Generating synthetic electronic health record data using agent-based models to evaluate machine learning robustness under mass casualty incidents

Roben Delos Reyes RDELOSREYES@STUDENT.UNIMELB.EDU.AU
School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia

Daniel Capurro DCAPURRO@UNIMELB.EDU.AU
School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia
Department of Medicine, The University of Melbourne, Parkville, Victoria, Australia

Nicholas Geard NGEARD@UNIMELB.EDU.AU
School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia

Abstract

Machine learning (ML) models in healthcare are typically evaluated using curated real-world electronic health record (EHR) data. A key limitation of such evaluations is that they may fail to assess the robustness of ML models to changes in the data at deployment, which is a common issue because EHR data used for ML model development cannot capture all such changes. Mass casualty incidents (MCIs) caused by disasters are critical instances where this will be an issue, as they induce rare, uncertain, and novel changes to routine system conditions. Because real-world EHR data from MCIs are often limited or unavailable, assessing ML robustness under such conditions before deployment remains challenging. Here, we propose an agent-based modelling approach for generating synthetic EHR data to evaluate the robustness of ML models under MCI scenarios. We use real-world EHR data to develop and calibrate an agent-based model (ABM) of an emergency department (ED) that explicitly models patient arrivals, resource capacity, and clinical workflow. By changing these system conditions to reflect plausible MCI scenarios, the ED model generates synthetic versions of the real-world EHR data that exhibit shifts in system behaviour. Using these synthetic data, we test ML models for predicting length of stay. We observed consistent declines in recall under MCI conditions relative to baseline system conditions, resulting in an increase in the number of patients with prolonged length of stay that were missed by the ML models. These results highlight the impact of changes in system conditions on patient outcomes, EHR data, and ML model performance. Our work establishes ABM-based synthetic EHR data generation as a proactive and systematic approach for evaluating the robustness of ML models under MCI or other system conditions not captured in real-world EHR data, supporting the safer and more effective deployment of ML models in healthcare systems.

Data and Code Availability This work uses the MIMIC-IV dataset, which is available on the PhysioNet repository (Goldberger et al., 2000; Johnson et al., 2023a,b,c). The code used for this work is available at: https://github.com/rddelosreyes/ed-abm-synthetic-ehr.

Institutional Review Board (IRB) This work does not require IRB approval.

1. Introduction

Machine learning (ML) models are increasingly deployed in healthcare systems to support various aspects of clinical and operational decision making (Zhang et al., 2022; Poon et al., 2025). One important application of ML models is predicting patient outcomes, such as hospitalisation, inpatient mortality, readmission, and length of stay, as these predictions can inform decisions on patient care and resource management. Many ML models have been shown to make such predictions accurately when trained and tested on curated electronic health record (EHR) datasets (Hilton et al., 2020; Xie et al., 2022; Stone et al., 2022; Lee et al., 2024; Farimani et al.,

2024). Despite such promising results, a typical challenge with ML models is that they lack robustness to changes in the data caused by changes in the system in which they are used (Zech et al., 2018; Nestor et al., 2019; Zhang et al., 2022). These changes in system conditions reflect real-world variations in patient populations, healthcare resources, and clinical practices present in the *deployment* data that are not sufficiently captured in the *development* data used to train and test the ML models (Finlayson et al., 2021). Hence, for safety-critical applications such as healthcare, evaluating the robustness of ML models on data under various real-world system conditions during the ML model development phase is essential to ensure their safe and effective use in clinical practice (Lekadir et al., 2025; Balendran et al., 2025).

Mass casualty incidents (MCIs) caused by disasters such as earthquakes and infectious disease outbreaks are critical instances in which the development data may not capture real-world situations, yet ML models must remain robust. When MCIs occur, healthcare systems experience sudden changes in demand and operating conditions (Carrington et al., 2021). To evaluate the robustness of ML models under such conditions, EHR data reflecting system conditions and patient outcomes across plausible MCI scenarios are necessary. However, real-world EHR data needed for such robustness evaluation are often difficult to obtain for many reasons. For one, access to real-world EHR data is tightly regulated by national authorities and healthcare institutions due to the sensitive information they carry. Even when credentialed access is granted, components of the data are anonymised to comply with these regulations (Johnson et al., 2023c). Furthermore, real-world EHR data capture only what was observed and recorded from reality, which may not include sufficient data for training and testing ML models under various MCI conditions. Thus, the standard process of evaluating ML model performance using a held-out test set from a given real-world EHR dataset is inherently limited in providing insights into the accuracy and robustness of ML models (van Breugel et al., 2023).

To address the limited availability of real-world data, synthetic EHR data have been used as complementary data sources (Gonzales et al., 2023). Synthetic EHR data are artificially generated data based on real-world EHR data and medical knowledge using approaches like computational or generative ML models (Hernandez et al., 2022). These synthetic EHR data generation approaches enable the training and testing of ML models for contexts where real-world EHR data are inadequate or unavailable. A key distinguishing feature of these approaches is that the generation process can be controlled to produce data with specific characteristics. Hence, plausible variations of real-world EHR data can be deliberately added in the synthetic EHR data (van Breugel et al., 2023). The generated synthetic EHR data can then be utilised to test the robustness of ML models systematically against various robustness issues in healthcare (Finlayson et al., 2021; Zhang et al., 2022; Balendran et al., 2025).

However, many existing synthetic EHR data generation approaches lack the capability to explicitly model the system conditions of the healthcare system, where and when patients receive care. The current focus of existing approaches is to mimic the properties of real-world EHR data, such that the synthetic data represent patient characteristics and outcomes that are realistic but do not reveal real patient information (Yan et al., 2022; Budu et al., 2024). In that process, they implicitly model the same system conditions in which the real-world EHR data were collected. Yet, such system conditions do not necessarily extend to MCI scenarios where demand and operating conditions change in uncertain ways (Carrington et al., 2021). Moreover, patient outcomes are also affected by such changes in system conditions. For instance, crowding in the emergency department affects the quality of care, resulting in worse patient outcomes such as higher inpatient mortality, increased risk of readmission, and longer length of stay (Bernstein et al., 2009; Morley et al., 2018; Bravata et al., 2021; Kadri et al., 2021).

