store, Section 5.2)

5.1.1. Comparative positioning

Table 11 positions the proposed framework qualitatively against three alternative approaches: expert systems built without RAG (the traditional approach this paper extends), standalone LLM/RAG systems without formal reasoning, and Bayesian network approaches to uncertainty quantification, an alternative comparator to Dempster-Shafer evidential reasoning. The comparison is qualitative and architectural, reflecting each approach's structural capabilities rather than measured performance; quantitative head-to-head benchmarking against these baselines, using a common prototype and dataset, is identified as future work in Section 5.4.

5.2. Practical implications

For supply chain managers, the framework offers a decision support tool capable of processing large volumes of heterogeneous information. This information includes regulatory guidelines, operational data, and expert opinions, allowing the tool to produce risk assessments with quantified uncertainty. The belief intervals in Table 7 tell managers not only which risks rank highest but also where expert consensus is strong and where it is weak.

The framework reduces dependency on expert availability for knowledge acquisition. Traditional SCRM studies require months of questionnaire distribution and interview scheduling. Wang et al. (2023) distributed 431 questionnaires between February and May 2017 to obtain 56 valid responses. In the executed pilot (Section 4.8), the RAG layer's retrieval-and-extraction step for a single query against the 18-document pilot corpus completed in 33–49 seconds, several orders of magnitude faster than the manual elicitation cycle described below, though this measures only the retrieval-and-extraction step against an already-assembled corpus, not the full cycle of corpus assembly, questionnaire design, and expert validation that a fair end-to-end comparison would require. Experts then validate and refine these assessments rather than building them from scratch. This four-month survey window was followed by a further round of interviews between June and August 2017 with a separate twelve-expert panel to establish risk interdependencies (Section 4.1.1). A further round followed with a twenty-expert panel for mitigation-strategy evaluation (Section 4.7). This was a combined effort spanning most of a calendar year across three distinct expert panels. This framework's RAG layer is designed to acquire the same knowledge from existing documents in a much shorter cycle, with experts then validating rather than originating that knowledge. This represents a substantially lighter burden.

The system is also maintainable. When new regulations, incident reports, or clinical guidelines are published, they can be chunked and added to the vector database without rebuilding the rule base. The retriever will surface newly relevant evidence in response to queries. This continuous updating capability addresses a limitation of static expert systems, whose knowledge bases become outdated as domains evolve.

5.3. Generalizability

The framework architecture is domain-independent. The six layers are defined in general terms. They include input, RAG, structured evidence extractor, knowledge base, fuzzy inference, and evidential reasoning synthesis. Domain specificity enters only through the input documents and the rule base content.

The original fuzzy-evidential expert system was developed and validated for offshore safety assessment (Ren et al. 2005), using 245 rules across three input dimensions assessed by five experts. The present paper demonstrates the same reasoning architecture applied to healthcare supply chain risk management, with different risk factors, different expert panels, and different document corpora. The architecture transfers without modification.

Other domains present natural extensions. Manufacturing supply chains face disruption risks from supplier failures, quality defects, and demand volatility. All of these risks can be assessed using the same fuzzy-evidential framework with appropriate document corpora. Energy supply chains, food supply chains, and pharmaceutical distribution networks present similar knowledge structures.

The RAG layer makes domain adaptation practical: rather than rebuilding the knowledge base from scratch for each domain, analysts populate the vector database with domain-relevant documents and let the retriever and grounded LLM extract structured evidence. Recent applications of formal AI methods in related safety-critical domains highlight the versatility of these approaches. Examples include Bayesian network models for offshore wind farm decommissioning (Shafiee and Adedipe 2023), LNG shipping route risk assessment (Z. Chang et al. 2023), and AI-enabled predictive safeguards for offshore oil facilities (Ren 2026), underscoring the breadth of sectors suitable for the framework's approach.

5.4. Limitations

Four limitations should be noted.

