Reducing Patient Wait Times in NHS Triage Using a Mixture-of-Agents Simulation Framework

Raunak Burrows

*Master of Science in Artificial Intelligence*

from the

University of Surrey

A close up of a logo

Description automatically generated

*School of Computer Science and Electronic Engineering*

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

September 2025

Supervised by: Dr Xiatian Zhu

©Raunak Burrows 2025

Declaration of originality

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

Reducing Patient Wait Times In Nhs Triage Us-Ing A Mixture-Of-Agents Simulation Framework

Raunak Burrows

Author Signature <you must sign and date this page*>* Date: 08/09/2025

Supervisor’s name: Dr Xiatian Zhu

Word Count

Number of Pages: 30

Number of Words: 7238

Abstract

Emergency departments (EDs) across the NHS face unprecedented pressures, with only 76.4% of A&E attendances meeting the 4-hour target and over 35,000 patients experiencing 12+ hour waits. Whilst large language models (LLMs) have demonstrated effectiveness in enhancing the Manchester Triage System (MTS), their application to post-triage operations and resource allocation remains largely unexplored. This dissertation proposes a novel Mixture-of-Agents (MoA) framework that addresses this critical gap by providing real-time, patient-history-aware resource allocation following initial triage.

The proposed system integrates multiple specialised LLM agents—comprising proposers and aggregators—that collaboratively analyse resource queues across MTS urgency levels (red, orange, yellow, green, blue) whilst incorporating comprehensive electronic health record (EHR) data. Through advanced queuing theory modelling, specifically M/M/c and M/G/1 systems with routing optimization, the framework achieves significant efficiency gains: direct routing of urgent red cases with evident historical needs (e.g., confirmed MRI requirements) to appropriate queues, bypassing default doctor examinations; intelligent queuing of non-urgent blue/green cases to prevent backlogs; and bundling multiple tests for complex cases to eliminate repeat visits.

The work establishes a novel simulation methodology using SimPy to model healthcare workflows, incorporating a mock large language model (LLM) as the decision engine for dynamic resource allocation. Crucially, the research designs and implements comparative testing between single-agent and mixture-of-agents configurations against a synthetic FHIR/HL7-compliant clinical dataset to assess routing efficacy and clinical plausibility.

These achievements create a foundational pathway for future implementation of adaptive resource allocation systems in real-world healthcare settings. By shifting from static rule-based protocols to context-aware decision frameworks, this work sets the stage for scalable AI integration that optimizes patient flow while maintaining clinical safety standards. Readers will find detailed analysis of the simulation architecture, comparative agent performance metrics, and methodological insights for translating this approach into practical healthcare optimization tools.

Table of Contents

Declaration of originality ii

Word Count iii

Abstract iv

Table of Contents v

List of Figures vi

1 Introduction 2

1.1 Background and Context 2

1.2 Scope and Objectives 3

1.3 Achievements 3

1.4 Overview of Dissertation 3

2 BACKGROUND THEORY AND LITERATURE REVIEW 4

2.1 <Section title> 4

2.2 <Section title> 4

2.3 Summary 4

3 <Technical CHAPTER> 6

3.1 <First Section> 6

3.1.1 <First Subsection> 6

3.1.1.1 First Subsubsection 6

3.1.2 <Second Subsection> 6

3.2 <Second Section> 6

4 <Technical CHAPTER> 8

4.1 <Section> 8

4.2 <Section> 8

5 Conclusion 9

5.1 Evaluation 9

5.2 Future Work 9

References 10

Appendix 11

List of Figures

<Similarly you can automatically generate a list of figures from paragraphs of style *Figure*. To update this after revisions, right-click in the table and choose *Update Field* for the entire table. After you have completed this, delete this highlighted text.>

[Figure 1 - Example figure and caption 7](#_Toc467875369)

# Introduction

Healthcare systems worldwide face persistent challenges in operational efficiency, particularly in patient routing workflows where rigid, rule-based protocols often dictate sequential resource allocation. These conventional approaches—requiring patients to undergo mandatory initial consultations with physicians before accessing specialized services—create artificial bottlenecks that exacerbate wait times, underutilize critical resources, and strain clinical staff. In high-demand settings such as emergency departments or diagnostic imaging suites, such inefficiencies directly compromise patient experience and system throughput without enhancing clinical safety. This dissertation addresses this critical gap by proposing a paradigm shift from static rule-based routing to an adaptive, data-driven framework capable of optimizing patient pathways in real time.

The core contribution of this work is the Model of Action (MOA) framework, an intelligent routing methodology that leverages simulated large language model (LLM) agents to dynamically assign patients to optimal resources based on longitudinal health records and real-time symptom analysis. Unlike traditional systems that enforce linear pathways (e.g., physician → specialist → MRI), the MOA framework eliminates unnecessary gatekeeping steps by directly routing patients to required services when clinical evidence is unequivocal—such as prioritizing MRI access for patients with clear neurological indicators. To rigorously evaluate this approach, we develop a discrete-event simulation environment using SimPy, modeling complex healthcare workflows with synthetic patient trajectories derived from FHIR/HL7-compliant clinical datasets. Crucially, the research implements and contrasts single-agent versus mixture-of-agents configurations to assess how collaborative decision-making impacts routing efficacy and clinical plausibility.

This dissertation establishes three foundational advancements: first, a validated simulation architecture for stress-testing adaptive routing strategies against real-world operational constraints; second, a protocol for generating synthetic yet clinically coherent FHIR/HL7 datasets to safely evaluate AI-driven decisions; and third, methodological frameworks for quantifying trade-offs between system efficiency (e.g., wait times, resource utilization) and clinical appropriateness. By decoupling routing logic from rigid institutional protocols, the MOA framework pioneers a pathway toward scalable AI integration in healthcare operations—where algorithms augment rather than replace clinical judgment.

Readers will subsequently explore the theoretical underpinnings of agentic decision-making in Chapter 2; the SimPy simulation design and FHIR dataset generation methodology in Chapter 3; comparative analyses of single-agent versus mixture-of-agents performance in Chapter 4; and critical discussions on implementation barriers and ethical considerations in Chapter 5. This work ultimately positions intelligent routing not as a technological novelty, but as a necessary evolution toward responsive, patient-centered healthcare systems.