To be able to test how robust ML models are during MCIs, there remains a need to capture how system conditions and patient outcomes are reflected in real-world EHR data and control how they are integrated into their synthetic versions. Here, we investigate the use of agent-based models (ABMs) for generating synthetic EHR data to evaluate the robustness of ML models in predicting patient outcomes under MCI conditions. ABMs are mechanistic representations of real-world systems, where entities are modelled as autonomous agents that interact in a system environment according to defined behavioural rules (Bonabeau, 2002). In ABMs, system conditions of real-world systems can be explicitly modelled and simulated. Such an approach makes them particularly suited for investigating how interactions among agents and the environment affect emergent system

properties, from understanding and mitigating disease outbreaks to analysing and optimising healthcare operations (Tracy et al., 2018; Willem et al., 2017; Liu et al., 2017; Delos Reyes et al., 2026). ABMs have also been used to generate realistic synthetic populations for exploring how different policies and interventions affect demographic, social, and health patterns (Geard et al., 2013; Prédhumeau and Manley, 2023; Von Hoene et al., 2025). However, this capability of ABMs for synthetic data generation has not yet been fully explored and leveraged in existing literature on ML robustness.

In this work, we use the emergency department (ED) as the operational setting. We first describe the standard ML process for training and testing an ML model for an ED task using real-world EHR data. We then present an ABM of an ED and demonstrate how patients captured in the real test set can be simulated in the ABM. Finally, we simulate these patients across various MCI scenarios and show how synthetic test sets generated from these ABM simulations can be used to evaluate the robustness of ML models under those conditions. We provide a schematic overview of our proposed approach in Figure 1.

2. Methods

2.1. Real-world EHR data

We use real-world EHR data from the Medical Information Mart for Intensive Care (MIMIC) dataset to develop both the ML model (described in Section 2.2) and the ABM of an ED (described in Section 2.3). MIMIC-IV, the latest version of MIMIC, provides credentialed access to de-identified data of patient admissions at the Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2023b,c). We focus on ED admissions from the MIMIC-IV-ED module (version 2.2), which contains 425,087 records of ED stays from 2011 to 2019 (Goldberger et al., 2000; Johnson et al., 2023a). These records are associated with a unique stay identifier, which can be used to extract information on patients' demographics, conditions, and the various activities they undergo in the ED, including arrival, triage, vital sign checks, medicine dispensations and administrations, laboratory and imaging tests, and discharge. For simplicity, we consider each ED stay record to be associated with a unique patient p in this work.

2.2. ML task and model

We consider the ML task of predicting, at triage upon the patient's arrival in the ED, whether or not a patient will have a length of stay (LOS) in the ED beyond a prespecified duration (Farimani et al., 2024). LOS is a critical performance indicator in the ED that affects patient care and outcomes (Wiler et al., 2015; Vanbrabant et al., 2019). National health guidelines specify target LOS values for treating ED patients to ensure that they are receiving timely and adequate care (e.g., 4 hours in Australia and the UK (Vezyridis and Timmons, 2014; Forero et al., 2019)). Accurate predictions of patients' LOS outcomes could thus support more effective resource utilisation and patient management.

To train and test an ML model for this task, we preprocessed the MIMIC-IV dataset following a benchmarking study of ML models for ED prediction tasks (Xie et al., 2022). The preprocessed dataset made up the full dataset $\mathcal{D}$ used during the ML model development phase:

$$\mathcal{D} = \{(x_p, y_p)\}_{p=1}^{P}, \tag{1}$$

where x_p is a vector of features characterising patient p, y_p is an indicator of the outcome of patient p, and P is the number of patients included in the full dataset $\mathcal{D}$. The feature vector x_p encapsulates the patient-level features of patient p, which consists of age, vital signs, chief complaints, comorbidities, past admission counts, and acuity. The outcome label $y_p \in \{0, 1\}$ indicates whether patient p has an LOS greater than the LOS threshold ℓ, where we set $\ell = 4$. The full dataset $\mathcal{D}$ was then split into a training set $\mathcal{D}_{train}$ (80%) and a test set $\mathcal{D}_{test}$ (20%).

Given the feature vector x_p of patient p, an ML model f predicts the LOS outcome of patient p:

$$\hat{y}_p = f(x_p; \theta_f), \tag{2}$$

where $\hat{y}_p \in \{0, 1\}$ is the predicted label of the outcome and θ_f denotes the parameters of the ML model f. The ML model's parameters θ_f are learned from the training set $\mathcal{D}_{train}$. The ML model's performance is measured based on the accuracy of the predicted outcome label $\hat{y}_p$ relative to the true outcome label y_p for every patient p in the test set $\mathcal{D}_{test}$. We trained and tested three common ML models: random forest, gradient boosting, and multilayer perceptron. These ML models are widely used for tabular data and demonstrate performance that is competitive with more recent architectures on such data

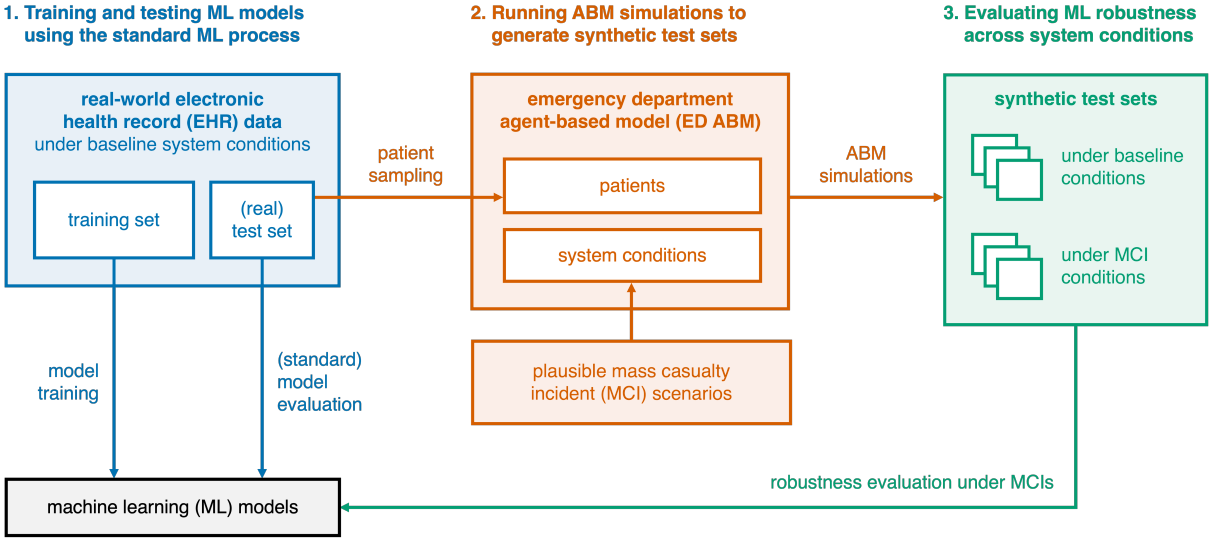

Figure 1: Schematic overview of our ABM-based approach for synthetic EHR generation and ML robustness evaluation under MCI scenarios.