First, this paper presents a conceptual framework with a case study demonstration, strengthened by a pilot execution of the RAG layer reported in Section 4.8. The pilot is real but bounded, and it validates the RAG component rather than the complete pipeline: it uses a smaller, differently-sourced corpus than the full input-layer scope described in Section 4.2, and evaluates two representative queries rather than the full range of risk identification, characterisation, relationship, and mitigation queries the framework is designed to support. A larger-scale prototype implementing all six layers, evaluated across the full query range and the full corpus scope, is still needed.

This limitation is scoped narrowly, not paper-wide: the risk factors, severity assessments, and mitigation-strategy ratings that anchor the case study are empirically validated through the expert panels and reliability testing summarised in Section 4.1.1, independent of the RAG pilot. Within the pilot itself, seven further caveats apply. (i) Extraction accuracy was scored by aligning extracted items to the panel-validated risk-factor and strategy codes using an LLM judge (Claude Sonnet 4.5, prompted separately from the extraction step), not independent human review; this is a defensible substitute at pilot scale but is not equivalent to expert adjudication, which remains the priority for full-scale validation. (ii) Because the Wang (2018) thesis is both a corpus source and the origin of the ground-truth codes, extraction accuracy against that ground truth is partly circular for corpus configurations that include it; Section 4.8 addresses this directly by also reporting results with the thesis excluded from retrieval. (iii) The 0.00 hallucination flag rate reported in Section 4.8 should be read cautiously and not generalised: given the limited number of pilot queries (two) and the small corpus (18 documents), it reflects only a narrow query set and a single LLM/embedding configuration, and the automated citation checker itself required two rounds of correction during development, after an initial naive character-matching approach produced false-positive hallucination flags from routine PDF-extraction artefacts rather than genuine ungrounded claims. This underscores that automated grounding checks are themselves imperfect instruments and should not be taken as a substitute for human review at production scale. (iv) The recall and precision figures reported in Section 4.8 measure extraction accuracy, whether the final structured output matches the panel-validated ground truth, not retrieval accuracy in isolation (conventional retrieval metrics, namely Precision@k, Recall@k, MRR and nDCG, were not computed, and every use of 'precision' and 'recall' in this paper refers to extraction-level scores); a weak retrieval that the LLM happens to compensate for from parametric knowledge, or a strong retrieval poorly extracted, would look identical

under this metric, and separating the two (e.g. precision@k over labelled relevant chunks) is a priority for future work. (v) The latency figures reported are measured at the pilot's fixed scale (18 documents, 1,564 chunks); scalability as corpus size grows or under concurrent queries has not been tested. (vi) The pilot validates retrieval, extraction, and the citation check; it does not extend end-to-end through the belief-mass-function construction, fuzzy inference, or Dempster-Shafer synthesis stages, which continue to operate on the real survey-derived assessments described in Section 4.1.1 rather than on pilot-extracted output. The experiment therefore validates the RAG component, not the end-to-end propagation of RAG-derived evidence through the fuzzy and Dempster-Shafer reasoning stages; how extraction errors or omissions would alter belief intervals and risk rankings downstream remains untested. (vii) On robustness in safety-critical settings: healthcare supply chain risk management is safety-adjacent rather than safety-critical in the immediate, real-time sense (a drug-shortage risk assessment informs procurement decisions rather than a point-of-care clinical action), but the downstream consequences of a wrong risk assessment, a misallocated mitigation effort or an overlooked shortage risk, are still patient-relevant. The framework's design responds to this with layered, not single-point, safeguards: no RAG-extracted claim enters the knowledge base without passing both the citation check and, for claims the model itself flags as insufficiently supported, human expert review; the pilot's 58.5% insufficient-evidence rate for mitigation-strategy extraction (Section 4.8) shows this mechanism engaging substantially in practice, not merely existing on paper. That said, this pilot does not constitute a safety-critical-deployment validation: it covers two queries, a single domain, and a single model configuration, with no adversarial or red-team testing of the grounding mechanism's failure modes, and the reported 0.00 hallucination rate reflects a mechanism that was never actually challenged to fail. Establishing the framework's suitability for safety-critical deployment would require, at minimum, adversarial testing designed to elicit hallucination rather than passive observation of a low rate, human-expert-adjudicated accuracy rather than the LLM-judge scoring used here (see caveat (i) above), and a staged, human-supervised rollout rather than autonomous operation.