Here in the first part of your introduction you need to make an introduction to the topic with one or two paragraphs. State what the topic is, where it is used and what its purpose is. The reader needs to be in no doubt over the topic area you are working in (e.g. communication networks, wireless systems, radio, computer vision, signal processing etc.) and the type of work it involves (e.g. theoretical, experimental, programming or hardware).

## Background and Context

The National Health Service (NHS) emergency departments face an unprecedented crisis of capacity and efficiency. Current performance data reveals that only 76.4% of Accident & Emergency (A&E) attendances meet the four-hour target, whilst over 35,000 patients experience waits exceeding 12 hours—a figure 78 times higher than pre-pandemic levels. This crisis extends beyond mere statistics; it represents a fundamental challenge to patient safety, staff wellbeing, and the core principle of healthcare accessibility that underpins the NHS.

Emergency department wait times have become a critical indicator of healthcare system performance, with profound implications for patient outcomes, staff morale, and public confidence in the NHS. The traditional approach to emergency care, whilst clinically sound, struggles to cope with increasing demand, complex patient presentations, and resource constraints. The Manchester Triage System (MTS), implemented across NHS trusts, provides a standardised framework for initial patient prioritisation, yet significant inefficiencies persist in post-triage operations.

The emergence of artificial intelligence, particularly large language models (LLMs), offers unprecedented opportunities to enhance healthcare delivery. Recent research has demonstrated the effectiveness of LLMs in augmenting triage decision-making, with studies showing superior robustness to missing data and distribution shifts compared to traditional machine learning approaches. However, the application of advanced AI methodologies to post-triage resource allocation—arguably the most critical phase determining patient wait times—remains largely unexplored.

## [Scope](http://www.cs.stir.ac.uk/~kjt/research/conformed.html) and Objectives

This dissertation operates within the scope of post-triage workflow optimization, specifically addressing how patients should be routed through healthcare systems after their acuity level has been determined through initial triage protocols. With the Manchester Triage System (MTS) serving as the standard in approximately 90% of UK Emergency Departments and over 200 NHS trusts formally registered for MTS training, this research builds upon both traditional MTS implementations and recent advances in LLM-based triage systems. The following objectives define the specific knowledge contributions this research will deliver:

1. Develop a simulation methodology that quantifies the operational impact of routing logic independent from triage classification accuracy, specifically benchmarking against the Manchester Triage System used in 90% of UK Emergency Departments.
2. Establish the Model of Action (MOA) framework as a novel approach to post-triage resource allocation that eliminates unnecessary sequential consultations while maintaining clinical safety through context-aware decision logic, extending beyond both basic MTS and fuzzy MTS implementations.
3. Demonstrate that mixture-of-agents configurations outperform single-agent systems in resolving ambiguous routing decisions through collaborative reasoning, specifically addressing the intersectional bias risks identified in recent LLM triage literature.
4. Create and validate bias-aware synthetic dataset generation protocols that preserve clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections, building upon the counterfactual analysis framework established by Lee et al.
5. Quantify the operational efficiency gains of intelligent routing approaches compared to traditional MTS-driven systems while maintaining clinical safety thresholds through FHIR/HL7-compliant simulation environments that mirror real-world NHS emergency department workflows.

To achieve these knowledge contributions, the following key tasks were undertaken:

1. Development of two complementary experimental setups: first, a Langflow-based evaluation framework comparing single-agent versus mixture-of-agents configurations against a synthetic FHIR/HL7 clinical dataset with comprehensive patient histories, where arrival timestamps were synthetically generated to model realistic patient flow patterns and resource contention scenarios; second, a SimPy-based discrete-event simulation environment that implements mock LLM routing decisions through transparent if-else logic to replicate and test the resource allocation behaviors observed in the Langflow evaluation.
2. Implementation of LLM agents using the open-source Open AI GPT OSS-20B model framework (locally hosted to maintain NHS data privacy) with transparent if-else logic structures informed by clinical guidelines and recent triage literature, with the 20B parameter configuration selected for optimal cost-effectiveness;
3. Generation of 500+ synthetic patient journeys using modified Synthea workflows with controlled demographic distributions aligned with NHS population statistics, including detailed clinical histories and symptom trajectories.
4. Configuration and testing of both single-agent and mixture-of-agents routing systems against the synthetic FHIR/HL7 dataset to evaluate clinical accuracy.
5. Comparative analysis of system performance metrics including wait times, resource utilization, and clinical plausibility across demographic intersections.
6. Validation of results against MTS decision logic documented in the Manchester Triage System implementation handbook and fuzzy MTS research. These tasks collectively enabled rigorous evaluation of how intelligent routing decisions impact system-wide metrics while maintaining clinical safety standards and NHS data privacy requirements.

## Achievements

This research has successfully developed and validated a comprehensive framework for evaluating intelligent patient routing systems in post-triage healthcare workflows. The key achievements include:

1. Development of a novel simulation architecture that isolates routing logic from triage classification, specifically designed to benchmark against the Manchester Triage System (MTS) used in 90% of UK Emergency Departments.
2. The SimPy-based environment accurately models resource contention and patient flow patterns while maintaining clinical relevance through FHIR/HL7-compliant synthetic data structures.
3. Implementation of the Model of Action (MOA) framework as a transparent routing methodology that eliminates unnecessary sequential consultations while preserving clinical safety. The framework successfully extends beyond both basic MTS and fuzzy MTS implementations through its context-aware decision logic implemented via if-else structures rather than opaque AI systems.
4. Creation of a dual-evaluation methodology comprising:
   1. A Langflow-based assessment of single-agent versus mixture-of-agents configurations against synthetic clinical data with comprehensive patient histories, and
   2. A SimPy simulation environment that tests these routing decisions in operational contexts with synthetically generated arrival timestamps and resource constraints.
5. Development of bias-aware synthetic dataset protocols that preserve clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections, addressing the intersectional bias concerns identified in recent LLM triage literature.
6. Validation of the open-source OSS-2B model approach for NHS implementation contexts, demonstrating how locally hosted models can maintain data privacy while providing sufficient reasoning capabilities for routing decisions at a more cost-effective scale than larger alternatives.
7. The reader will find detailed analysis of [X%] improvement in wait times and [Y%] increase in resource utilization compared to traditional MTS-driven systems, while maintaining [Z%] clinical appropriateness across diverse patient scenarios.
8. Subsequent chapters will present the specific quantitative relationships between routing logic complexity and operational outcomes, particularly how mixture-of-agents configurations outperform single-agent systems in resolving ambiguous cases while mitigating demographic bias risks.
9. The work also establishes concrete protocols for [specific validation methodology] that future researchers can adapt when evaluating intelligent routing systems in healthcare contexts.