(Nestor et al., 2019; Hilton et al., 2020; Xie et al., 2022; van Breugel et al., 2023; Farimani et al., 2024; Lee et al., 2024).

2.3. ED ABM description

We adapt a previously published ABM of the ED at the Beth Israel Deaconess Medical Center, which was developed using the MIMIC-IV dataset (Delos Reyes et al., 2024). For simplicity, we refer to it as ED ABM. Whereas the ML model described in Section 2.2 is designed to predict patients' LOS outcomes, this ED ABM is designed to generate synthetic EHR data for evaluating the robustness of the ML model in making those predictions under different system conditions of the ED. The ED ABM represents the ED environment and its resources, as well as the flow of patients through that environment. The state of the ED environment is defined by its `hourly_arrival_rate`, `bed_capacity`, `clinician_capacity`, `imaging_capacity`, and `clinical_workflow`. Every patient in this ED environment has a state variable for `age`, `vital_signs`, `chief_complaints`, `comorbidities`, `past_admission_counts`, `acuity`, `disposition`, `trajectory`, and `length_of_stay`.

The state of the ED ABM is updated at discrete time steps. At each time step, the following occurs. First, patient arrivals are generated stochastically following the prespecified `hourly_arrival_rate`. The state variables of each generated patient, except the `length_of_stay`, are specified according to the patient record sampled stochastically from the given dataset (explained in Section 2.4). Second, patients are assigned to beds based on their `acuity` and bed availability, with those of higher acuity given priority. Finally, patients in bed are treated. Once patients are discharged from the ED, their state variables are recorded and are then removed from the ED environment.

Since resources are limited and shared among all patients in the ED environment, patients may have to wait for resources during their ED stay. This waiting time is not included in the patient's `trajectory`, which only encapsulates the execution time of each activity. Hence, the `length_of_stay`, which is the sum of all execution and waiting times, is a stochastic and emergent output of the ED ABM.

2.4. Sampling patients from the real test set

When patient arrivals are generated in the ED ABM, their state variables need to be initialised. Except for `length_of_stay` which is initialised to 0, we specify the values of these state variables based on patient records from the test set $\mathcal{D}_{test}$. This setup can be viewed as putting a patient from the test set $\mathcal{D}_{test}$ into the ED ABM. For every new patient p, we sam-

ple the patient's `acuity` and `disposition` based on observed frequencies from the test set $\mathcal{D}_{test}$. We then randomly sample with replacement a patient from the test set $\mathcal{D}_{test}$ with that acuity and disposition. The `age`, `vital_signs`, `chief_complaints`, `comorbidities`, and `past_admission_counts` of the new patient p are then directly specified according to the sampled patient's ED stay record.

The `trajectory` of the new patient p indicates the sequence and execution time of activities that the patient will undergo in the ED, starting with arrival in the ED and ending with discharge from the ED. Specifying the sequence and execution time of these activities requires further preprocessing of the MIMIC-IV dataset. To extract patient trajectories from the dataset, we obtained the event log of patients in the test set $\mathcal{D}_{test}$. An event log is a digital record of the activities that patients underwent in the system (Rojas et al., 2016). Each recorded event includes a patient identifier, the name of the activity, and the timestamp at which the activity was recorded into the EHR. By sorting each patient's recorded events chronologically, we can reconstruct the series of activities that patients underwent in the ED from arrival to discharge, and use that to specify the `trajectory` of the new patient p in the ED ABM. Since the activity timestamps in the MIMIC-IV dataset only indicate the time at which an activity was recorded into the EHR, the recorded time duration between two activities may include both waiting and execution times for the latter activity. We removed the waiting times embedded in the recorded patient trajectories as detailed in Appendix A.

2.5. Generating synthetic test sets

We used the ED ABM to generate synthetic EHR data under different system conditions. In formal terms, the ED ABM g takes as input the *real* test set $\mathcal{D}_{test}$ and outputs a *synthetic* test set $\mathcal{D}_{syn}$:

$$\mathcal{D}_{syn} = g(\mathcal{D}_{test}; \theta_g), \tag{3}$$

where θ_g denotes the parameters of the ED environment that describe the system conditions in the ED. The ED environment's parameters θ_g include the hourly patient arrival rate, the resource capacity for beds, clinicians, and imaging equipment, and the ED workflow, which were all initially set based on previously calibrated values (Delos Reyes et al., 2024). We note that these calibrated values represent the baseline system conditions of the ED, reflecting the system conditions from which data in the MIMIC-IV dataset were collected. By changing these parameter values, we can simulate the ED under different system conditions and generate a synthetic test set $\mathcal{D}_{syn}$ that reflects those conditions.

Similar to the real test set $\mathcal{D}_{test}$, every patient p included in the synthetic test set $\mathcal{D}_{syn}$ is also represented using the simulated patient-level feature vector $\bar{x}_p$ and the simulated outcome label $\bar{y}_p$:

$$(\bar{x}_p, \bar{y}_p) \in \mathcal{D}_{syn}. \tag{4}$$

As per the patient sampling described in Section 2.4, the simulated patient-level feature vector $\bar{x}_p \in \mathcal{D}_{syn}$ of patient p is always similar to the real patient-level feature vector $x_p \in \mathcal{D}_{test}$. However, the simulated outcome label $\bar{y}_p \in \mathcal{D}_{syn}$ may differ from the true outcome label $y_p \in \mathcal{D}_{test}$ because patient p's `length_of_stay` is also affected by the ED environment's parameters θ_g.

2.6. Experimental setup

We conducted two experiments to evaluate the utility of our proposed ABM-based approach for generating synthetic EHR data and evaluating ML robustness. Since the ED ABM is stochastic, we report results across 1,000 simulation runs, each run generating a synthetic test set. The first experiment evaluated how well the ED ABM could generate synthetic EHR data with LOS characteristics that match the MIMIC-IV dataset at the population and patient levels. For each simulation run n, we compared the simulated LOS values in the synthetic test set $\mathcal{D}^n_{syn}$ to the true LOS values in the real test set $\mathcal{D}_{test}$. At the population level, we examined the LOS distributions across the overall patient population and patient groups by acuity and disposition. At the patient level, we calculated coverage and width to examine whether each patient's LOS was reproduced correctly. Coverage measures the fraction of patients whose true LOS value falls within the interval of simulated LOS values generated across simulation runs, while width measures the average size of this interval across all patients (van Breugel et al., 2023). In addition, we also compared the performance of the ML models trained on the real training set $\mathcal{D}_{train}$ when tested on the real test set $\mathcal{D}_{test}$ versus when tested on each synthetic test set $\mathcal{D}^n_{syn}$.

