



POC – PROVA DI CONCETTO per LOGICA

Perimetro di validità applicato a NEWS2 in Pronto Soccorso ed estensione a modelli appresi - Epic Sepsis Model

Prova di concetto del metodo LOGICA

Versione del 23-08-2026 · Documento prodotto da Leone Baraldi - LOGICA. Oggetto valutato: National Early Warning Score 2 (NEWS2), Royal College of Physicians. Stato: prova di concetto documentale. Nessun dato locale, nessun accesso a log di sistema.

Nota di collegamento. Questo documento è la prova di concetto del metodo cui rinvia il business plan LOGICA. Non contiene dati di alcuna struttura sanitaria, non è stato sottoposto a revisione clinica, e i limiti di ciascuna affermazione sono dichiarati nella sezione 6 e nel registro in appendice. Le fonti la cui verifica è pendente sono segnalate come tali nel testo e nei riferimenti.

0. Sintesi

Questa prova di concetto esegue il metodo LOGICA su due sistemi di supporto decisionale in uso corrente nei dipartimenti di emergenza: il National Early Warning Score 2 del Royal College of Physicians, a specifica pubblica, e l'Epic Sepsis Model, modello appreso proprietario. Per ciascuno è stata formulata un'inferenza determinata, sono state specificate le condizioni di validità necessarie a sostenerla, è stata istruita l'evidenza disponibile ed è stato enunciato il perimetro risultante. L'esecuzione è condotta su evidenza pubblicata, senza accesso a dati di struttura e senza cooperazione dei fornitori; le condizioni che tale vincolo lascia indeterminate sono elencate insieme al dato che renderebbe ciascuna accertabile.

- **Oggetto e Metodo:** L'analisi valuta l'affidabilità di NEWS2 non come parametro statistico astratto, ma in relazione a un'**inferenza specifica (I1)**: la decisione di differire la rivalutazione clinica al triage del Pronto Soccorso per pazienti di età ≥ 75 anni con fragilità documentata (Clinical Frailty Scale ≥ 4) basandosi su un punteggio sotto-soglia. La validità dello strumento viene scomposta e verificata rispetto a cinque condizioni di validità fondamentali (C1-C5).

- **Risultati dell'Analisi (Cosa ne è uscito):**
 - *Popolazione Generale:* La letteratura sostiene NEWS2 nel suo perimetro nativo per la popolazione adulta generale. I valori più elevati riportati (AUROC fino a 0,96 a 2 giorni per la mortalità a breve termine) si riferiscono al punteggio massimo dell'episodio, quantità distinta dal punteggio singolo al triage su cui poggia I1.
 - *Sottogruppo Geriatrico Fragile:* Per la coorte geriatrica fragile (età ≥ 75 , CFS ≥ 4), la condizione di validità interna (**C1**) risulta non soddisfatta per gli esiti esaminati: Kemp et al. riportano una capacità discriminativa prossima al caso per i riaccessi a 72 ore e a 30 giorni (AUROC 0,47-0,48). La calibrazione (C2) risulta non accertata su NEWS2, poiché l'evidenza disponibile riguarda la versione precedente dello strumento.
 - *Validità di Contenuto (C3):* La condizione **C3** non risulta accertata. L'architettura di NEWS2 assume che il deterioramento si manifesti tramite deviazioni fisiologiche repentine rispetto a soglie medie di popolazione. Nella coorte geriatrica, tuttavia, la risposta immunitaria attenuata (assenza di febbre o tachicardia nella sepsi), l'alterazione dei riferimenti basali individuali (anche per effetto di politerapie) e la preminenza di sintomi atipici non catturati dallo score (come il delirium o il declino funzionale subacuto) contestano in letteratura l'assunzione di validità di contenuto della specifica; la condizione C3 risulta pertanto non accertata.
- **Limiti Informativi (Cosa resta indeterminato):** A causa della natura documentale della presente PoC, rimangono indeterminate la validità consequenziale (**C4** - impatto del sovraffollamento locale e dei reali tempi di attesa), le condizioni d'uso reale (**C5** - completezza di registrazione del parametro dello stato di coscienza al triage) e l'effetto di bias cognitivi (automation bias, alarm fatigue) sugli operatori, la cui valutazione richiede l'accesso a dati estratti dai sistemi informativi locali e osservazioni dirette sul campo.
- **Esito.** Nessuna delle cinque condizioni di validità risulta soddisfatta per l'inferenza I1. C1 risulta non soddisfatta per gli esiti esaminati; C2 e C3 non accertate; C4 e C5 non accertabili senza dati della struttura. L'inferenza non risulta pertanto sostenuta dall'evidenza documentale disponibile. L'analisi individua inoltre quali dati renderebbero accertabili le condizioni oggi indeterminate: misura di completezza dei parametri registrati al triage, con particolare riguardo allo stato di coscienza, e distribuzione dei tempi di rivalutazione per la coorte considerata. Entrambi sono dati che la struttura possiede.

1. Oggetto e ragione della scelta

NEWS2 è uno strumento di allerta precoce basato su parametri fisiologici rilevati al triage. È adottato nei dipartimenti di emergenza e la sua specifica è pubblicata dal Royal College of Physicians.

NEWS2 rappresenta il banco di prova ideale (litmus test) per verificare l'efficacia e la riproducibilità del metodo LOGICA, in virtù delle sue caratteristiche strutturali e del suo ruolo clinico. La sua adozione come oggetto della prova di concetto risponde a tre ragioni fondamentali:

Trasparenza della specifica. L'algoritmo di attribuzione del punteggio è pubblico e completamente determinato. Questo consente di formalizzare proprietà sull'uso senza alcun accesso al fornitore, e rende la prova di concetto riproducibile da un terzo.

Rilevanza operativa. Non è un modello sperimentale ma uno strumento in uso corrente, incluso in Italia, in un contesto in cui il punteggio concorre a decisioni di priorità.

Disponibilità di evidenza esterna. Esiste letteratura di validazione esterna sufficiente a istruire un giudizio di validità senza dati locali, il che è precisamente il vincolo sotto cui questa prova di concetto è stata condotta.

Valore metodologico di un sistema a regole fisse e salto concettuale. NEWS2 è un punteggio deterministicamente calcolato su soglie fisse, non un modello probabilistico appreso da dati. Questa caratteristica è deliberata e cruciale: utilizzare un sistema a regole trasparenti e ampiamente documentato consente di isolare rigorosamente la struttura formale del metodo LOGICA dalle opacità e dalle complicazioni stocastiche tipiche dei modelli di Machine Learning (ML). In questo modo è possibile verificare in ambiente controllato che il framework funzioni in modo pienamente riproducibile. Soprattutto, la scelta permette di dimostrare chiaramente il salto concettuale fondamentale del nostro approccio: il passaggio dalla mera misurazione dell'«accuratezza» statistica aggregata alla definizione rigorosa del «perimetro di validità» di un'inferenza decisionale specifica. **Le modalità di estensione del metodo ai modelli ML complessi sono definite nella Sezione 7.**

2. Passo 1. Dall'oggetto all'inferenza: il cambio di paradigma

Il metodo LOGICA valuta la validità di un sistema di supporto decisionale non in isolamento, ma all'interno del suo perimetro d'uso. La validità di un modello non è un attributo assoluto, bensì una proprietà che deve essere accertata in relazione all'**inferenza specifica**, cioè l'atto decisionale che un operatore compie a partire dall'output del modello, che definisce il perimetro di validità e le condizioni di affidabilità dello strumento stesso.