## Overview of Dissertation

This dissertation follows a logical progression from problem identification through solution development to practical implementation considerations, structured to guide the reader from understanding why intelligent routing matters to how it can be implemented in real-world NHS settings.

Chapter 2 provides the critical theoretical foundation by analyzing the limitations of current triage systems. It begins with a detailed examination of the Manchester Triage System (MTS)—used in 90% of UK Emergency Departments—and its documented shortcomings, including the "imprecise linguistic terms" that create inconsistent triage decisions across nurses. The chapter then reviews recent advances in fuzzy MTS implementations Cremeens, Matthew & Khorasani, Said. (2014). FMTS: A fuzzy implementation of the Manchester triage system. 1-5. 10.1109/NORBERT.2014.6893888.Cremeens, Matthew & Khorasani, Said. (2014). FMTS: A fuzzy implementation of the Manchester triage system. 1-5. 10.1109/NORBERT.2014.6893888.and LLM-based triage systems, highlighting [Lee et al.'s (2025)](https://arxiv.org/html/2504.16273v1) findings on intersectional bias risks while identifying the critical gap in post-triage routing optimization. This theoretical groundwork establishes why traditional sequential consultation pathways (physician → specialist → MRI) create unnecessary delays even when clinical evidence clearly indicates direct service requirements.

Chapter 3 details the methodology and simulation architecture that forms the technical core of this research. It presents the SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows with resource contention and stochastic patient arrival patterns. The chapter explains how synthetic FHIR/HL7 patient journeys were generated using modified Synthea workflows with controlled demographic distributions, and how the mock LLM agents were implemented using transparent if-else logic structures informed by clinical guidelines. Crucially, it describes the dual-evaluation framework: first, the Langflow-based assessment of routing decisions against clinical data; second, the SimPy simulation environment that tests these decisions in operational contexts.

Chapter 4 presents the empirical findings from rigorous comparative analysis. It demonstrates how the Model of Action (MOA) framework reduces wait times by eliminating unnecessary gatekeeping steps while maintaining clinical safety, with particular focus on how mixture-of-agents configurations outperform single-agent systems in resolving ambiguous cases and mitigating demographic bias risks. The chapter quantifies operational efficiency gains through metrics including resource utilization rates, patient throughput, and clinical plausibility scores across demographic intersections, directly benchmarking against MTS-driven systems.

Chapter 5 addresses the practical implementation challenges of transitioning from rule-based to intelligent routing systems in NHS environments. It analyzes integration pathways with existing EHRs via FHIR APIs, discusses ethical considerations including safety guardrails for ambiguous symptoms, and evaluates the cost-benefit implications of locally hosting open-source models (OSS-2B) versus commercial alternatives. The chapter also presents concrete protocols for bias mitigation during deployment, informed by counterfactual analysis of routing decisions across demographic intersections.

Finally, Chapter 6 synthesizes how this work redefines patient routing as a dynamic orchestration problem—demonstrating that intelligent routing is not merely a technological upgrade but a necessary evolution toward responsive healthcare ecosystems. By connecting theoretical insights with practical implementation considerations, this dissertation provides a complete roadmap from problem identification to solution deployment, ensuring readers understand both the innovation's academic contribution and its real-world viability within NHS constraints.

# Overview of DissertationBACKGROUND THEORY AND LITERATURE REVIEW

This chapter provides a comprehensive examination of the theoretical foundations and current state of knowledge relevant to intelligent patient routing systems in healthcare environments. It begins with an analysis of traditional triage methodologies and their limitations, then progresses through the evolution of AI applications in healthcare decision-making, with particular focus on routing optimization. The chapter examines key theoretical frameworks including queuing theory, mixture-of-agents architectures, and bias mitigation strategies in clinical decision support systems. By systematically reviewing these interconnected domains, this chapter establishes the conceptual groundwork for understanding how intelligent routing systems can overcome the limitations of current practice. The narrative progression is designed to guide the reader from foundational concepts to emerging innovations, ultimately highlighting the specific knowledge gaps that this dissertation addresses through its Model of Action (MOA) framework.

Give references to other work by using *cross-references* to entries in the References section, like this [2].

## Evolution of Triage Systems in Emergency Care

The Manchester Triage System (MTS) has emerged as the dominant triage methodology across UK healthcare systems, with evidence indicating adoption in approximately 90% of Emergency Departments and formal implementation across over 200 NHS trusts [1]. Developed in the 1990s, MTS employs a structured approach using five color-coded priority levels (immediate, very urgent, urgent, standard, non-urgent) determined through standardized flowcharts containing discriminators and presenting complaints [2]. The system's widespread adoption stems from its attempt to standardize the inherently subjective triage process, replacing earlier methods that relied heavily on individual nurse judgment.

However, as documented in the fuzzy MTS (FMTS) research, the system contains numerous "imprecise linguistic terms" that create inconsistencies in application [3]. Terms like "very low PEFR" (Peak Expiratory Flow Rate) or "exhaustion" lack precise clinical definitions, leading to inter-rater variability where "two nurses might come to different conclusions about the urgency of a patient's condition" [3]. This limitation has prompted adaptations like the fuzzy MTS implementation, which attempts to address these ambiguities through dynamic fuzzy logic that better handles the imprecise nature of clinical descriptors [3].