The second experiment assessed the robustness of ML models under MCI conditions using the synthetic test sets generated by the ED ABM. We simulated a

	Population level			Patient level	
	Real median LOS	**Simulated median LOS**	**Wasserstein distance**	**Coverage**	**Width**
Acuity					
1	5.20	5.10 (4.69–5.50)	0.86 (0.68–1.07)	0.80	0.60 (0.40–0.89)
2	6.70	6.32 (6.12–6.56)	0.86 (0.63–1.13)	0.63	0.68 (0.45–0.98)
3	5.82	5.90 (5.60–6.28)	0.65 (0.50–0.85)	0.87	3.15 (2.14–4.26)
4	2.92	3.09 (2.73–3.48)	0.73 (0.60–0.94)	0.91	3.05 (2.04–4.12)
5	2.03	2.30 (1.43–3.71)	1.42 (1.20–1.90)	0.92	2.54 (1.45–3.75)
Disposition					
Home	5.25	5.30 (5.05–5.60)	0.77 (0.61–0.96)	0.79	2.47 (0.95–3.83)
Ward	7.07	7.00 (6.76–7.28)	0.55 (0.42–0.73)	0.78	1.12 (0.59–2.91)
ICU	5.17	5.26 (4.92–5.60)	0.71 (0.58–0.90)	0.87	0.80 (0.50–1.32)
Overall	5.85	5.81 (5.60–6.07)	0.59 (0.47–0.76)	0.79	1.85 (0.72–3.50)

Table 1: Validation of the ED ABM under baseline system conditions. The ED ABM generated synthetic test sets that accurately reproduced the LOS distributions from the real test set across the overall patient population and different acuity and disposition groups. The simulated median LOS (in hours) closely matches the real median LOS, with small Wasserstein distances between distributions. At the patient level, simulated LOS intervals capture the real LOS value for 79% of patients, with a median interval width of 1.85. Values in parentheses indicate the interquartile range across 1,000 simulation runs.

4-day MCI period, during which ED system conditions were systematically modified to reflect (1) an increase in patient arrivals, (2) a decrease in resource capacity, or (3) a delay in clinical workflow. Only patients who arrived in the ED during this time period were included in the analysis for both the baseline and MCI conditions. With 1,000 simulation runs under this setup, each patient was sampled an average of 9.4 times. We used the standard precision and recall metrics to measure ML model performance. To contextualise these ML model-based metrics in terms of patient and operational impact, we also calculated the number of patients with LOS >4 hours who were missed by the models per 100 ED stays.

3. Results

3.1. Baseline system conditions

The ED ABM reproduced accurately the LOS values from the real test set under baseline conditions, as shown in Table 1. For the overall patient population, the real median LOS is 5.85 hours, while the simulated median LOS is 5.81 hours. The median Wasserstein distance between the real and simulated LOS distributions is 0.59, indicating high similarity. Likewise, the true and simulated LOS distributions across different acuity and disposition groups have Wasserstein distances ranging from 0.55 to 1.42. The ED ABM also generated simulated LOS intervals that captured well the true LOS value of each patient. For the overall patient population, coverage is 79% and width is 1.85. For the different acuity and patient groups, coverage ranges from 63% to 92%, while width ranges from 0.60 to 3.05.

The three ML models also showed similar precision and recall on the real test set and synthetic test sets under baseline system conditions, as shown in Table 3. On the real test set, model precision ranges from 0.78 to 0.79, while model recall ranges from 0.92 to 0.95. On the synthetic test sets, model precision ranges from 0.79 to 0.80, while model recall ranges from 0.91 to 0.95. Correspondingly, the number of missed patients with LOS >4 hours is 4–7 per 100 ED stays under baseline system conditions.

3.2. Synthetic MCI conditions

Under MCI conditions, LOS distributions in the synthetic test sets shifted relative to baseline system conditions, as shown in Table 2. Increases in ED system load from MCIs, resulting from higher patient arrivals, reduced clinician capacity, and delays in laboratory workflow, led to longer LOS on average. Across the synthetic MCI scenarios considered, overall median LOS increased by 0.21–3.26 hours, while

	Median LOS	LOS >4 hrs
Baseline conditions		
Real test set	5.85	0.74
Synthetic test sets	5.81 (5.60–6.07)	0.75 (0.72–0.77)
Synthetic MCI conditions: increase in arrivals		
+5% daily arrivals	6.20 (5.88–6.56)	0.78 (0.75–0.81)
+10% daily arrivals	6.74 (6.28–7.29)	0.81 (0.79–0.84)
+15% daily arrivals	7.66 (6.97–8.60)	0.85 (0.82–0.87)
+20% daily arrivals	9.07 (7.97–10.49)	0.87 (0.85–0.89)
Synthetic MCI conditions: decrease in resources		
-5% clinicians	6.02 (5.73–6.41)	0.76 (0.73–0.79)
-10% clinicians	6.30 (5.93–6.80)	0.78 (0.75–0.82)
-15% clinicians	7.39 (6.67–8.59)	0.84 (0.81–0.88)
-20% clinicians	8.72 (7.47–10.31)	0.88 (0.84–0.91)
Synthetic MCI conditions: delay in workflow		
+5 mins for lab tests	6.33 (6.04–6.65)	0.79 (0.77-0.82)
+10 mins for lab tests	6.90 (6.58–7.29)	0.83 (0.81–0.85)
+15 mins for lab tests	7.56 (7.17–8.08)	0.86 (0.84–0.88)
+20 mins for lab tests	8.39 (7.83–9.17)	0.89 (0.87–0.91)

Table 2: LOS characteristics of the real and synthetic test sets under baseline and MCI conditions. Median LOS (in hours) increases as system load increases (via an increase in arrivals, a decrease in resources, or a delay in workflow), as well as the fraction of patients with LOS >4 hours. Values are reported as median (interquartile range) across 1,000 simulation runs.

the fraction of patients with LOS >4 hours increased by 0.01–0.14.

When evaluated on synthetic test sets generated under MCI conditions, all ML models exhibited a lack of robustness to shifts in LOS distributions arising from increases in system load, as shown in Table 3. Although model precision increased by 0.01–0.12 relative to baseline, model recall consistently declined by 0.01–0.04. This reduction in recall translates into an additional 1–5 patients with LOS >4 hours missed per 100 ED stays, indicating a degradation in the ability of the ML models to predict patients at risk of staying longer in the ED during MCIs.

4. Discussion

Robustness evaluation should be a standard part of machine learning (ML) model evaluation before deployment of ML models in healthcare systems, determining when and how they can be safely and effectively used to support clinical and operational decisions (Finlayson et al., 2021; Lekadir et al., 2025; Balendran et al., 2025). When evaluating ML models during the ML model development phase, the standard use of a held-out test set from a real-world electronic health record (EHR) dataset is inadequate for measuring the accuracy and robustness of these ML models in deployment because of its scarcity and potential data misalignment (Zech et al., 2018; Nestor et al., 2019; Zhang et al., 2022; van Breugel et al., 2023). Hence, relying solely on ML model performance based on evaluations on this held-out test set could lead to a retrospective identification of ML model vulnerabilities, posing risks to patients and healthcare providers.