Second, RAG grounding is imperfect. Despite instructions to use only retrieved context, LLMs can occasionally introduce information from their parametric knowledge. The expert validation checkpoint (Section 3.4) mitigates this risk but does not eliminate it. Advances in constrained decoding and faithful generation will strengthen this component.

Third, Dempster's rule of combination has known limitations with highly conflicting evidence. When expert assessments strongly disagree, the normalisation factor (1K) can produce counterintuitive results (Zadeh 1979). Alternative combination rules may be more appropriate in high-conflict scenarios. Examples include Yager's rule and the cautious rule. The framework's modular design allows substitution of the combination rule without changing other components.

Fourth, the case study uses empirical data from a specific geographic and temporal context (UK and China hospitals, 2017). The risk factors and their relative priorities may differ in other healthcare systems, time periods, or regulatory environments. The framework is general; the specific results are context-dependent.

6. Conclusion

This paper proposes a general framework that integrates Retrieval-Augmented Generation with fuzzy inference and Dempster-Shafer evidential reasoning for supply chain risk management. The framework addresses the knowledge acquisition bottleneck that limits traditional expert systems by using a RAG layer, which includes a vector database, semantic retriever, and grounded LLM, to automatically extract and structure domain knowledge from heterogeneous unstructured sources.

The framework was demonstrated through a healthcare supply chain risk management case study using empirical data from UK and China hospitals, and through a pilot execution of the RAG layer (Section 4.8). The RAG layer identified and structured the priority risk factors and mitigation strategies from a corpus of published research and regulatory reports, recovering the panel-validated risk factors and strategies at extraction recall rates that exceeded a no-retrieval baseline, with no hallucination flagged by the pilot's citation check (a result to be read cautiously given the pilot's two queries and 18-document corpus). The pilot evaluated the RAG component only; end-to-end propagation of RAG-derived evidence through the fuzzy and Dempster-Shafer stages was not tested. The fuzzy inference engine produced risk scores consistent with empirical expert assessments, computed independently of the RAG pilot. The Dempster-Shafer

synthesis module combined evidence from multiple expert groups, yielding risk rankings with quantified uncertainty intervals. The results identified drug shortage and counterfeiting as the highest-priority risks, consistent with independent empirical findings.

The framework's main contribution is architectural: it establishes a principled integration between modern LLM-based knowledge extraction and established formal reasoning methods. The RAG layer does not replace the expert system; it feeds it. The expert system does not limit RAG's knowledge acquisition capability, rather it grounds it in mathematical rigour. This complementarity addresses a gap in the current literature, where LLM-based and expert-system-based approaches have developed largely in isolation.

Three directions for future work are identified. First, the pilot RAG evaluation (Section 4.8) should be scaled to the full corpus and query range described in Section 4.2 – including NHS/WHO/MHRA primary regulatory documents and supplier audit reports – its extraction accuracy should be validated by independent human expert review rather than the LLM-judge scoring used at pilot scale, retrieval quality should be reported with conventional metrics (Precision@k, Recall@k, MRR, nDCG), and the evaluation should be extended end-to-end so that RAG-derived evidence is propagated through the fuzzy inference and Dempster-Shafer stages. Second, comparative studies should benchmark the framework against pure-LLM and pure-expert-system approaches to quantify the value of integration. Third, the framework should be applied to additional domains. These include manufacturing, energy, and food supply chains, allowing its cross-domain generalisability to be validated through independent case studies.

Author contributions

CRedit: **Jun Ren:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft; **Lei Wang:** Conceptualization, Data curation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing; **Jirapan Liangrokapart:** Data curation, Investigation, Writing – review & editing; **Duangpun Kritchanhai:** Data curation, Investigation, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project is partially supported by European Union's Horizon 2020 Research and Innovation Programme RISE under grant agreement [823759] REMESH.