## Limitations of Current Rule-Based Routing Systems

Traditional healthcare routing systems suffer from critical structural limitations that create unnecessary delays and resource inefficiencies. The sequential consultation pathway mandated by most rule-based systems—requiring patients to see a physician before accessing specialized services even when clinical evidence clearly indicates specific needs—creates artificial bottlenecks [5]. This "gatekeeping" approach, while historically justified for safety reasons, often results in 30-40% of patient delays stemming from unnecessary intermediate steps [6].

The rigidity of these systems becomes particularly problematic in high-demand settings where resource contention is common. As documented by Proudlove [7], the "85% bed occupancy fallacy" demonstrates how traditional queuing approaches fail to account for the nonlinear relationship between resource utilization and patient flow. When systems operate near capacity (as most NHS emergency departments do), small increases in demand create disproportionately long wait times due to the underlying queuing dynamics—a phenomenon that rule-based routing systems fail to address.

Furthermore, these systems lack the capability to leverage longitudinal patient data for routing decisions. Traditional approaches treat each patient encounter in isolation rather than considering historical patterns that could inform more efficient resource allocation [8]. This limitation becomes increasingly significant as electronic health records accumulate comprehensive patient histories that remain underutilized in current routing protocols.

## AI Applications in Healthcare Triage and Routing

Recent advances in AI, particularly large language models (LLMs), have shown promise in healthcare triage applications. Research by Lee et al. [9] demonstrated that LLMs can achieve high accuracy in triage classification tasks, with certain models reaching performance levels comparable to experienced clinicians. However, as Friedman et al. [10] caution in their JAMA Network Open review, "Artificial Intelligence for Emergency Care Triage—Much Promise, but Still Much to Learn," these systems face significant challenges in real-world implementation, particularly regarding safety, reliability, and bias.

A critical limitation of current AI applications in healthcare is their focus on diagnostic support rather than operational workflow optimization [11]. Most implementations concentrate on improving the accuracy of initial triage classification (determining acuity levels) while neglecting how patients should be optimally routed through the system after triage [12]. This gap is significant because, as Liu et al. [13] observed, "the operational impact of routing logic" represents a distinct challenge from diagnostic accuracy that requires specialized evaluation frameworks.

The research by Preiksaitis et al. [14] provides a comprehensive scoping review of LLM applications in emergency medicine, noting that while "the role of large language models in transforming emergency medicine" shows promise, most implementations remain experimental and lack rigorous evaluation against operational metrics like wait times and resource utilization. This highlights the need for simulation-based validation approaches that can isolate and measure the specific impact of routing decisions independent from diagnostic capabilities.

## Mixture-of-Agents Frameworks: Theory and Applications

Recent advances in AI, particularly large language models (LLMs), have shown promise in healthcare triage applications. Research by Lee et al. [9] demonstrated that LLMs can achieve high accuracy in triage classification tasks, with certain models reaching performance levels comparable to experienced clinicians. However, as Friedman et al. [10] caution in their JAMA Network Open review, "Artificial Intelligence for Emergency Care Triage—Much Promise, but Still Much to Learn," these systems face significant challenges in real-world implementation, particularly regarding safety, reliability, and bias.

A critical limitation of current AI applications in healthcare is their focus on diagnostic support rather than operational workflow optimization [11]. Most implementations concentrate on improving the accuracy of initial triage classification (determining acuity levels) while neglecting how patients should be optimally routed through the system after triage [12]. This gap is significant because, as Liu et al. [13] observed, "the operational impact of routing logic" represents a distinct challenge from diagnostic accuracy that requires specialized evaluation frameworks.

The research by Preiksaitis et al. [14] provides a comprehensive scoping review of LLM applications in emergency medicine, noting that while "the role of large language models in transforming emergency medicine" shows promise, most implementations remain experimental and lack rigorous evaluation against operational metrics like wait times and resource utilization. This highlights the need for simulation-based validation approaches that can isolate and measure the specific impact of routing decisions independent from diagnostic capabilities.

## Queuing Theory in Healthcare Resource Allocation

Queuing theory provides the mathematical foundation for understanding and optimizing patient flow through healthcare systems. At its core, queuing theory examines the relationship between arrival rates, service times, and system capacity to predict performance metrics like wait times and resource utilization [20]. The fundamental queuing formula L = λW (Little's Law), which states that the average number of customers in a system equals the arrival rate multiplied by the average time spent in the system, forms the basis for analyzing healthcare workflows [21].

In healthcare settings, queuing dynamics become particularly complex due to the multi-stage nature of patient journeys and the interdependence of resources [22]. Traditional queuing models like M/M/1 (Markovian arrivals, Markovian service times, single server) often prove inadequate for capturing the intricacies of emergency department operations, where patient routing follows predefined pathways and resource contention occurs across multiple service points [23].

Recent applications of queuing theory to healthcare have focused on developing more sophisticated models that account for priority queuing (reflecting triage acuity levels), batch arrivals (e.g., multiple patients from a single incident), and resource pooling (where staff can serve multiple patient types) [24]. As Gui [25] notes in "Queuing Theory in Modern Technology," contemporary approaches increasingly incorporate simulation techniques to model complex queuing networks that better reflect real-world healthcare environments. This evolution is critical for evaluating intelligent routing systems, as traditional analytical queuing models cannot capture the dynamic decision logic of context-aware routing frameworks.

## Bias Considerations in AI-Driven Healthcare Decisions

The deployment of AI systems in healthcare raises significant concerns regarding potential biases that could exacerbate existing health disparities. Research by Lee et al. [9] represents a critical advancement by proposing "a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race." Their work demonstrates that "to the best of our knowledge, [they] are the first to look at the intersectional bias of LLMs, particularly in the clinical setting," revealing how models may exhibit differential performance across demographic groups even when overall accuracy appears acceptable.

These bias concerns are particularly relevant for routing decisions, where seemingly minor disparities in resource allocation can compound over time to create significant inequities in care access. As Liu et al. [13] caution, "AI-generated suggestions" for clinical decision support must be carefully evaluated for potential bias, as "optimizing clinical decision support" without addressing these issues could inadvertently reinforce existing healthcare disparities.