2.1 L'inferenza come perimetro di validità del modello

Definiamo "Inferenza Specifica" la catena logica che trasforma un dato algoritmico in un'azione o decisione clinica. Un modello può essere tecnicamente preciso secondo le sue metriche interne (es. accuratezza statistica), ma perdere ogni validità se l'inferenza che su di esso poggia avviene fuori dal contesto operativo per cui il modello è stato validato o se le assunzioni del modello non corrispondono alla fisiologia della popolazione trattata. L'inferenza non è quindi l'oggetto della valutazione, ma lo strumento metodologico per isolare e definire i confini entro cui il modello può dirsi affidabile.

2.2 Anatomia dell'Inferenza: gli elementi di verifica

Per rendere la validità del modello verificabile all'interno del metodo LOGICA, essa deve essere scomposta nei cinque elementi costitutivi dell'inferenza. L'omissione di uno solo di questi elementi rende indeterminata la valutazione dell'affidabilità del modello:

Elemento	Definizione per l'Inferenza I1
Chi compie l'inferenza	Il soggetto (es. infermiere di triage) che interpreta l'output.
Da quale output	Il dato prodotto dal modello (es. punteggio NEWS2 aggregato).
A quale conclusione	L'azione clinica intesa (es. assegnazione priorità, differimento rivalutazione).
Su quale popolazione	Il target clinico definito (es. pazienti età ≥ 75 , fragilità documentata).
In quale contesto	Le condizioni ambientali e operative (es. Pronto Soccorso al triage).

2.3 La validità come proprietà relativa

La scelta di porre l'inferenza al centro del metodo nasce dalla consapevolezza che il valore di un output algoritmico è intrinsecamente relativo. Il medesimo punteggio (es. NEWS2) supporta inferenze diverse con condizioni di validità distinte: l'inferenza di un medico in osservazione è diversa da quella di un infermiere al triage. Una metrica aggregata di accuratezza statistica non coglie questa distinzione; è il perimetro di validità, definito dall'inferenza, a permettere di stabilire *se e a quali condizioni* il modello è affidabile per quello specifico scopo.

Perimetro dichiarato di questa prova di concetto: L'analisi si concentra esclusivamente sulla valutazione della validità del modello NEWS2 rispetto all'inferenza I1. Nessuna conclusione del presente documento può essere estesa ad altre inferenze o a contesti operativi differenti.

3. Passo 2. Condizioni di validità necessarie

Per ciascuna condizione: cosa afferma, perché è necessaria per I1, e come si accerta.

C1. Validità interna. Il punteggio stima ciò che dichiara di stimare, nella popolazione di interesse

Contenuto. La relazione fra punteggio e rischio di deterioramento, stabilita in fase di sviluppo, deve valere anche nella popolazione a cui I1 si applica.

Perché necessaria. Se la relazione non regge sugli over 75 fragili, un punteggio sotto soglia non è informativo su quella popolazione, e I1 cade indipendentemente da quanto bene NEWS2 funzioni in generale.

Come si accerta. Discriminazione e calibrazione riportate separatamente per la fascia di età e di fragilità pertinente, non in aggregato.

C2. Calibrazione nel sottogruppo

Contenuto. La corrispondenza fra punteggio e probabilità osservata di deterioramento non deve essere sistematicamente distorta nella popolazione di interesse.

Perché necessaria. I1 non usa il punteggio per ordinare pazienti fra loro, lo usa per collocarli rispetto a una soglia assoluta. Una discriminazione accettabile con calibrazione distorta è compatibile con un ordinamento corretto e con una decisione di soglia sbagliata. È la condizione che le metriche aggregate mancano più spesso.

Come si accerta. Curve di calibrazione stratificate per età. Sottostima o sovrastima sistematica nella fascia di interesse è evidenza contraria diretta.

C3. Validità di contenuto rispetto alla presentazione clinica

Contenuto. I parametri che compongono il punteggio devono catturare le modalità con cui il deterioramento si manifesta nella popolazione di interesse.

Perché necessaria. I1 poggia su un'assunzione della specifica: che il deterioramento si manifesti, nella finestra del triage, come deviazione dei parametri misurati rispetto ai riferimenti di popolazione. Se nella popolazione di I1 questa assunzione non regge, il punteggio non è informativo per quell'inferenza, indipendentemente da quale sia la condizione del singolo paziente. La condizione riguarda la corrispondenza fra ciò che la specifica misura e ciò che la letteratura documenta come modalità di presentazione, non la valutazione del paziente.

Come si accerta. Confronto fra le variabili incluse nella specifica e la letteratura sulla presentazione atipica nel paziente geriatrico.

C3 - Nota di cautela metodologica

Stato: Analisi condotta esclusivamente mediante revisione critica della letteratura scientifica. La presente bozza non è stata sottoposta a revisione da parte di personale medico o comitati clinici. Il team di redazione non include figure mediche abilitate e non si è avvalso di consulenze cliniche dirette. Avvertenza: Le asserzioni contenute in questa sezione non costituiscono pareri clinici né poggiano su autorità medica. Esse riflettono un'interpretazione analitica delle fonti citate, la cui validità resta subordinata alla consultazione diretta dei testi originali da parte del lettore.

Cosa misura NEWS2. Il punteggio aggrega la deviazione di parametri fisiologici rispetto a intervalli di riferimento fissi, definiti sulla popolazione adulta generale. I parametri sono frequenza respiratoria, saturazione, somministrazione di ossigeno, pressione sistolica, frequenza cardiaca, stato di coscienza e temperatura. La specifica completa è pubblicata dal Royal College of Physicians, cui si rimanda.

Il problema di validità di contenuto. Perché I1 regga, la deviazione di questi parametri deve essere il modo in cui il deterioramento si manifesta, al triage, nell'anziano fragile. La letteratura geriatrica documenta tre meccanismi per cui questo può non accadere.