References

- Bak, O., S. Shaw, C. Colicchia, and V. Kumar. 2022. "A Systematic Literature Review of Supply Chain Resilience in Small Medium Enterprises (SMEs): A Call for Future Research." In , 328–341 doi:10.1109/TEM.2020.3016988.
- Bandaly, D., A. Satir, Y. Kahyaoglu, and L. Shanker. 2012. "Supply Chain Risk Management – I: Conceptualization, Framework and Planning Process." *Risk Management* 14 (4): 249–271. <https://doi.org/10.1057/rm.2012.7>.
- Breen, L. 2008. "A Preliminary Examination of Risk in the Pharmaceutical Supply Chain (PSC) in the National Health Service (NHS) (UK)." *Journal of Service Science and Management* 1 (2): 193–199. <https://doi.org/10.4236/jssm.2008.12020>.
- Chang, D. Y. 1996. "Applications of the Extent Analysis Method on Fuzzy AHP." *European Journal of Operational Research* 95 (3): 649–655. [https://doi.org/10.1016/0377-2217\(95\)00300-2](https://doi.org/10.1016/0377-2217(95)00300-2).
- Chang, Z., X. He, H. Fan, W. Guan and L. He. 2023. "Leverage Bayesian Network and Fault Tree Method on Risk Assessment of LNG Maritime Transport Shipping Routes: Application to the China–Australia Route." *Journal of Marine Science and Engineering* 11 (9): 1722 doi:10.3390/jmse11091722.
- Chopra, S., and M. S. Sodhi. 2004. "Managing Risk to Avoid Supply-Chain Breakdown." *MIT Sloan Management Review* 46:53–61 <https://sloanreview.mit.edu/article/managing-risk-to-avoid-supplychain-breakdown/>.
- Christopher, M., and H. Peck. 2004. "Building the Resilient Supply Chain." *The International Journal of Logistics Management* 15 (2): 1–14. <https://doi.org/10.1108/09574090410700275>.
- Cohen, R. J., and M. E. Swerdlik. 2010. *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. 7th ed. New York: McGraw-Hill.