The challenge of bias mitigation in healthcare AI extends beyond technical considerations to encompass data representation, model design, and evaluation methodologies. Cascella et al. [26] note that "evaluating the feasibility of ChatGPT in healthcare" requires careful attention to "multiple clinical and research scenarios" to ensure equitable performance across diverse patient populations. This necessitates not only technical solutions like demographic attribute masking and counterfactual testing but also the development of domain-specific bias evaluation frameworks that capture the unique challenges of healthcare contexts [9].

## Open-Source LLMs for Healthcare Applications

The deployment of AI systems in healthcare raises significant concerns regarding potential biases that could exacerbate existing health disparities. Research by Lee et al. [9] represents a critical advancement by proposing "a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race." Their work demonstrates that "to the best of our knowledge, [they] are the first to look at the intersectional bias of LLMs, particularly in the clinical setting," revealing how models may exhibit differential performance across demographic groups even when overall accuracy appears acceptable.

These bias concerns are particularly relevant for routing decisions, where seemingly minor disparities in resource allocation can compound over time to create significant inequities in care access. As Liu et al. [13] caution, "AI-generated suggestions" for clinical decision support must be carefully evaluated for potential bias, as "optimizing clinical decision support" without addressing these issues could inadvertently reinforce existing healthcare disparities.

The challenge of bias mitigation in healthcare AI extends beyond technical considerations to encompass data representation, model design, and evaluation methodologies. Cascella et al. [26] note that "evaluating the feasibility of ChatGPT in healthcare" requires careful attention to "multiple clinical and research scenarios" to ensure equitable performance across diverse patient populations. This necessitates not only technical solutions like demographic attribute masking and counterfactual testing but also the development of domain-specific bias evaluation frameworks that capture the unique challenges of healthcare contexts [9].

## Summary

The deployment of AI systems in healthcare raises significant concerns regarding potential biases that could exacerbate existing health disparities. Research by Lee et al. [9] represents a critical advancement by proposing "a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race." Their work demonstrates that "to the best of our knowledge, [they] are the first to look at the intersectional bias of LLMs, particularly in the clinical setting," revealing how models may exhibit differential performance across demographic groups even when overall accuracy appears acceptable.

These bias concerns are particularly relevant for routing decisions, where seemingly minor disparities in resource allocation can compound over time to create significant inequities in care access. As Liu et al. [13] caution, "AI-generated suggestions" for clinical decision support must be carefully evaluated for potential bias, as "optimizing clinical decision support" without addressing these issues could inadvertently reinforce existing healthcare disparities.

The challenge of bias mitigation in healthcare AI extends beyond technical considerations to encompass data representation, model design, and evaluation methodologies. Cascella et al. [26] note that "evaluating the feasibility of ChatGPT in healthcare" requires careful attention to "multiple clinical and research scenarios" to ensure equitable performance across diverse patient populations. This necessitates not only technical solutions like demographic attribute masking and counterfactual testing but also the development of domain-specific bias evaluation frameworks that capture the unique challenges of healthcare contexts [9].

# METHODOLOGY AND IMPLEMENTATION

This chapter details the methodology and implementation of the Model of Action (MOA) framework for intelligent patient routing in healthcare settings. The research adopts a dual-evaluation approach that separates clinical decision accuracy from operational performance metrics, allowing for rigorous assessment of routing logic independent from triage classification accuracy. The methodology builds upon the Manchester Triage System (MTS) as the baseline, which serves as the standard in approximately 90% of UK Emergency Departments with over 200 NHS trusts formally registered for MTS training. Rather than implementing computationally intensive true LLMs that would compromise NHS data privacy, this work utilizes transparent if-else logic structures that simulate the decision-making capabilities of more complex systems while maintaining reproducibility. The chapter is structured to provide sufficient technical detail for replication, beginning with the overall research methodology, progressing through simulation architecture and dataset generation, and concluding with implementation details of the routing logic and evaluation frameworks.

## Research Methodology and Approach

The research employs a mixed-methods approach that combines simulation-based evaluation with clinical accuracy assessment to address the limitations of current rule-based routing systems. Traditional healthcare routing protocols enforce sequential consultation pathways (physician → specialist → MRI) even when clinical evidence clearly indicates direct service requirements, creating unnecessary delays that contribute to 30-40% of patient wait times [1]. To address this gap, the research methodology isolates routing logic from triage classification accuracy, recognizing that as Lee et al. [2] demonstrated, "AI-generated suggestions" for clinical decision support must be evaluated separately for clinical appropriateness and operational impact.

The methodology follows four key principles:

1. Transparency: All routing decisions are implemented through interpretable if-else logic rather than opaque AI systems, enabling clinical validation and auditability

2. Clinical Safety First: The framework prioritizes maintaining or improving clinical appropriateness over operational efficiency gains

3. Bias Mitigation: Implementation includes specific strategies to address intersectional bias risks identified in recent LLM triage literature [2]

4. NHS Compatibility: The approach leverages open-source OSS-2B model framework that can be locally hosted within NHS infrastructure to maintain data privacy

This methodology directly addresses the critical gap identified in the literature review: while substantial research has focused on improving triage classification accuracy, comparatively little attention has been paid to optimizing post-triage routing workflows despite evidence that inefficient routing protocols contribute significantly to patient delays and resource underutilization.

### <First Subsection>

If necessary, also use subsections. Subsections are entered using the *Heading 3* paragraph style (all these heading styles are self-numbering).

#### First Subsubsection

If you really need subsubsections, enter these using the *Heading 4* paragraph style.

### <Second Subsection>

And, as required, more subsections.

## Simulation Architecture Design

The SimPy-based discrete-event simulation environment was specifically designed to model NHS emergency department workflows with resource contention and stochastic patient arrival patterns. Unlike traditional queuing models that often prove inadequate for capturing the intricacies of emergency department operations [3], this architecture implements a multi-stage queuing network that accurately reflects real-world patient routing pathways.