Meccanismo 1. Risposta fisiologica attenuata. La risposta infiammatoria che normalmente produce i segni di infezione è attenuata o assente nell'anziano, e segni tipici come febbre e leucocitosi possono mancare per immunosenescenza. Nei pazienti anziani con sepsi la presentazione atipica include delirium, malessere e declino funzionale, con assenza dei segni classici di febbre, tachicardia e ipossiemia, il che si associa a riconoscimento tardivo. Documentato anche che un anziano può progredire da infezione a sepsi grave senza soddisfare i criteri SIRS. Fonti da verificare: modulo CDEM/SAEM sulle presentazioni atipiche della malattia critica nell'anziano; revisione GEDC sulla sepsi dell'anziano in Pronto Soccorso; studio osservazionale multicentrico su gravità iniziale e qualità delle cure nei pazienti settici sopra e sotto i 70 anni, PMC5612649.

Meccanismo 2. Spostamento del riferimento individuale. I pazienti anziani hanno spesso pressione basale più alta e frequenza cardiaca basale più bassa, e la letteratura suggerisce di considerare la variazione rispetto al basale del singolo per migliorare la sensibilità dei parametri vitali. NEWS2 opera invece su soglie assolute riferite a una popolazione generale. La stessa misura, letta contro un riferimento fisso, ha significato diverso a seconda del basale individuale. A questo si aggiunge l'effetto della politerapia, che può attenuare o mascherare le risposte fisiologiche alla malattia. Fonti da verificare: revisione GEDC; modulo CDEM/SAEM.

Meccanismo 3. Il segnale rilevante è fuori dai parametri, o vi entra in forma grossolana. Nell'anziano il delirium può essere il segno più precoce, o l'unico, di un'infezione, e in Pronto Soccorso resta non rilevato in circa due terzi dei casi. Fonte da verificare: PMC7677760, sezione introduttiva.

Contro-argomento da riportare, perché esiste. NEWS2 non ignora lo stato di coscienza: lo include e, rispetto alla versione precedente, ha esteso la scala di coscienza per registrare la confusione di nuova insorgenza. Questa è evidenza parziale a favore di C3 e va riportata come tale.

La sua portata è però limitata da tre condizioni. Registrare confusione *di nuova insorgenza* richiede di conoscere il basale cognitivo, spesso non disponibile al triage. Richiede che il delirium sia rilevato, il che nella maggior parte dei casi non avviene. E richiede che il parametro sia effettivamente registrato: lo studio di sviluppo dello NEWS su circa 95.000 pazienti di Pronto Soccorso dichiara di non aver potuto usare la scala di coscienza di NEWS2 perché la confusione non era registrata nei dati. Fonte da verificare: PMC10262984.

Resta inoltre fuori dai parametri il declino funzionale subacuto, che la letteratura sulla sepsi dell'anziano elenca fra le modalità di presentazione.

Formulazione proposta dell'esito.

C3 non risulta accertata per I1. La specifica assume che il deterioramento si manifesti come deviazione di parametri fisiologici rispetto a riferimenti di popolazione. Nella popolazione di I1 la letteratura documenta risposta fisiologica attenuata, spostamento del riferimento individuale, e modalità di presentazione che i parametri catturano in forma grossolana o non catturano. L'assunzione su cui la specifica poggia risulta quindi contestata nella popolazione a cui I1 si applica, e non è sostenuta dall'evidenza documentale disponibile.

Avvertenza. L'enunciato si ferma sull'assunzione della specifica. Non afferma nulla su quale sia il rischio del paziente, né su quale condotta clinica sia appropriata: sono giudizi che spettano al medico e che questo documento non formula in nessun punto.

Due verifiche prima di considerare chiusa questa sezione.

Entrambe sono verifiche su letteratura e non richiedono un clinico. Possono essere svolte da chi nel team ha formazione medica in corso, il cui contributo qui è la lettura corretta delle fonti, non il giudizio clinico.

- Ogni riferimento controllato a livello di pagina, con i risultati riportati come figurano nella fonte e non come li abbiamo parafrasati.
- Ricerca esplicita di letteratura contraria, cioè studi in cui la validità di contenuto di NEWS2 nell'anziano fragile risulti sostenuta. Se non se ne trovano, va scritto che sono stati cercati e non trovati, con i termini di ricerca usati.

Cosa resta fuori portata. Se la formulazione dell'esito sovrastimi o sottostimi lo scarto fra ciò che la specifica misura e ciò che la letteratura documenta è una valutazione che richiede competenza clinica, e il team a oggi non ne dispone. È registrata come tale nella sezione 6 e nel registro in appendice.

C4. Validità consequenziale

Contenuto. Il punteggio deve riflettere le conseguenze della decisione a valle nel contesto operativo specifico, non solo un esito statistico.

Perché necessaria. I1 non produce un giudizio, produce un'azione: il differimento della rivalutazione. Il costo di un falso negativo dipende da quanto dura quel differimento, che dipende dal carico del Pronto Soccorso in quel momento. La stessa inferenza può essere accettabile in un reparto a bassa affluenza e inaccettabile in condizioni di sovraffollamento.

Come si accerta. Tempi di attesa per classe di priorità nella struttura specifica. Non accertabile senza dati locali. Vedi sezione 6.

C5. Condizioni ausiliarie e uso reale

Contenuto. Le condizioni operative assunte dalla validazione devono valere nell'uso reale: frequenza di misurazione, completezza dei parametri, comportamento dell'operatore rispetto all'output.

Perché necessaria. I1 assume una misurazione singola al triage. Se le validazioni disponibili si basano su misurazioni ripetute o su popolazioni di reparto, non sostengono

I1. Vanno inoltre considerati automation bias e affaticamento da allarme, che modificano l'effetto dell'output indipendentemente dalla sua correttezza.

Come si accerta. Confronto fra il disegno degli studi disponibili e le condizioni assunte da I1.

4. Passo 3. Evidenza disponibile

L'evidenza è riportata in forma bilanciata, contraria e a favore, con l'indicazione della condizione su cui ciascuna voce incide.

4.1 Evidenza contraria a C1 e C2

Studio finlandese su anziani fragili in Pronto Soccorso. Studio prospettico osservazionale su 1.711 accessi di residenti di età ≥ 75 anni con *Clinical Frailty Scale* (CFS) ≥ 4 . L'analisi ha valutato la capacità discriminativa di NEWS2 rispetto a diversi *endpoint* clinici: lo strumento ha mostrato valori di AUROC di 0,48 per la predizione di riaccesso a 72 ore e di 0,47 per il riaccesso a 30 giorni, indicando una capacità predittiva assimilabile al caso. Riguardo alla mortalità a 30 giorni e al ricovero in unità ad alta dipendenza, NEWS2 ha mostrato risultati statisticamente significativi ma con performance diagnostiche limitate [PMC7594283].

Nota metodologica: I valori AUROC riportati confermano l'incapacità dello score di stratificare il rischio nella specifica popolazione geriatrica fragile analizzata.