- Dempster, A. P. 1968. "A Generalization of Bayesian Inference." *Journal of the Royal Statistical Society, Series B* 30 (2): 205–247. <https://doi.org/10.1111/j.2517-6161.1968.tb00722.x>.
- El Baz, J., and S. Ruel. 2021. "Can Supply Chain Risk Management Practices Mitigate the Disruption Impacts on Supply chains' Resilience and Robustness? Evidence from an Empirical Survey in a COVID-19 Outbreak Era." *International Journal of Production Economics* 233:107972. <https://doi.org/10.1016/j.ijpe.2020.107972>.
- Elleuch, H., W. Hachicha, and H. Chabchoub. 2014. "A Combined Approach for Supply Chain Risk Management: Description and Application to a Real Hospital Pharmaceutical Case Study." *Journal of Risk Research* 17 (5): 641–663. <https://doi.org/10.1080/13669877.2013.815653>.
- Enyinda, C. I., F. Gebremikael, and A. O. Ogbuehi. 2014. "An Analytical Model for Healthcare Supply Chain Risk Management." *African Journal of Business and Economic Research* 9 (1): 13–27.
- Feigenbaum, E. A. 1977. "The Art of Artificial Intelligence: Themes and Case Studies in Knowledge Engineering." In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* 22–25 August 1977 Massachusetts Institute of Technology (MIT), 1014–1029.
- Fenies, P., M. Gourgand, and N. Tchernev. 2006. "A Framework for Supply Chain Performance Evaluation 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM 2006) Saint-Étienne, France Saint-Étienne, France." 39 : 535–540.
- Ghadge, A., S. Dani, M. Chester, and R. Kalawsky. 2013. "A Systems Approach for Modelling Supply Chain Risks." *Supply Chain Management: An International Journal* 18 (5): 523–538. <https://doi.org/10.1108/SCM-11-2012-0366>.
- Haigh, J. 2004. "Parallel trade in medicines: the results of a Social Market Foundation discussion seminar." *Social Market Foundation*. June.
- Ho, W., T. Zheng, H. Yildiz, and S. Talluri. 2015. "Supply Chain Risk Management: A Literature Review." *International Journal of Production Research* 53 (16): 5031–5069. <https://doi.org/10.1080/00207543.2015.1030467>.
- Jüttner, U., H. Peck, and M. Christopher. 2003. "Supply Chain Risk Management: Outlining an Agenda for Future Research." *International Journal of Logistics: Research and Applications* 6 (4): 197–210. <https://doi.org/10.1080/13675560310001627016>.
- Kanyoma, K. E., J. K. Khomba, E. J. Sankhulani, and R. Hanif. 2013. "Sourcing Strategy and Supply Chain Risk Management in the Healthcare Sector: A Case Study of Malawi's Public Healthcare Delivery Supply Chain." *Journal of Management and Strategy* 4 (3): 16–26. <https://doi.org/10.5430/jms.v4n3p16>.
- Kleindorfer, P. R., and G. H. Saad. 2005. "Managing Disruption Risks in Supply Chains." *Production & Operations Management* 14 (1): 53–68. <https://doi.org/10.1111/j.1937-5956.2005.tb00009.x>.
- Klüver, C. 2024. "A Requirements Model for AI Algorithms in Functional Safety-Critical Systems Security and Safety 3 2024020 doi:<https://doi.org/10.1051/sands/2024020>." In.
- Kreysa, U., and J. Denecker. 2009. "GS1 standards in healthcare: raising the bar on patient safety and supply chain efficiency." In *GS1 Healthcare Reference Book 2010*: (World Scientific Publishing) 1–5.
- Lewis, P., E. Perez, A. Piktus. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *NeurIPS* 33:9459–9474 doi:[10.48550/arXiv.2005.11401](https://doi.org/10.48550/arXiv.2005.11401).
- Liang, B., F. Jiang, and J. Li. 2026. "Multimodal Intelligent Hyperautomation for Sustainable Healthcare Supply Chains: A Blockchain-IoMT-AI Integration Perspective." *Enterprise Information Systems* 20 (5–6): 2677565. <https://doi.org/10.1080/17517575.2026.2677565>.
- Liu, J., J. B. Yang, J. Wang, and H. S. Sii. 2005. "Engineering System Safety Analysis Using Fuzzy Rule-Based Evidential Reasoning." *Quality and Reliability Engineering International* 21 (4): 387–411. <https://doi.org/10.1002/qre.668>.
- Lord Carter. 2016. *Operational Productivity and Performance in English NHS Acute Hospitals: Unwarranted Variations*. UK: Department of Health.
- Raut, S., N. U. I. Hossain, and S. A. Fazio. 2025. "Healthcare Resilience: Bibliometric and Network Analysis." *Enterprise Information Systems* 19 (10): 2464752. <https://doi.org/10.1080/17517575.2025.2464752>.
- Ren, J. 2026. "A Skill-Enhanced Retrieval-Augmented Generation Expert System Framework for Offshore Accident Analysis: Combining Fuzzy Inference, Dempster-Shafer Evidential Reasoning, and Expert-Gated Skill Adaptation." *Ocean Engineering* 364:126908. <https://doi.org/10.1016/j.oceaneng.2026.126908>.
- Ren, J., I. Jenkinson, H. S. Sii, J. Wang, D. L. Xu, and J. B. Yang. 2005. "An Offshore Safety Assessment Framework Using Fuzzy Reasoning and Evidential Synthesis Approaches." *Journal of Marine Engineering & Technology* 2005 (1): 3–16. <https://doi.org/10.1080/20464177.2005.11020183>.
- Ren, J., I. Jenkinson, and O. O. Tobora. 2026. "Grounded Expert Systems for Offshore Safety: Enhancing Rule-Based Risk Assessment with Retrieval-Augmented Generation (RAG) for Auditable Explanations." In *Proceedings of the International Conference on Electronics Technology (ICET)*, 29–31. May, Chengdu, China.
- Ren, J., D. L. Xu, J. B. Yang, I. Jenkinson, and J. Wang. 2009. "An Offshore Risk Analysis Method Using Fuzzy Bayesian Network." *Journal of Offshore Mechanics & Arctic Engineering* 131 (4): 1101–1112 doi:[10.1115/1.3124123](https://doi.org/10.1115/1.3124123).
- Ritchie, B., and C. Brindley. 2007. "Supply Chain Risk Management and Performance." *International Journal of Operations and Production Management* 27 (3): 303–322. <https://doi.org/10.1108/01443570710725563>.
- Samvedi, A., V. Jain, and T. S. Chan. 2013. "Quantifying Risks in a Supply Chain Through Integration of Fuzzy AHP and Fuzzy TOPSIS." *International Journal of Production Research* 51 (8): 2433–2442. <https://doi.org/10.1080/00207543.2012.741330>.
- Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.