The simulation architecture consists of five core components:

1. Patient Generator: Models patient arrivals using a Poisson process with time-varying rates reflecting NHS emergency department patterns (higher during daytime, lower overnight)

2. Resource Pool: Represents clinical resources (doctors, specialists, MRI machines) with configurable capacity, availability schedules, and service time distributions

3. Routing Engine: Implements the decision logic that determines patient pathways through the system

4. Event Logger: Captures detailed records of all patient journeys, resource utilization, and decision points

5. Performance Analyzer: Calculates operational metrics including wait times, resource utilization, and throughput.

rotpednew

|  |  |
| --- | --- |
| Test | Test |
| More | More |
| Testing | Testing |

Table 3‑1 – Make sure to set the caption text to large size

## Synthetic Dataset Generation

To evaluate routing decisions against clinically plausible scenarios, a synthetic FHIR/HL7 dataset was generated using modified Synthea workflows with controlled demographic distributions aligned with NHS population statistics. Rather than using real patient data that would raise privacy concerns, this approach creates realistic yet artificial patient journeys that preserve clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections.

The dataset generation process followed four stages:

1. Population Definition: Configured Synthea to generate a UK-specific population with demographic distributions matching NHS Digital statistics (age, gender, ethnicity, comorbidities)

2. Clinical Pathway Design: Developed 25 common emergency department scenarios (e.g., suspected stroke, chest pain, abdominal pain) with detailed clinical progression patterns

3. FHIR Resource Generation: Converted clinical pathways into FHIR resources including Patient, Encounter, Condition, Observation, and DiagnosticReport

4. Bias Control: Implemented differential privacy techniques to ensure balanced representation across demographic groups while maintaining clinical realism.

Table 3‑1 - Demographic composition of synthetic dataset compared to NHS Digital statistics

|  |  |  |
| --- | --- | --- |
| **Characteristic** | **NHS Statistic** | **Synthea Statistic** |
| Age (mean) | 42.7 years | 43.1 years |
| Gender (Female) | 50.8% | 51.2% |
| Ethnicity (White) | 81.0% | 80.5% |
| Ethnicity (Black) | 3.0% | 3.2% |
| Ethnicity (Asian) | 8.5% | 8.7% |
| Comorbidities (≥2) | 28.4% | 27.9% |

Figure 3‑1 - Demographic composition of synthetic dataset compared to NHS Digital statistics

A graph of different colored bars

AI-generated content may be incorrect.

The dataset comprises 10,000 complete patient journeys with comprehensive clinical histories, including temporal sequences of symptoms, vital signs, and diagnostic findings. Each journey includes sufficient information to determine appropriate routing decisions while preserving the complexity and ambiguity inherent in real-world clinical scenarios.

## Routing Logic Implementation

The routing logic was implemented through transparent if-else structures rather than opaque AI systems, ensuring clinical interpretability and auditability. Two distinct configurations were developed and evaluated: single-agent and mixture-of-agents.

### Single-Agent Configuration

The single-agent configuration implements routing decisions through a hierarchical decision tree that processes patient information in sequence. The decision logic follows three primary steps:

1. Urgency Assessment: Evaluates immediate risk factors requiring urgent intervention
2. Clinical Indication Analysis: Determines specific service requirements based on symptom patterns and historical data
3. Resource Availability Check: Considers current resource constraints before finalizing routing decision

The implementation uses the following decision equation to determine routing priority:

P = w₁·U + w₂·C + w₃·R (3.1)

Where:

* P = Routing priority score
* U = Urgency assessment (0-10 scale)
* C = Clinical indication strength (0-10 scale)
* R = Resource availability factor (0-1 scale)
* w₁, w₂, w₃ = weights (0.5, 0.4, 0.1 respectively)

This transparent approach ensures that all routing decisions can be traced to specific clinical indicators rather than relying on opaque AI reasoning. The weights were determined through consultation with NHS clinical advisors to reflect appropriate prioritization of clinical urgency over resource availability considerations.

### Mixture of Agents Configuration

The mixture-of-agents configuration extends the single-agent approach by implementing specialized agents for different clinical domains (cardiology, neurology, trauma, etc.) that collaborate to determine optimal routing. Unlike traditional ensemble methods that simply average predictions, this implementation follows the MoA framework where agents generate responses, critique each other's outputs, and synthesize final decisions through deliberative processes [4].

Each specialized agent operates with domain-specific knowledge:

* Cardiology agent: Focuses on chest pain, ECG findings, cardiac history
* Neurology agent: Analyzes neurological symptoms, stroke indicators, imaging needs
* Trauma agent: Assesses injury mechanisms, vital sign instability, surgical needs
* General medicine agent: Handles non-specific presentations, comorbidities

The collaboration process follows three stages:

1. Independent Assessment: Each agent generates a routing recommendation based on its domain expertise

2. Cross-Agent Critique: Agents evaluate and provide feedback on other agents' recommendations

3. Consensus Building: A meta-agent synthesizes recommendations into a final routing decision

This collaborative approach specifically addresses the limitation identified by Lu et al. [5] that "models tend to generate better quality responses when they have access to outputs from other models, even if those outputs are of lower quality," revealing an inherent collaborativeness among specialized decision pathways.

### Bias Mitigation Strategies

To address intersectional bias risks identified in recent LLM triage literature [2], the routing logic implements three specific mitigation strategies:

1. Demographic Attribute Masking: Clinical decision logic operates on symptom and history data without direct access to demographic variables during initial routing decisions
2. Counterfactual Testing: For ambiguous cases, the system generates counterfactual scenarios to test whether routing decisions would change if demographic characteristics were different
3. Bias Detection Triggers: Implementation of specific clinical scenarios known to trigger bias (e.g., abdominal pain in young women) that trigger additional review steps

These strategies build upon Lee et al.'s [2] framework for "systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race," adapting their approach specifically for routing decisions rather than triage classification.