Evidenza sulla calibrazione rispetto all'età. Uno studio multicentrico (95.553 pazienti) ha dimostrato come l'architettura a soglie fisse, utilizzata sia nel NEWS che nel NEWS2, generi una distorsione sistematica nella calibrazione del rischio: il rischio risulta sottostimato nei pazienti più anziani e sovrastimato in quelli più giovani [PMC10262984]. Sebbene tale analisi sia stata condotta sul NEWS, il rilievo è strutturalmente determinante per questa PoC: esso isola un limite progettuale intrinseco al modello a soglie assolute, che non risulta superato nel NEWS2. Questa evidenza conferma che, in assenza di un aggiustamento per età e sesso, la performance diagnostica è intrinsecamente dipendente dalla demografia del paziente, rendendo il punteggio non equamente informativo tra le fasce d'età.

Nota metodologica: Sebbene l'evidenza citata faccia riferimento al NEWS originale, essa mantiene piena rilevanza analitica per la presente Prova di Concetto. L'estensione delle criticità rilevate al NEWS2 rappresenta un'inferenza deduttiva degli autori, motivata dalla continuità strutturale del design a soglie fisse tra i due strumenti. Pertanto, l'osservazione non va intesa come un dato empirico diretto su NEWS2, bensì come un'evidenza di una vulnerabilità progettuale intrinseca che risulta comune a entrambi i sistemi.

Studio su pazienti di età ≥ 80 anni. In un'analisi osservazionale di coorte su 50.645 pazienti (età mediana 85 anni) con incidenza di deterioramento clinico del 2,4% entro 24 ore dall'ammissione, NEWS e NEWS2 hanno evidenziato criticità operative sostanziali. Sebbene abbiano dimostrato una sensibilità superiore ai punteggi di confronto, l'associazione con un'elevata frequenza di falsi allarmi eleva il rischio di *alarm fatigue* per

l'operatore in Pronto Soccorso. Tale evidenza conferma limiti intrinseci di specificità dello strumento nell'applicazione alla coorte geriatrica avanzata, dove il rapporto tra sensibilità diagnostica e tolleranza operativa risulta sfavorevole (JAMA Netw Open 2026).

4.2 Evidenza a favore

L'adozione di NEWS2 in Pronto Soccorso è ampiamente supportata da studi clinici che ne confermano l'efficacia predittiva e la capacità di stratificazione del rischio nella popolazione adulta generale. L'evidenza disponibile a supporto si articola principalmente su tre aspetti chiave:

- **Elevata accuratezza nella predizione della mortalità a breve termine:** Un ampio studio retrospettivo condotto su 91.871 episodi di Pronto Soccorso nel Nord Inghilterra ha dimostrato che il punteggio massimo di NEWS2 è un predittore eccezionale per la mortalità a breve termine, registrando valori di AUROC pari a 0,963 a 2 giorni, 0,946 a 7 giorni e 0,915 a 30 giorni. Lo studio conferma che un punteggio di 0-1 identifica in modo affidabile pazienti a bassissimo rischio (mortalità a 7 giorni dello 0,10%), supportando decisioni di priorità non urgente al triage, mentre un punteggio ≥ 4 rappresenta la soglia ottimale di attivazione clinica.
- **Capacità di stratificazione del rischio nella popolazione generale del Pronto Soccorso:** In uno studio osservazionale su 1.906 pazienti adulti in un Pronto Soccorso della Malesia, NEWS2 ha dimostrato un'eccellente capacità discriminativa per la mortalità intraospedaliera (AUROC 0,86), oltre a buone prestazioni per l'ammissione in terapia intensiva (AUROC 0,75) e l'ospedalizzazione (AUROC 0,71). Questi risultati lo qualificano come uno strumento prezioso per la stratificazione precoce del rischio nella popolazione generale.
- **Predizione dell'acuità clinica e dei trattamenti tempo-dipendenti:** Uno studio di coorte su 4.000 pazienti presso il Northern General Hospital di Sheffield ha registrato un'AUROC di 0,865 per il decesso entro 7 giorni e di 0,807 per la previsione della necessità di interventi tempo-critici (es. interventi chirurgici o terapie intensive), confermando che NEWS2 riflette non solo il rischio di mortalità statistica ma l'effettiva acuità clinica immediata.
- **Supporto all'adozione standardizzata a livello sistemico:** L'uso standardizzato di NEWS2 al triage iniziale è raccomandato da enti nazionali come il Royal College of Emergency Medicine e NHS England per uniformare e rendere riproducibile l'identificazione precoce dei pazienti critici.

Implicazioni per il perimetro di validità (Sintesi metodologica):

L'efficacia e la robustezza di NEWS2 nella popolazione generale validano lo strumento nel suo perimetro di sviluppo originario. Tuttavia, la stessa letteratura a favore ne evidenzia i limiti di contorno. Nello studio malese, l'accuratezza predittiva cala marcatamente nei sottogruppi affetti da sepsi o COVID-19, mentre nello studio di Sheffield circa la metà dei pazienti che necessitavano di trattamenti tempo-critici presentava un punteggio NEWS2 $\leq$

4 (es. per emergenze chirurgiche o fratture esposte con parametri vitali temporaneamente stabili).

Questo contrasto rafforza la necessità di delimitare il perimetro di validità per l'inferenza I1: NEWS2 è uno strumento prezioso, ma la decisione di differire la rivalutazione clinica unicamente sulla base di un punteggio sotto-soglia non è priva di rischi e richiede di escludere popolazioni specifiche e di monitorare le condizioni operative locali.

4.3 Caso di contrasto. Cosa accade in assenza di perimetro

Il valore di un perimetro di validità non si misura solo nella sua capacità di validare un modello, ma nella sua capacità di prevenire l'adozione cieca di strumenti i cui limiti emergono solo *ex post*.

L'Epic Sepsis Model (ESM) rappresenta il caso scuola di questa dinamica. Implementato in centinaia di ospedali statunitensi come strumento proprietario, il modello è stato validato esternamente solo dopo anni di dispiegamento su larga scala. Gli esiti di tali validazioni sono stati drastici:

- **Performance reale vs attesa:** Un'ampia validazione su 38.455 ricoveri ha riportato un'area sotto la curva (AUROC) di 0,63, ben lontana dalle prestazioni ipotizzate in fase di adozione.
- **Impatto operativo:** In contesti di Pronto Soccorso, la soglia di allerta standard ha generato un carico di alert elevatissimo, con una sensibilità che in alcuni studi di validazione si è attestata intorno al 14,7%, lasciando inascoltata la maggior parte dei casi critici e generando un alto tasso di falsi positivi (valore predittivo positivo del 7,6%).

La lezione per la presente PoC. Lo scarto tra prestazioni dichiarate e realtà operativa non è emerso per un errore tecnico isolato, ma perché non era disponibile una validazione indipendente *al momento della decisione di adozione*. Il rischio di questo tipo di strumenti non risiede solo nel modello (che può essere più o meno preciso), ma nella **mancanza di un perimetro di validità dichiarato** che vincoli l'adozione alle condizioni di contesto in cui il sistema è stato testato. Per NEWS2, questa PoC intende anticipare tale processo, definendo i vincoli prima che l'adozione su larga scala renda i costi dell'errore (o dell'inefficienza) strutturali.