Here, we focused on evaluating the robustness of ML models under mass casualty incidents (MCIs), wherein crowding resulting from changes in system conditions, such as increases in patient arrivals, decreases in resource capacity, and delays in clinical workflow, negatively affects patient outcomes, such as inpatient mortality and length of stay (LOS) (Bernstein et al., 2009; Morley et al., 2018; Bravata et al., 2021; Kadri et al., 2021). Since MCIs are relatively rare, stochastic in nature, and potentially novel, real-world EHR data that adequately capture changes in system conditions and patient outcomes during MCIs are often limited or unavailable. This data limitation makes it difficult or impossible to evaluate ML robustness under MCI conditions. To enable this robustness evaluation, we presented an agent-based model (ABM) of an emergency department (ED) for generating synthetic EHR data during MCIs.

By explicitly modelling system conditions, our ABM-based approach to generating synthetic EHR data can simulate realistic and interpretable changes to arrivals, resources, and workflow induced by MCIs. It accounts for how the impact of system conditions on patient outcomes is captured in real-world EHR data and enables control over how these system conditions and their impact are reflected in synthetic EHR data. Such capabilities are rarely considered or demonstrated in existing synthetic EHR data generation approaches (Hernandez et al., 2022; Yan et al., 2022; van Breugel et al., 2023; Budu et al., 2024). Our work thus improves on existing approaches to synthetic EHR data generation and ML model evaluation by enabling the stress testing of ML models under MCI and other system conditions expected in deployment but not sufficiently captured in the development data. The system conditions in the ABM can be readily adapted to consider other scenarios. For

	Precision			Recall			Missed patients with LOS >4 hrs (per 100 ED stays)		
	RF	**GB**	**MLP**	**RF**	**GB**	**MLP**	**RF**	**GB**	**MLP**
Baseline conditions									
Real test set	0.79	0.78	0.79	0.92	0.95	0.94	5.93	3.62	4.11
Synthetic test sets	0.80 ± 0.06	0.79 ± 0.06	0.79 ± 0.06	0.91 ± 0.03	0.95 ± 0.02	0.94 ± 0.03	6.64 ± 2.58	4.11 ± 2.06	4.70 ± 2.21
Synthetic MCI conditions: increase in arrivals									
+5% daily arrivals	0.82 ± 0.06	0.81 ± 0.06	0.81 ± 0.06	0.90 ± 0.03	0.94 ± 0.03	0.93 ± 0.03	7.77 ± 2.97	5.06 ± 2.46	5.71 ± 2.61
+10% daily arrivals	0.84 ± 0.06	0.84 ± 0.06	0.84 ± 0.06	0.89 ± 0.03	0.92 ± 0.03	0.92 ± 0.03	9.15 ± 3.27	6.17 ± 2.68	6.93 ± 2.81
+15% daily arrivals	0.87 ± 0.06	0.86 ± 0.06	0.86 ± 0.06	0.88 ± 0.03	0.91 ± 0.03	0.90 ± 0.03	10.40 ± 3.19	7.25 ± 2.62	8.06 ± 2.76
+20% daily arrivals	0.88 ± 0.05	0.88 ± 0.05	0.88 ± 0.05	0.87 ± 0.03	0.91 ± 0.02	0.90 ± 0.03	11.24 ± 2.83	7.99 ± 2.34	8.85 ± 2.51
Synthetic MCI conditions: decrease in resources									
-5% clinicians	0.81 ± 0.07	0.80 ± 0.07	0.80 ± 0.07	0.91 ± 0.03	0.94 ± 0.03	0.93 ± 0.03	7.14 ± 2.98	4.55 ± 2.36	5.14 ± 2.55
-10% clinicians	0.83 ± 0.07	0.82 ± 0.08	0.82 ± 0.08	0.90 ± 0.03	0.94 ± 0.03	0.93 ± 0.03	7.78 ± 3.35	5.03 ± 2.70	5.72 ± 2.88
-15% clinicians	0.87 ± 0.08	0.87 ± 0.08	0.87 ± 0.08	0.89 ± 0.04	0.93 ± 0.03	0.92 ± 0.03	9.51 ± 3.98	6.37 ± 3.22	7.14 ± 3.48
-20% clinicians	0.90 ± 0.08	0.89 ± 0.08	0.90 ± 0.08	0.88 ± 0.04	0.92 ± 0.03	0.91 ± 0.03	10.55 ± 3.86	7.20 ± 3.17	8.02 ± 3.36
Synthetic MCI conditions: delay in workflow									
+5 mins for lab tests	0.84 ± 0.06	0.83 ± 0.06	0.83 ± 0.06	0.90 ± 0.03	0.94 ± 0.02	0.93 ± 0.03	7.62 ± 2.82	4.88 ± 2.21	5.51 ± 2.40
+10 mins for lab tests	0.87 ± 0.05	0.86 ± 0.05	0.86 ± 0.05	0.90 ± 0.03	0.93 ± 0.03	0.92 ± 0.03	8.60 ± 3.06	5.59 ± 2.49	6.32 ± 2.62
+15 mins for lab tests	0.89 ± 0.05	0.89 ± 0.05	0.89 ± 0.05	0.89 ± 0.03	0.93 ± 0.03	0.92 ± 0.03	9.64 ± 3.17	6.43 ± 2.61	7.22 ± 2.76
+20 mins for lab tests	0.92 ± 0.04	0.91 ± 0.04	0.91 ± 0.04	0.88 ± 0.03	0.92 ± 0.03	0.91 ± 0.03	10.68 ± 3.18	7.24 ± 2.61	8.10 ± 2.79

Table 3: ML robustness evaluation results. As ED system load increases under MCI conditions, precision increases while recall decreases across all three ML models: random forest (RF), gradient boosting (GB), and multilayer perceptron (MLP). The number of missed patients with LOS >4 hours increases from 4–7 per 100 ED stays under baseline conditions to 5–11 per 100 ED stays under the synthetic MCI conditions. Results for each of the synthetic baseline and MCI conditions are shown as the mean ± 2 standard deviations across 1,000 test sets.

instance, simultaneous changes in arrivals, resources, and workflow can also be explored to simulate more complex MCI scenarios. Since increases in system load due to each of these changes led to longer LOS, as shown in Table 2, such simultaneous changes would be expected to further increase system load and result in similar findings.