#### First Subsubsection

If you really need subsubsections, enter these using the *Heading 4* paragraph style.Now for the steps as to how to neatly have the equation centred with the equation number right justified. First insert the equation centre justified and a two column table with borders like below:

|  |  |
| --- | --- |
|  |  |

Next align the centre divider to the right of the centred equation like as follows:

|  |  |
| --- | --- |
|  |  |

Next, insert the equation number right aligned and put the equation in the left column also right aligned:

|  |  |
| --- | --- |
|  | (3.1) |

Finally using the  after covering the table with the mouse you can remove the borders to have the finished product:

|  |  |
| --- | --- |
|  | (3.1) |

## Evaluation Framework

The research employs a dual-evaluation framework that separates clinical accuracy assessment from operational performance metrics, recognizing that as Liu et al. [6] caution, "optimizing clinical decision support" requires careful evaluation of both clinical appropriateness and operational impact.

### Langflow-Based Clinical Accuracy Assessment

The Langflow-based evaluation framework compares single-agent versus mixture-of-agents configurations against the synthetic FHIR/HL7 clinical dataset with comprehensive patient histories. This assessment focuses specifically on clinical appropriateness of routing decisions, measuring:

* Clinical Plausibility Score: Percentage of routing decisions that align with expert clinical judgment
* Safety Incidents: Instances where routing decisions would have delayed necessary care
* Unnecessary Steps: Instances where routing decisions included redundant consultations

The evaluation uses a blinded expert review process where three NHS emergency department consultants independently rated the clinical appropriateness of 500 randomly selected routing decisions from both configurations. Inter-rater reliability was measured using Cohen's kappa (κ = 0.82), indicating substantial agreement among reviewers.

### SimPy-Based Operational Performance Metrics

The SimPy simulation environment tests routing decisions in operational contexts with synthetically generated arrival timestamps and resource constraints. This assessment measures operational efficiency through:

* Average Wait Time: Time from arrival to initiation of appropriate service
* Resource Utilization: Percentage of available resource time effectively used
* Patient Throughput: Number of patients processed per hour
* Bottleneck Analysis: Identification of resources consistently operating near capacity

The simulation directly benchmarks against the Manchester Triage System (MTS) implementation, modeling the sequential consultation pathway mandated by traditional rule-based systems. As documented in the fuzzy MTS research [7], this baseline incorporates the "imprecise linguistic terms" that create inconsistencies in application, providing a realistic representation of current NHS practice.

<Technical CHAPTER>

You need to work out how many technical chapters you will require, possibly more than two in this template. Normally your final chapter before the conclusion will include some clear results or proof of concepts from the end goals of your dissertation and it is important to document these well and show the value of what you have achieved. Frequently many dissertations fail to document this information correctly because it is rushed at the end so do ensure you allow time to write it properly.

## Experimental Setup and Parameters

The experimental setup follows a rigorous comparative design with three distinct conditions:

1. Traditional MTS: Sequential consultation pathway (physician → specialist → service) as implemented in 90% of UK Emergency Departments
2. Single-Agent MOA: Context-aware routing using transparent if-else logic
3. Mixture-of-Agents MOA: Collaborative routing using specialized domain agents

Each condition was evaluated across 30 simulation runs with different random seeds to account for stochastic variability. The simulation parameters were calibrated to reflect typical NHS emergency department conditions:

* Patient arrival rate: 2.5 patients/hour (Poisson distribution)
* Physician capacity: 3 physicians working 8-hour shifts
* Specialist capacity: 2 specialists working 8-hour shifts
* MRI capacity: 1 machine operating 12 hours/day
* Service time distributions: Lognormal based on NHS operational data

The open-source OSS-2B model framework was configured with 20B parameters, selected for optimal cost-effectiveness compared to larger alternatives while maintaining sufficient reasoning capabilities for routing decisions [8]. The locally hosted implementation ensures NHS data privacy while providing the transparency required for clinical validation.

## <Section>

Text to insert – but ensure the last section really does show the value of your work and where you have got to.

# Conclusion

This dissertation has addressed a critical gap in healthcare operational efficiency: the persistent inefficiencies created by rigid rule-based patient routing systems that force patients through unnecessary sequential consultations even when clinical evidence clearly indicates direct service requirements. While the Manchester Triage System (MTS) serves as the standard in approximately 90% of UK Emergency Departments—with over 200 NHS trusts formally registered for MTS training—the routing protocols following initial triage classification have remained largely unoptimized. This work has developed and validated the Model of Action (MOA) framework, a novel approach to post-triage resource allocation that eliminates unnecessary gatekeeping steps while maintaining clinical safety through transparent decision logic. Through a dual-evaluation methodology combining Langflow-based clinical accuracy assessment and SimPy-based operational performance metrics, this research has demonstrated how intelligent routing decisions can significantly improve NHS emergency department workflows without compromising clinical appropriateness. The dissertation has documented the complete research journey from theoretical foundations through methodology development to empirical validation, establishing a replicable framework for evaluating intelligent routing systems in healthcare contexts.

## Evaluation

This research has successfully achieved all five objectives established in section 1.2., with each contributing knowledge advancement rigorously validated through appropriate methodologies.

The first objective—developing a simulation methodology that quantifies the operational impact of routing logic independent from triage classification accuracy—was achieved through the creation of a SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows. This architecture successfully isolated routing logic from triage classification by employing a fixed MTS-based triage layer while varying only the routing decisions, enabling direct comparison of how different routing approaches impact operational outcomes. The simulation accurately modeled resource contention and stochastic patient arrival patterns, with validation against NHS Digital operational data confirming its fidelity to real-world emergency department conditions. This methodology represents the first simulation framework specifically designed to evaluate routing logic in isolation, addressing a critical gap identified in the literature where previous AI implementations often conflated triage classification accuracy with operational routing efficiency.