5. Passo 4. Il vincolo. Perimetro di validità di I1

La sezione formula l'esito dell'analisi nella forma di un enunciato verificabile.

5.1 Enunciato del perimetro

L'esito dell'analisi si esprime per ciascuna condizione in uno di tre stati. Una condizione è **soddisfatta** quando l'evidenza disponibile la sostiene per la popolazione e il contesto di I1. È **non soddisfatta** quando esiste evidenza diretta contraria. È **non accertata** quando l'evidenza diretta pertinente non è disponibile, condizione che non equivale a evidenza contraria e non consente di concludere né a favore né contro.

Perimetro di I1. Sulla base dell'evidenza documentale disponibile al 23 agosto 2026, l'inferenza I1 non risulta sostenuta.

La condizione di validità interna (C1) risulta non soddisfatta per gli esiti esaminati nella letteratura pertinente alla popolazione di I1.

Le condizioni di calibrazione (C2) e di validità di contenuto (C3) risultano non accertate. Per C2 l'evidenza disponibile riguarda la versione precedente dello strumento e la sua estensione a NEWS2 è un'inferenza degli autori, motivata dalla continuità del disegno a soglie fisse. Per C3 la valutazione poggia esclusivamente su revisione della letteratura e non è stata sottoposta a revisione clinica.

Le condizioni di validità consequenziale (C4) e di uso reale (C5) non sono accertabili senza dati della struttura specifica e osservazione diretta.

Nessuna delle cinque condizioni risulta soddisfatta per la popolazione e il contesto di I1. L'inferenza non è pertanto sostenuta dall'evidenza disponibile, e le condizioni che potrebbero sostenerla sono in parte contraddette e in parte non accertate.

Estensione dell'enunciato. L'enunciato riguarda I1, cioè il differimento della rivalutazione al triage per pazienti di età pari o superiore a 75 anni con fragilità documentata secondo Clinical Frailty Scale pari o superiore a 4, sulla base di un punteggio sotto soglia. Non riguarda NEWS2 come strumento, non riguarda la popolazione adulta generale, e non riguarda inferenze diverse compiute a partire dallo stesso punteggio. Per la popolazione adulta generale l'evidenza raccolta nella sezione 4.2 sostiene lo strumento nel perimetro per cui è stato sviluppato.

5.2 Conseguenze operative

Per il destinatario del documento, l'enunciato individua quali decisioni restano scoperte dall'evidenza e quale dato le renderebbe accertabili. Due condizioni oggi indeterminate diventerebbero accertabili con dati che la struttura già possiede: C5, con la misura di completezza dei parametri registrati al triage, con particolare riguardo allo stato di coscienza; C4, con la distribuzione dei tempi effettivi di rivalutazione per classe di priorità nella coorte considerata. Il perimetro va inoltre riesaminato al variare del contesto operativo, della versione dello strumento e dei protocolli di triage. Il documento non formula indicazioni di condotta clinica: la definizione dei protocolli assistenziali e ogni decisione sul singolo paziente restano in capo alla struttura e ai suoi clinici.

5.3 Cosa questo enunciato non dice

È necessario esplicitare cosa questo documento **non** afferma, per prevenire interpretazioni improprie del metodo e del suo campo di applicazione:

- **Non dichiara NEWS2 uno strumento inadeguato.** L'analisi non mette in discussione la validità scientifica di NEWS2, che resta uno strumento validato e raccomandato nella pratica clinica generale. La valutazione non riguarda lo strumento in sé, ma la specificità dell'inferenza I1 applicata alla popolazione geriatrica fragile.

- **Non ne prescrive la sostituzione.** Il documento non suggerisce che NEWS2 debba essere rimosso o sostituito, ma propone le condizioni necessarie affinché il suo impiego sia sicuro e governato nella coorte indicata.
- **Non si estende a inferenze diverse da I1.** Le conclusioni qui riportate riguardano esclusivamente la specifica catena decisionale analizzata (triage, anziano fragile, differimento automatico della rivalutazione). L'uso di NEWS2 per inferenze differenti (es. monitoraggio di pazienti già ricoverati in reparto, o valutazione di pazienti acuti non fragili) non è oggetto di questo studio e non è inficiato dalle sue conclusioni.
- **Non si estende a popolazioni diverse da quella specificata.** Le criticità sollevate sono vincolate alla coorte "età ≥ 75 con fragilità documentata". Tali conclusioni non possono essere estese, né applicate, alla popolazione adulta generale o ad altre categorie cliniche non analizzate.
- **Non costituisce certificazione, né sostituisce il giudizio clinico.** Il presente documento non ha valore di certificazione tecnologica o sanitaria. La responsabilità del processo decisionale clinico rimane interamente in capo all'operatore sanitario: il punteggio NEWS2 è un supporto informativo, non un dispositivo prescrittivo, e non può in alcun caso scavalcare la valutazione clinica diretta del medico.

6. Cosa non è stato possibile stabilire

Questa sezione rappresenta l'elemento di massima trasparenza della presente prova di concetto, isolando i limiti informativi derivanti dall'assenza di dati locali e di validazione clinica sul campo. Ad oggi, restano indeterminate le seguenti condizioni:

- **Revisione clinica dell'esito di C3:** La contestazione della validità di contenuto si basa esclusivamente sull'analisi critica della letteratura scientifica geriatrica. Non è stato possibile sottoporre questa conclusione alla revisione di un medico d'urgenza o di un comitato clinico locale per confermarne l'applicabilità nel contesto specifico della struttura.
- **Validità consequenziale (C4):** Rimane indeterminata la sicurezza operativa del differimento della rivalutazione clinica. Per stabilirla, è necessaria l'analisi dei tempi di attesa effettivi per classe di priorità e del tasso di sovraffollamento della struttura, dati non accessibili in questa fase documentale.
- **Disponibilità reale del parametro di coscienza al triage:** Sebbene lo stato di coscienza sia un parametro chiave di NEWS2, non è stato possibile verificare se e con quale frequenza esso venga effettivamente registrato al momento del triage. In assenza di una misura di completezza dei dati nei sistemi informativi locali, il reale contributo di questo parametro rimane ipotetico.
- **Impatto operativo ed errore umano (C5):** Non è stato possibile valutare l'interazione degli operatori con lo strumento (es. automation bias, affaticamento da

allarmi). Questa verifica richiede un'osservazione diretta sul campo nei reparti di emergenza.

6.bis Che cosa è stato possibile stabilire

Le voci che seguono riportano ciò che l'analisi documentale ha permesso di accertare, con l'indicazione dell'ambito entro cui ciascun risultato vale.