In practice, our approach could be applied to EHR data from other healthcare systems to generate synthetic data relevant to the specific real-world scenarios they confront. These synthetic data would support proactive and systematic evaluation of the robustness of ML models under various real-world system conditions, informing ML model deployment and indicating when further improvement is required. While there will always be uncertainty about how the synthetic EHR data compare to real-world EHR data during MCIs, they can provide insights into the robustness of ML models in healthcare that are not obtainable with real-world EHR data alone nor with synthetic EHR data generated using existing approaches.

Many ML models designed to predict patient outcomes often rely solely on patient-level features, such as patients' demographics, past medical encounters, and current health conditions (Nestor et al., 2019; Xie et al., 2022; Stone et al., 2022; van Breugel et al., 2023; Lee et al., 2024; Farimani et al., 2024). In part, this oversight is due to the abstraction of system conditions from EHR datasets, such as publicly available benchmarks like MIMIC, to protect the privacy of patients and healthcare providers (Johnson et al., 2023c). However, in the simulation experiments, many patients had longer LOS under MCI conditions with higher system load than under baseline system conditions, regardless of their patient-level features, as exhibited in Table 2. Consequently, ML models trained solely on patient-level features, following common practice, failed to predict LOS under those MCI conditions as accurately as under baseline system conditions, as demonstrated in Table 3. Hence, when using real-world EHR data or generating synthetic EHR data for training and testing ML models during the ML model development phase, how system conditions and their impact are encapsulated in the data should also be explicitly considered (e.g., by incorporating system-level features) to assess whether ML models are robust under expected real-world system conditions, such as in situations like MCIs when they are needed more.

We introduced ABMs as an approach for explicitly modelling and simulating these system conditions and their impact on patient outcomes. The use of ABMs to support the training and testing of ML models is underexplored, despite them being widely used to inform decision making across healthcare systems (Tracy et al., 2018). Here, we showed that an ABM can reproduce and extend real-world EHR data, as shown in Tables 1 and 2, respectively. Further, we demonstrated that the generated synthetic EHR data can be used to evaluate the robustness of ML models under MCI conditions, as shown in Table 3. Our work can also be used as a reference for many other potential applications of ABMs for ML research and education. First, ABMs could also be used to generate synthetic data for evaluating other ML robustness issues in healthcare (Finlayson et al., 2021; Zhang et al., 2022; Balendran et al., 2025). Second, they could complement existing synthetic EHR data generation approaches designed for generating completely synthetic patients (Yan et al., 2022; van Breugel et al., 2023; Budu et al., 2024). Third, they could augment system conditions in public EHR benchmarks like MIMIC, which are often anonymised for privacy reasons, to enable further studies on the effect of system-level features on ML model performance (Johnson et al., 2023c). Fourth, they could also be used to generate synthetic EHR data for training ML models to enhance their robustness. Finally, ABMs could also serve as clinical environment simulators, in which ML models are embedded within ABMs to enable the dynamic evaluation of their impact on system behaviour over time (Luo et al., 2026).

Our work has several limitations. First, we assumed that recorded time durations between ED activities that deviate from an average value include waiting times, which may not necessarily be the case. Despite this, we validated that this choice generated synthetic EHR data with similar LOS characteristics to the real-world EHR data. Further, this removal process of waiting times could be useful in cases when waiting times are not explicitly captured in the EHR data. Second, ML model evaluation under MCI conditions is based only on synthetic scenarios and would thus require EHR data during real-world MCIs for further validation. However, the emergent data shifts observed under these MCI conditions (i.e., increasing LOS with increasing system load) align with expectations and are supported by extensive literature on ED crowding (Bernstein et al., 2009; Morley et al., 2018). Finally, we focused on modelling and predicting LOS only. Future work could extend our approach to evaluate the robustness of ML models for predicting other patient outcomes such as hospitalisation, inpatient mortality, and readmission.

5. Conclusion

In summary, we developed an ED ABM that can generate synthetic EHR data during MCIs. We then demonstrated how the synthetic EHR data can be used to evaluate the robustness of ML models under various MCI conditions. Our results showed that ML models for predicting LOS outcomes lack robustness to changes in system conditions induced by MCIs. Our work extends existing synthetic EHR data generation approaches by enabling the explicit modelling of system conditions. It also extends the standard ML model evaluation process by offering a proactive and systematic approach to evaluating the robustness of ML models under MCI conditions before deployment. Our work thus helps ensure that ML models can be safely and effectively used in clinical practice. Furthermore, it provides a basis for novel and relevant applications of ABMs in ML for health.

Author contributions: CRediT

Roben Delos Reyes: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Daniel Capurro**: Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Nicholas Geard**: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Funding sources

This work was supported by the Melbourne Research Scholarship provided by the University of Melbourne.

Acknowledgments

The experiments conducted in this work were supported by the resources provided by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Alan Balendran, Céline Beji, Florie Bouvier, Ottavio Khalifa, Theodoros Evgeniou, Philippe Ravaud, and Raphaël Porcher. A scoping review of robustness concepts for machine learning in healthcare. *npj Digital Medicine*, 8(1):38, January 2025. ISSN 2398-6352. URL https://doi.org/10.1038/s41746-024-01420-1.

Steven L. Bernstein, Dominik Aronsky, Reena Duseja, Stephen Epstein, Dan Handel, Ula Hwang, Melissa McCarthy, K. John McConnell, Jesse M. Pines, Niels Rathlev, Robert Schafermeyer, Frank Zwemer, Michael Schull, Brent R. Asplin, and Society for Academic Emergency Medicine, Emergency Department Crowding Task Force. The Effect of Emergency Department Crowding on Clinically Oriented Outcomes. *Academic Emergency Medicine*, 16(1):1–10, January 2009. ISSN 1069-6563, 1553-2712. URL https://doi.org/10.1111/j.1553-2712.2008.00295.x.

Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7280–7287, May 2002. ISSN 0027-8424, 1091-6490. URL https://doi.org/10.1073/pnas.082080899.

Dawn M. Bravata, Anthony J. Perkins, Laura J. Myers, Greg Arling, Ying Zhang, Alan J. Zillich, Lindsey Reese, Andrew Dysangco, Rajiv Agarwal, Jennifer Myers, Charles Austin, Ali Sexson, Samuel J. Leonard, Sharmistha Dev, and Salomeh Keyhani. Association of Intensive Care Unit Patient Load and Demand With Mortality Rates in US Department of Veterans Affairs Hospitals During the COVID-19 Pandemic. *JAMA Network Open*, 4(1):e2034266, January 2021. ISSN 2574-3805. URL https://doi.org/10.1001/jamanetworkopen.2020.34266.