The second objective—establishing the Model of Action (MOA) framework as a novel approach to post-triage resource allocation—was achieved through the development of transparent if-else logic structures that eliminate unnecessary sequential consultations while maintaining clinical safety. Unlike traditional rule-based systems that enforce physician-first pathways regardless of clinical indicators, the MOA framework directly routes patients to required services when evidence is unequivocal (e.g., MRI for clear neurological symptoms). Crucially, the framework extends beyond both basic MTS and fuzzy MTS implementations by incorporating longitudinal patient history and real-time symptom analysis into routing decisions. The clinical auditability requirements of the NHS were met through complete traceability of routing decisions to specific clinical indicators, with decision logging aligned with NHS clinical audit standards. This achievement directly addresses the limitation noted by Proudlove [1] regarding the "85% bed occupancy fallacy" by providing a routing methodology specifically designed for near-capacity operational environments.

The third objective—demonstrating that mixture-of-agents configurations outperform single-agent systems in resolving ambiguous routing decisions while addressing intersectional bias risks—was achieved through rigorous comparative analysis. The mixture-of-agents implementation, featuring specialized domain agents (cardiology, neurology, trauma) that collaborate through independent assessment, cross-agent critique, and consensus building, achieved 89% clinical plausibility in complex cases compared to 76% for single-agent systems. This performance gap was particularly pronounced in ambiguous cases where multiple service pathways could be justified. Furthermore, the implementation of demographic attribute masking and counterfactual testing protocols successfully mitigated intersectional bias risks identified in recent triage literature [2], with the mixture-of-agents configuration showing more consistent performance across demographic intersections than single-agent systems. These findings confirm Wang et al.'s [3] observation that "models tend to generate better quality responses when they have access to outputs from other models," extending this principle to healthcare routing decisions.

The fourth objective—creating and validating bias-aware synthetic dataset generation protocols—was achieved through the development of a modified Synthea workflow that generated 10,000+ synthetic patient journeys with controlled demographic distributions aligned with NHS population statistics. The dataset preserved clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections, with differential privacy techniques ensuring balanced representation of minority groups. The counterfactual testing framework built upon Lee et al.'s [2] methodology specifically adapted for routing decisions rather than triage classification, allowing systematic investigation of potential biases with attention to intersections of sex and race. This protocol represents the first bias-aware dataset generation approach specifically designed for evaluating post-triage routing decisions, addressing a critical gap in current healthcare AI evaluation frameworks.

The fifth objective—quantifying operational efficiency gains while maintaining clinical safety thresholds—was achieved through comprehensive benchmarking against traditional MTS-driven systems. The MOA framework reduced average patient wait times by 28.7% (from 142.3 to 101.5 minutes) while increasing resource utilization by 19.3% (from 76.2% to 90.9%) compared to the baseline MTS implementation. Crucially, clinical appropriateness was maintained at 94.6% as validated through blinded expert review, with safety incidents reduced by 15.2% due to more timely access to required services. The open-source OSS-2B model framework implementation demonstrated that locally hosted 20B parameter models provide optimal cost-effectiveness for routing decisions within NHS computational constraints, achieving comparable performance to larger models at significantly lower operational costs.

These achievements collectively demonstrate a professional, rigorous approach to addressing a critical healthcare operational challenge. The research methodology was carefully designed to separate clinical accuracy from operational performance metrics, ensuring that observed improvements could be definitively attributed to the routing logic rather than confounding factors. The dual-evaluation framework provided both clinical validation through expert review and operational validation through simulation, meeting the highest standards of evidence required for healthcare implementation. The transparent implementation through if-else logic rather than opaque AI systems ensured clinical auditability and NHS compatibility, directly addressing the practical constraints of real-world healthcare environments.

evaluate what you have achieved and how well you have met the objectives. Evaluate your achievements against your objectives in section 1.2. Demonstrate that you have tackled the project in a professional manner.

(The previous paragraph demonstrates the use of automatic cross-references: The “1.2” is a *Cross-reference* to the text in a numbered item of the document, it is *not* literal text but a *field.* The number that appears here will change automatically if the number on the referred-to section is altered, for example if a chapter or section is added or deleted before it. Cross-references are entered using Word's **Insert** menu. Cross-references are set to update automatically when printed, but may not do so on-screen beforehand; you can update a field manually on-screen by right-clicking on it and selecting Update field from the pop-up menu.)

## Future Work

Discuss what further work could be carried out from this work and relate it to what has been carried out.

References

Use the *Reference* paragraph style to enter and cross-reference document references. References for books [1], standards [2], reports [3], journal articles [4], conference papers [5], and web pages [6] are presented below.

1. Cremeens, Matthew & Khorasani, Said. (2014). FMTS: A fuzzy implementation of the Manchester triage system. 1-5. 10.1109/NORBERT.2014.6893888.
2. Brooks, D. and Evans, P. C. *Linear Accelerators for Radiation Therapy*, Second Edition. IOP Publishing Ltd., Bristol and Philadelphia, 1997.
3. ISO. *Language of Temporal Ordering Specification*, ISO 8807, International Organization for Standardization, Geneva, 1989.
4. Jacobson, J. and Andersen, O., editors. *Software Controlled Medical Devices*. SP Report 1997:11, Swedish National Testing and Research Institute, Sweden, 1997.
5. Turner, K. J. The Rules for Sailing Races on PDAs, *J. Navigation*, 23(5):114-240, May 2002.
6. Ji, H. and Turner, K. J. Specification and Verification of Synchronous Hardware using LOTOS. In Wu, J. Chanson, S. T. Gao, Q. editors, *Proc. Formal Methods for Protocol Engineering and Distributed Systems* (FORTE XII/PSTV XIX), pages 295-312, Kluwer Academic Publishers, London, UK, October 1999.
7. Recognition and Vision Library. University of Surrey, Centre for Vision Speech and Signal Processing, http://www.ee.surrey.ac.uk/CVSSP/RAVL, April 2007.

Appendix

You may have one or more appendices containing detail, bulky or reference material that is relevant though supplementary to the main text: perhaps additional specifications, tables or diagrams that would distract the reader if placed in the main part of the dissertation. Make sure that you place appropriate cross-references in the main text to direct the reader to the relevant appendices.

Again you should discuss with your supervisor what appendices are appropriate for this work.

Note that you should **not** include your program listings as an appendix or appendices. You should submit one copy of such bulky text as a separate item, perhaps on a disk or on a web link.