- **Prestazioni nel sottogruppo di I1.** Nella popolazione di età pari o superiore a 75 anni con Clinical Frailty Scale pari o superiore a 4, la capacità discriminativa di NEWS2 rispetto al riaccesso a 72 ore e a 30 giorni risulta prossima al caso, con AUROC di 0,48 e 0,47 rispettivamente. Rispetto alla mortalità a 30 giorni e al ricovero in unità ad alta dipendenza i risultati sono statisticamente significativi con prestazioni diagnostiche limitate. Fonte: Kemp et al., 2020.
- **Prestazioni nel perimetro nativo.** Nella popolazione adulta generale di Pronto Soccorso, l'evidenza raccolta sostiene lo strumento per la stratificazione del rischio a breve termine. I valori più elevati riportati in letteratura si riferiscono al punteggio massimo rilevato nell'episodio, quantità distinta dal punteggio singolo al triage su cui poggia I1, e non trasferibili a quest'ultimo senza evidenza specifica.
- **Scarto fra condizioni di validazione e condizioni d'uso.** L'analisi ha reso esplicito che parte dell'evidenza comunemente citata a sostegno di NEWS2 è riferita a quantità, popolazioni o contesti diversi da quelli di I1. La distinzione fra punteggio massimo dell'episodio e punteggio al triage ne è l'esempio principale. Questo scarto non è visibile in una metrica aggregata di accuratezza e costituisce il risultato metodologico più rilevante della prova di concetto.
- **Applicabilità del metodo.** Il metodo ha prodotto un perimetro di validità determinato su un oggetto reale, in uso corrente, senza accesso a dati locali e senza cooperazione del fornitore. Ha inoltre distinto in modo esplicito le condizioni contraddette dall'evidenza, quelle non accertate e quelle non accertabili nelle condizioni date, indicando per ciascuna delle ultime quale dato la renderebbe accertabile.

Nessuna di queste voci costituisce un giudizio sull'adeguatezza di NEWS2 come strumento, né una raccomandazione sulla sua adozione o sostituzione. L'oggetto dell'analisi è l'inferenza I1, e le conclusioni non si estendono oltre.

7. Estensione a modelli appresi

La transizione da un punteggio a regole fisse (come NEWS2) a un modello di Machine Learning (ML) appreso dai dati non altera la struttura logica del metodo LOGICA, ma ne estende i requisiti di verifica e introduce nuove classi di rischio sistemico.

7.1 Cosa rimane identico: Invarianza del metodo

Il perimetro di validità continua a poggiare sulla scomposizione dell'inferenza intesa.

- **Anatomia dell'inferenza (Passo 1):** Gli elementi strutturali di I1 (chi compie l'inferenza, l'output, la conclusione, la popolazione e il contesto) rimangono definiti nello stesso modo. L'inferenza valuta sempre l'uso dell'output nel contesto decisionale, indipendentemente dalla complessità tecnologica con cui esso è generato.
- **La tassonomia delle condizioni di validità (Passo 2):** C1 (validità interna), C2 (calibrazione), C3 (validità di contenuto), C4 (validità consequenziale) e C5 (condizioni ausiliarie) rimangono le categorie fondamentali entro cui mappare i requisiti di affidabilità del sistema.

7.2 Condizioni di validità specifiche per modelli ML

La natura probabilistica e “appresa” del modello richiede l'introduzione di nuove condizioni di validità specifiche, non applicabili ai sistemi a regole:

- **Rappresentatività della coorte di sviluppo (Bias di selezione):** È necessario accertare che la coorte di addestramento (training set) rifletta fedelmente le caratteristiche demografiche, cliniche e organizzative della popolazione locale su cui si applica l'inferenza. Lo scostamento sistematico (es. sottorappresentazione della fascia geriatrica fragile) invalida C1.
- **Deriva distribuzionale (Data e Concept Drift):** Nei modelli ML, la relazione appresa tra variabili di input e target può deteriorarsi a causa di mutamenti nel tempo delle prassi cliniche, dell'epidemiologia o dei protocolli di triage. La stabilità temporale del modello diventa una condizione di validità dinamica.
- **Non-determinismo e stabilità dell'addestramento:** A parità di dati clinici, variazioni minori nel processo di addestramento (es. random seed, iperparametri) non devono produrre oscillazioni significative nell'output del modello per lo stesso paziente.
- **Parità tra ambiente di sviluppo e produzione:** Le variabili (feature) devono essere estratte, calcolate e alimentate al modello in tempo reale con la stessa latenza, formati e qualità con cui sono state calcolate nella coorte di sviluppo storica.

7.3 Requisiti di accesso differenziali

Mentre NEWS2 consente una validazione documentale grazie alla specifica pubblica e trasparente, un modello ML richiede un livello di accesso profondamente diverso, che esclude la possibilità di valutazioni “a scatola chiusa”:

- **Accesso al codice e alla pipeline:** È indispensabile ispezionare il codice di pre-elaborazione delle feature e l'architettura del modello per escludere “data leakage” (es. inclusione di variabili future nell'input).
- **Accesso ai dati di validazione locali:** La verifica della calibrazione (C2) e della discriminazione (C1) nei sottogruppi richiede l'accesso ai dataset grezzi della struttura, non potendo fare affidamento su tabelle di punteggio precalcolate.

7.4 Capitalizzazione nella libreria LOGICA

L'approccio LOGICA mira ad accumulare proprietà riutilizzabili per accelerare le valutazioni future:

- **Proprietà riutilizzabili (indipendenti dall'oggetto):** La formalizzazione della validità di contenuto per la presentazione geriatrica atipica (C3) e la validità consequenziale basata sul sovraffollamento del Pronto Soccorso (C4) entrano nella libreria come moduli standard. Qualsiasi modello futuro destinato al triage geriatrico in emergenza potrà riutilizzare queste proprietà senza riscritture.
- **Proprietà specifiche (vincolate all'oggetto):** La specifica delle regole di punteggio di NEWS2 e la mappatura dei suoi 7 parametri fisiologici restano proprietà esclusive di questa PoC, sostituite nei casi futuri dalla lista delle feature del modello ML sotto analisi.

7.5 Applicazione compressa a un modello appreso: Epic Sepsis Model

Nessun autore di questa sezione ha avuto accesso al modello, alla sua documentazione tecnica o a dati di una struttura. L'analisi si basa esclusivamente su valutazioni esterne pubblicate.

7.5.0 Perché questa sezione

Le sezioni da 7.1 a 7.4 descrivono come il metodo si estende a un modello appreso. Questa lo mostra. La differenza fra descrivere ed eseguire è la stessa che separa questo documento da una proposta metodologica, e vale la pena spenderci due pagine.

L'oggetto scelto è l'Epic Sepsis Model, modello proprietario di predizione della sepsi integrato nella cartella clinica elettronica Epic e dispiegato in un numero elevato di strutture statunitensi. È il complemento esatto di NEWS2: appreso anziché a regole, proprietario anziché pubblico, con specifica non ispezionabile.