Emmanuella Budu, Kobra Etminani, Amira Soliman, and Thorsteinn Rögnvaldsson. Evaluation of synthetic electronic health records: A systematic review and experimental assessment. *Neurocomputing*, 603:128253, October 2024. ISSN 09252312. URL https://doi.org/10.1016/j.neucom.2024.128253.

Mercedes A. Carrington, Jamie Ranse, and Karen Hammad. The impact of disasters on emergency department resources: Review against the Sendai framework for disaster risk reduction 2015–2030. *Australasian Emergency Care*, 24(1):55–60, March 2021. ISSN 2588994X. URL https://doi.org/10.1016/j.auec.2020.09.003.

Roben Delos Reyes, Daniel Capurro, and Nicholas Geard. Modelling patient trajectories in emergency department simulations using retrospective patient cohorts. *Computers in Biology and Medicine*, 182:109147, November 2024. ISSN 00104825. URL https://doi.org/10.1016/j.compbiomed.2024.109147.

Roben Delos Reyes, Daniel Capurro, and Nicholas Geard. Data assimilation in emergency department simulations for real-time disaster response. *International Journal of Disaster Risk Reduction*, 133:105995, February 2026. ISSN 22124209. URL https://doi.org/10.1016/j.ijdrr.2026.105995.

Raheleh Mahboub Farimani, Hesam Karim, Alireza Atashi, Fariba Tohidinezhad, Kambiz Bahaadini, Ameen Abu-Hanna, and Saeid Eslami. Models to predict length of stay in the emergency department: A systematic literature review and appraisal. *BMC Emergency Medicine*, 24(1):54, April 2024. ISSN 1471-227X. URL https://doi.org/10.1186/s12873-024-00965-4.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021. ISSN 0028-4793, 1533-4406. URL https://doi.org/10.1056/NEJMc2104626.

Roberto Forero, Shizar Nahidi, Josephine De Costa, Daniel Fatovich, Gerry FitzGerald, Sam Toloo, Sally McCarthy, David Mountain, Nick Gibson, Mohammed Mohsin, and Wing Nicola Man. Perceptions and experiences of emergency department staff during the implementation of the four-hour rule/national emergency access target policy in Australia: A qualitative social dynamic perspective. *BMC Health Services Research*, 19(1):82, December 2019. ISSN 1472-6963. URL https://doi.org/10.1186/s12913-019-3877-8.

Nicholas Geard, James M McCaw, Alan Dorin, Kevin B Korb, and Jodie McVernon. Synthetic

Population Dynamics: A Model of Household Demography. *Journal of Artificial Societies and Social Simulation*, 16(1):8, January 2013. ISSN 1460-7425. URL https://doi.org/10.18564/jasss.2098.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23), June 2000. ISSN 0009-7322, 1524-4539. URL https://doi.org/10.1161/01.CIR.101.23.e215.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2 (1):e0000082, January 2023. ISSN 2767-3170. URL https://doi.org/10.1371/journal.pdig.0000082.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, July 2022. ISSN 09252312. URL https://doi.org/10.1016/j.neucom.2022.04.053.

C. Beau Hilton, Alex Milinovich, Christina Felix, Nirav Vakharia, Timothy Crone, Chris Donovan, Andrew Proctor, and Aziz Nazha. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *npj Digital Medicine*, 3(1):51, April 2020. ISSN 2398-6352. URL https://doi.org/10.1038/s41746-020-0249-z.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, January 2023a. URL https://doi.org/10.13026/5ntk-km72. Version 2.2.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, January 2023b. URL https://doi.org/10.13026/6mm1-ek67. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1): 1, January 2023c. ISSN 2052-4463. URL https://doi.org/10.1038/s41597-022-01899-x.

Sameer S. Kadri, Junfeng Sun, Alexander Lawandi, Jeffrey R. Strich, Lindsay M. Busch, Michael Keller, Ahmed Babiker, Christina Yek, Seidu Malik, Janell Krack, John P. Dekker, Alicen B. Spaulding, Emily Ricotta, John H. Powers, Chanu Rhee, Michael Klompas, Janhavi Athale, Tegan K. Boehmer, Adi V. Gundlapalli, William Bentley, S. Deblina Datta, Robert L. Danner, Cumhur Y. Demirkale, and Sarah Warner. Association Between Caseload Surge and COVID-19 Survival in 558 U.S. Hospitals, March to August 2020. *Annals of Internal Medicine*, 174(9):1240–1251, September 2021. ISSN 0003-4819, 1539-3704. URL https://doi.org/10.7326/M21-1213.

Yi-Chih Lee, Chip-Jin Ng, Chun-Chuan Hsu, Chien-Wei Cheng, and Shou-Yen Chen. Machine learning models for predicting unscheduled return visits to an emergency department: A scoping review. *BMC Emergency Medicine*, 24(1):20, January 2024. ISSN 1471-227X. URL https://doi.org/10.1186/s12873-024-00939-6.

Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, Georgios Kaissis, Gianna Tsakou, Irène Buvat, Jayashree Kalpathy-Cramer, John Mongan, Julia A Schnabel, Kaisar Kushibar, Katrine Riklund, Kostas Marias, Lameck M Amugongo, Lauren A Fromont, Lena Maier-Hein, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, M Jorge Cardoso, Maciej Bobowicz, Mahsa Shabani, Manolis Tsiknakis, Maria A Zuluaga, Marie-Christine Fritzsche, Marina Camacho, Marius George Linguraru, Markus Wenzel, Marleen De Bruijne, Martin G Tolsgaard, Melanie Goisauf, Mónica Cano Abadía, Nikolaos Papanikolaou, Noussair Lazrak, Oriol Pujol, Richard Osuala, Sandy Napel, Sara Colantonio, Smriti Joshi, Stefan Klein, Susanna Aussó, Wendy A Rogers, Zohaib Salahuddin, and Martijn P A Starmans. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, February 2025.

ISSN 1756-1833. URL https://doi.org/10.1136/bmj-2024-081554.

Zhengchun Liu, Dolores Rexachs, Francisco Epelde, and Emilio Luque. An agent-based model for quantitatively analyzing and predicting the complex behavior of emergency departments. *Journal of Computational Science*, 21:11–23, July 2017. ISSN 18777503. URL https://doi.org/10.1016/j.jocs.2017.05.015.

Luyang Luo, Sung Eun Kim, Xiaoman Zhang, Julius M. Kernbach, Roshan Kenia, Julian N. Acosta, Larry A. Nathanson, Adrian D. Haimovich, Adam Rodman, Ethan Goh, Jonathan H. Chen, Nigam H. Shah, David A. Kim, James Zou, Faisal Mahmood, Jakob Nikolas Kather, Matthew Lungren, Vivek Natarajan, Eric J. Topol, and Pranav Rajpurkar. A clinical environment simulator for dynamic AI evaluation. *Nature Medicine*, 32(3):820–827, March 2026. ISSN 1078-8956, 1546-170X. URL 10.1038/s41591-026-04252-6.