- Shafiee, M., and T. Adedipe. 2023. "A Bayesian Network Model for the Probabilistic Safety Assessment of Offshore Wind Farm Decommissioning Operations." *Wind Engineering* 47 (1): 3–21 doi:[10.1177/0309524X221122569](https://doi.org/10.1177/0309524X221122569).
- Sheffi, Y. 2005. *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage*. Cambridge, MA: MIT Press.
- Sijtsma, K. 2009. "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha." *Psychometrika* 74 (1): 107–120. <https://doi.org/10.1007/s11336-008-9101-0>.
- Tang, C. S. 2006. "Robust Strategies for Mitigating Supply Chain Disruptions." *International Journal of Logistics: Research and Applications* 9 (1): 33–45. <https://doi.org/10.1080/13675560500405584>.
- Thompson, D. W., S. Raut, A. Sokolov, and N. U. I. Hossain. 2025. "Supply Chain Risk Management: A Review, Bibliometric, and Network Analysis Immediate Pre and Post COVID Era." *Enterprise Information Systems* 19 (12): 2568657. <https://doi.org/10.1080/17517575.2025.2568657>.
- Tucker, A. L. 2004. "The Impact of Operational Failures on Hospital Nurses and Their Patients." *Journal of Operations Management* 22 (2): 151–169. <https://doi.org/10.1016/j.jom.2003.12.006>.
- Wang, L. 2018. "Research on Risk Management for Healthcare Supply Chain in Hospital." PhD thesis, Liverpool, UK: Liverpool John Moores University.
- Wang, L., J. Ren, J. Wang, D. Paraskevakakis, and C. Morecroft. 2023. "An in-Depth Examination of Multi-Faceted Supply Chain Risk Management Strategies in the Healthcare Industry." In *7th International Conference on Transportation Information and Safety (ICTIS)*. Xi'an, China, 1247–1254. Piscataway, NJ, USA: IEEE. 4–6 August 2023.
- Yang, J. B. 2001. "Rule and Utility Based Evidential Reasoning Approach for MADA Under Uncertainties." *European Journal of Operational Research* 131 (1): 31–61. [https://doi.org/10.1016/S0377-2217\(99\)00441-5](https://doi.org/10.1016/S0377-2217(99)00441-5).
- Yang, J. B., and D. L. Xu. 2002. "On the Evidential Reasoning Algorithm for MADA Under Uncertainty." *IEEE Transactions on Systems, Man, and Cybernetics, Part A* 32 (3): 289–304 doi:[10.1109/TSMCA.2002.802746](https://doi.org/10.1109/TSMCA.2002.802746).
- Yang, L. H., Y. M. Wang, Y. G. Fu, J. Qin, C. Nugent, H. Wang, and L. Martínez. 2022. "Highly Explainable Cumulative Belief Rule-Based System with Interpretable Rule Weighting." *Knowledge-Based Systems* 240 (107895): 107805. <https://doi.org/10.1016/j.knosys.2021.107805>.
- Zadeh, L. A. 1965. "Fuzzy sets." *Information and Control* 8 (3): 338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X).
- Zadeh, L. A. 1979. *On the Validity of Dempster's Rule of Combination of Evidence*, Memo M 79/24. Berkeley: University of California.