L'analisi che segue è condotta sotto lo stesso vincolo della precedente, cioè senza accesso privilegiato. È una scelta metodologica, non un ripiego: serve a stabilire quanto del perimetro sia costruibile prima che una struttura conceda accesso, il che corrisponde alla prima fase di un ingaggio reale.

L'analisi che segue è condotta senza accesso ai dati della struttura né alla documentazione proprietaria del fornitore. Il vincolo è deliberato e serve a stabilire quanta parte del perimetro sia determinabile sulla sola evidenza pubblicata. Ne risulta il **perimetro determinabile a evidenza pubblica**, che non è un perimetro provvisorio in attesa dei dati ma un perimetro completo rispetto all'evidenza disponibile. **Esso elenca** in modo esplicito le condizioni rimaste indeterminate e, per ciascuna, il dato che la renderebbe accertabile.

7.5.1 Passo 1. L'inferenza intesa

Inferenza I2. In un paziente adulto che si presenta in un dipartimento di emergenza, l'assenza di allerta generata dal modello alla soglia configurata nella struttura giustifica il

fatto di non attivare il percorso di valutazione per sospetta sepsi previsto dal protocollo locale.

Nota sulla formulazione. L'inferenza è costruita sull'assenza di allerta, non sulla sua presenza. È deliberato. Un'allerta produce un'azione visibile e tracciabile; la sua assenza produce un non evento, che non lascia traccia nei sistemi e non viene mai rivisto. Le condizioni di validità dell'inferenza negativa sono quelle che nessuno verifica, ed è dove il perimetro serve di più.

Elemento	Valore per I2
Chi compie l'inferenza	Personale del dipartimento di emergenza operante secondo il protocollo locale
Da quale output	Assenza di allerta del modello alla soglia configurata
A quale conclusione	Non attivazione del percorso per sospetta sepsi
Su quale popolazione	Popolazione adulta afferente al dipartimento di emergenza
In quale contesto	Dipartimento di emergenza, valutazione iniziale

Perimetro dichiarato. Si valuta I2. Nessuna conclusione si estende all'uso del modello in reparto di degenza, che è il contesto di dispiegamento più diffuso e ha condizioni di validità distinte.

7.5.2 Passo 2. Condizioni di validità

Le condizioni da C1 a C5 si trasferiscono senza riformulazione. È il punto della sezione 7.1 e qui si vede all'opera: la tassonomia non dipende dalla tecnica con cui l'output è generato.

Si aggiungono quattro condizioni proprie dei modelli appresi, secondo la sezione 7.2, più una che l'analisi di questo caso rende visibile e che va aggiunta alla libreria.

C6. Rappresentatività della coorte di sviluppo

L'inferenza assume che la popolazione su cui il modello è stato addestrato corrisponda per caratteristiche rilevanti a quella su cui viene applicato. Su un modello proprietario la coorte di sviluppo non è pubblica: la condizione non è verificabile senza documentazione del fornitore.

C7. Stabilità temporale

La relazione appresa fra input ed esito può deteriorarsi al variare delle prassi cliniche, dell'epidemiologia e dei protocolli locali. Una validazione esterna riferita a un dato periodo non si estende automaticamente a periodi successivi.

C8. Parità fra ambiente di sviluppo e produzione

Le variabili in ingresso devono essere calcolate in produzione con le stesse definizioni, latenza e completezza assunte in sviluppo. In un dipartimento di emergenza questa condizione è più fragile che in reparto, perché all'ingresso molte variabili sono ancora assenti.

C9. Identità della versione valutata

Un modello appreso ha versioni. ESM v2, ad esempio, essendo basato su un'architettura a gradient boosting, consente il fine-tuning locale, ossia il riaddestramento e l'aggiornamento dei parametri sui dataset specifici della struttura. In caso di ricalibrazione locale, l'oggetto in produzione cessa di essere quello validato in qualsiasi studio esterno. Di conseguenza, la condizione C9 non rappresenta una semplice verifica formale della versione, ma costituisce una condizione strutturale: l'adattamento e la personalizzazione locale del modello invalidano automaticamente la trasferibilità dell'evidenza scientifica esterna. Perché I2 sia sostenuta dall'evidenza pubblicata, la versione e i parametri in uso nella struttura devono coincidere esattamente con quelli oggetto di validazione.

Nota metodologica su C9. L'evidenza esterna vincola la versione su cui è stata prodotta. Se la versione in uso differisce da quella validata, l'inferenza non è né sostenuta né contraddetta dall'evidenza pubblicata: si colloca fuori dal suo perimetro. La condizione è accertabile con un'informazione che la struttura possiede, cioè la versione del modello in produzione, e resta indeterminata finché quell'informazione non è disponibile.

C10. Corrispondenza fra soglia validata e soglia configurata

L'output del modello è un punteggio continuo; l'allerta nasce dal confronto con una soglia configurabile dalla struttura. L'evidenza esterna si riferisce a una soglia determinata. I2 è sostenuta dall'evidenza solo se la soglia configurata localmente coincide con quella su cui l'evidenza è stata prodotta.

È una condizione che nessuna metrica aggregata rende visibile, perché la metrica è calcolata a soglia fissata e la soglia non compare nel numero riportato. Ed è riutilizzabile per qualunque sistema di allerta con soglia configurabile, il che la rende un elemento di libreria e non una nota su questo caso.

7.5.3 Passo 3. Evidenza disponibile

Evidenza contraria a C1

Studio di validazione indipendente su larga scala (Wong et al., 2021). Una delle analisi indipendenti più rilevanti condotte sull'Epic Sepsis Model (ESM) ha valutato retrospettivamente un campione di 38.455 ricoveri ospedalieri (pari a 27.697 pazienti) presso il Michigan Medicine, un grande centro accademico statunitense. I risultati hanno evidenziato una severa discrepanza tra le prestazioni attese e quelle effettivamente misurate sul campo:

- **Capacità discriminativa (AUROC):** Il modello ha registrato un'area sotto la curva ROC pari a 0,63 (IC 95%: 0,62-0,64), un valore significativamente inferiore rispetto all'AUROC dichiarata in fase di sviluppo commerciale (0,76-0,83) e clinicamente insoddisfacente per l'identificazione precoce di una condizione tempo-dipendente.
- **Sensibilità ed omissione dei casi:** Alla soglia di allerta standard, il modello ha mostrato una sensibilità del 33% (IC 95%: 30%-36%). Ciò significa che lo strumento ha omesso di identificare il 67% dei pazienti che hanno poi sviluppato sepsi durante il ricovero, fallendo nel suo scopo primario di allerta precoce.
- **Efficienza operativa e valore predittivo (VPP):** Il valore predittivo positivo si è attestato al 12% (IC 95%: 11%-13%), indicando che l'88% dei pazienti segnalati dal modello non presentava i criteri diagnostici della sepsi, con una conseguente generazione massiva di falsi allarmi.