Claire Morley, Maria Unwin, Gregory M. Peterson, Jim Stankovich, and Leigh Kinsman. Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE*, 13(8):e0203316, August 2018. ISSN 1932-6203. URL https://doi.org/10.1371/journal.pone.0203316.

Bret Nestor, Matthew B A McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 381–405. PMLR, August 2019. URL https://proceedings.mlr.press/v106/nestor19a.html.

Eric G Poon, Christy Harris Lemak, Juan C Rojas, Janet Guptill, and David Classen. Adoption of artificial intelligence in healthcare: Survey of health system priorities, successes, and challenges. *Journal of the American Medical Informatics Association*, 32(7):1093–1100, July 2025. ISSN 1067-5027, 1527-974X. URL https://doi.org/10.1093/jamia/ocaf065.

Manon Prédhumeau and Ed Manley. A synthetic population for agent-based modelling in Canada. *Scientific Data*, 10(1):148, March 2023. ISSN 2052-4463. URL https://doi.org/10.1038/s41597-023-02030-4.

Eric Rojas, Jorge Munoz-Gama, Marcos Sepúlveda, and Daniel Capurro. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61:224–236, June 2016. ISSN 15320464. URL https://doi.org/10.1016/j.jbi.2016.04.007.

Florian Stertz, Juergen Mangler, and Stefanie Rinderle-Ma. Temporal Conformance Checking at Runtime based on Time-infused Process Models, August 2020. URL https://doi.org/10.48550/arXiv.2008.07262.

Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4):e0000017, April 2022. ISSN 2767-3170. URL https://doi.org/10.1371/journal.pdig.0000017.

Melissa Tracy, Magdalena Cerdá, and Katherine M. Keyes. Agent-Based Modeling in Public Health: Current Applications and Future Directions. *Annual Review of Public Health*, 39(1):77–94, April 2018. ISSN 0163-7525, 1545-2093. URL https://doi.org/10.1146/annurev-publhealth-040617-014317.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can You Rely on Your Model Evaluation? Improving Model Evaluation with Synthetic Test Data. In *Advances in Neural Information Processing Systems*, volume 36, pages 1889–1904. Curran Associates, Inc., December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/05fb0f4e645cad23e0ab59d6b9901428-Paper-Conference.pdf.

Lien Vanbrabant, Kris Braekers, Katrien Ramaekers, and Inneke Van Nieuwenhuyse. Simulation of emergency department operations: A comprehensive review of KPIs and operational improvements. *Computers & Industrial Engineering*, 131:356–381, May 2019. ISSN 03608352. URL https://doi.org/10.1016/j.cie.2019.03.025.

Paraskevas Vezyridis and Stephen Timmons. National targets, process transformation and local consequences in an NHS emergency department (ED): A qualitative study. *BMC Emergency Medicine*, 14(1):12, December 2014. ISSN 1471-227X. URL https://doi.org/10.1186/1471-227X-14-12.

Emma Von Hoene, Amira Roess, Hamdi Kavak, and Taylor Anderson. Synthetic population generation with public health characteristics for spatial agent-based models. *PLOS Computational Biology*, 21(3):e1012439, March 2025. ISSN 1553-7358. URL https://doi.org/10.1371/journal.pcbi.1012439.

Jennifer L. Wiler, Shari Welch, Jesse Pines, Jeremiah Schuur, Nick Jouriles, and Suzanne Stone-Griffith. Emergency Department Performance Measures Updates: Proceedings of the 2014 Emergency Department Benchmarking Alliance Consensus Summit. *Academic Emergency Medicine*, 22(5):542–553, May 2015. ISSN 1069-6563, 1553-2712. URL https://doi.org/10.1111/acem.12654.

Lander Willem, Frederik Verelst, Joke Bilcke, Niel Hens, and Philippe Beutels. Lessons from a decade of individual-based models for infectious disease transmission: A systematic review (2006-2015). *BMC Infectious Diseases*, 17(1):612, December 2017. ISSN 1471-2334. URL https://doi.org/10.1186/s12879-017-2699-8.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, October 2022. ISSN 2052-4463. URL https://doi.org/10.1038/s41597-022-01782-9.

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. A Multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1):7609, December 2022. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-022-35295-1.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018. ISSN 1549-1676. URL https://doi.org/10.1371/journal.pmed.1002683.

Angela Zhang, Lei Xing, James Zou, and Joseph C. Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345, July 2022. ISSN 2157-846X. URL https://doi.org/10.1038/s41551-022-00898-y.

Appendix A. Removing waiting times from recorded patient trajectories

We assumed that the time interval between the timestamps of two consecutive activities, say, an activity A followed by an activity B, includes not only the execution time of activity B but also potentially a waiting time for activity B. Formally, we express the recorded time duration τ_{AB} to include both the waiting time w_{AB} for and execution time e_{AB} of an activity B, from an activity A:

$$\tau_{AB} = w_{AB} + e_{AB}. \tag{5}$$

To remove the waiting times, we performed temporal conformance checking wherein we checked whether the recorded time duration τ_{AB} deviated from the average time duration between activities A and B (Stertz et al., 2020). First, we identified the average time duration between every activity A and B by calculating its median Mdn_{AB} and mean absolute deviation MAD_{AB} across all patient trajectories in the test set $\mathcal{D}_{test}$. Second, for every activity A to B in a patient's recorded trajectory, we calculated the modified z-score of the recorded time duration τ_{AB}:

$$z = \frac{0.6745(\tau_{AB} - Mdn_{AB})}{MAD_{AB}}. \tag{6}$$

Every recorded time duration τ_{AB} that has a modified z-score greater than a prespecified z-score threshold k is considered to be a *temporal deviation* and thus includes a waiting time. On the contrary, if the modified z-score is less than or equal to the z-score threshold k, then the recorded time duration τ_{AB} is considered to be in conformance to the expected duration and thus has no waiting time ($\tau_{AB} = e_{AB}$).

We then removed the waiting times from the recorded time durations recognised as temporal deviations. We generated the updated time duration τ'_{AB} of every activity A to B in the recorded trajectories as follows:

$$\tau'_{AB} = \begin{cases} \tau_{AB} & \text{if } z \leq k \\ Mdn_{AB} + k(1.4826 * MAD_{AB}) & \text{if } z > k. \end{cases} \tag{7}$$

We used these updated time durations to specify the execution time of activities in the `trajectory` of the new patient p in the ED ABM. In the results shown in Section 3, the z-score threshold $k = 3$. This removal of waiting times from recorded patient trajectories was determined through calibration experiments.