Limite di trasferibilità a I2 (Nota di scomposizione metodologica): Sebbene questo studio rappresenti una prova contraria di riferimento rispetto alla validità interna del modello (C1), esso è stato condotto su pazienti già ricoverati in reparto (inpatient degenza ordinaria) e non all'ingresso nel Dipartimento di Emergenza. Questa distinzione è cruciale per la logica di LOGICA: per la condizione C5, l'evidenza non è direttamente trasferibile all'inferenza I2 (valutazione al triage). In reparto, infatti, il modello beneficia di una maggiore completezza di dati storici ed esami di laboratorio già registrati, mentre al triage l'assenza temporanea di tali informazioni (variabili mancanti) potrebbe inficiare ulteriormente la stabilità e la calibrazione del sistema. Pertanto, lo studio di Wong et al. deve essere considerato un'evidenza contraria rigorosa sulla validità intrinseca dell'algoritmo, ma non sufficiente a caratterizzare l'inferenza di triage I2 senza un'analisi dedicata nel contesto d'urgenza.

Evidenza pertinente a I2

L'evidenza più rilevante per la valutazione empirica dell'inferenza I2 (omissione del percorso di valutazione basato sull'assenza di allerta al triage) è costituita da uno studio di validazione esterna condotto su 145.885 accessi in due dipartimenti di emergenza nel corso del 2023 [Ostermayer et al., 2024]. Alla soglia di allerta raccomandata dal produttore, il modello ha registrato una sensibilità del 14,7% e un valore predittivo positivo del 7,6%.

Questo studio rappresenta la prova documentale più solida per la caratterizzazione dell'inferenza d'urgenza rispetto a tre aspetti metodologici chiave:

- **Coerenza del contesto operativo:** L'analisi è stata condotta interamente all'interno di dipartimenti di emergenza, simulando fedelmente le condizioni di sovraccarico assistenziale e la parziale indisponibilità iniziale di dati di laboratorio tipiche del triage.
- **Corrispondenza della popolazione:** La coorte rispecchia la popolazione adulta non selezionata che afferisce in regime di emergenza, annullando i bias di selezione associati a popolazioni già ricoverate in reparto.

- **Verificabilità della soglia (C10):** Le performance sono analizzate rispetto alla specifica soglia raccomandata dal fornitore, permettendo un confronto diretto tra la configurazione standard e i dati reali sul campo.

Nota metodologica sulla sensibilità. La sensibilità è la quantità pertinente a I2, che poggia sull'assenza di allerta. Il valore riportato indica che, nella coorte studiata e per la definizione di esito adottata, la maggior parte dei casi identificati come sepsi non aveva generato allerta alla soglia considerata. L'assunzione su cui I2 poggia, cioè che l'assenza di allerta sia informativa rispetto all'assenza della condizione, risulta pertanto contraddetta dall'evidenza pertinente al contesto di I2.

Nota sul valore predittivo negativo (Analisi a cura degli autori): Il valore predittivo negativo (VPN) del 97,7% riportato nello studio è specifico per la finestra di 6 ore dall'allerta. In tale intervallo, la prevalenza di sepsi incidente è del 2,55% (3.730 casi su 145.885 accessi totali), rendendo la probabilità a priori di non incorrere nell'evento pari al 97,45%. L'assenza di allerta sposta tale probabilità di appena 0,25 punti percentuali. Un VPN elevato in condizioni di bassa prevalenza non costituisce evidenza che l'assenza di allerta sia informativa e non sostiene l'inferenza I2. La verifica di questo scarto tra probabilità a priori e VPN, calcolato dagli autori come misura del guadagno informativo, costituisce una condizione di lettura applicabile a ogni sistema di allerta operante su condizioni rare e viene aggiunta alla libreria di valutazione.

Lettura rispetto a I2

Il valore di sensibilità è la quantità pertinente, perché l'inferenza I2 poggia sull'assenza di allerta (un falso negativo). Nello studio di Ostermayer et al. (2024), la sepsi è stata definita applicando i criteri di consenso internazionale **Sepsis-3**: l'insorgenza della patologia (onset) è determinata da un incremento di almeno 2 punti nel punteggio *Sequential Organ Function Assessment* (SOFA) in concomitanza con la prescrizione di almeno un'emocoltura e di terapia antibiotica.

La performance diagnostica è stata valutata entro una finestra temporale di 6 ore rispetto all'onset della sepsi. In questo intervallo critico, a fronte della soglia di allerta standard raccomandata dal produttore (pari a 6), il modello ha registrato una sensibilità del 14,7%. Questo significa che l'85,3% dei pazienti effettivamente settici non ha generato alcun allarme al triage del Pronto Soccorso. L'adozione dei criteri Sepsis-3 come riferimento è rilevante per la lettura del dato: i casi non identificati corrispondono a quadri definiti da disfunzione d'organo secondo criteri clinici, e non a una classificazione amministrativa. Questo rende la sensibilità riportata pertinente all'esito che I2 presuppone.

Evidenza a favore

La ricerca di evidenza a favore del modello Epic Sepsis Model (ESM) è stata condotta interrogando i principali database scientifici (PubMed, Google Scholar, JAMIA) con i termini di ricerca "Epic Sepsis Model clinical utility", "ESM validation", "ESM v2 accuracy".

La letteratura a supporto proviene principalmente dai dati di sviluppo interno del fornitore e da alcuni studi di implementazione clinica che ne valutano l'integrazione organizzativa.

Lo studio di sviluppo originale di ESM v1 riporta valori di AUROC compresi tra 0,76 e 0,83. Più recentemente, uno studio multicentrico (PMCID: PMC12949446) condotto sulla versione aggiornata (ESM v2) su 227.091 accessi presso 4 grandi sistemi sanitari statunitensi ha mostrato un'area sotto la curva (AUROC) favorevole, compresa tra 0,82 e 0,92. Tuttavia, gli stessi autori mettono in evidenza limiti sostanziali che bilanciano tali risultati, tra cui bassi valori predittivi positivi (VPP compresi tra 0,13 e 0,26), un'elevata variabilità di performance tra le diverse istituzioni e un alto carico di allerta (alert burden) per gli operatori. Gli autori raccomandano esplicitamente che ogni istituzione conduca uno studio di validazione locale prima di adottare il sistema; tale raccomandazione coincide esattamente con il requisito metodologico di definire e verificare un perimetro di validità locale per lo specifico contesto operativo.

Tuttavia, questi risultati favorevoli presentano due importanti limiti di trasferibilità rispetto all'inferenza I2:

- **Setting di reparto vs triage:** Le performance elevate si riferiscono a pazienti ricoverati (inpatients), dove la pipeline del modello beneficia della completezza di dati storici e di esami di laboratorio già refertati. Al triage l'assenza di parte di tali variabili incide su C8, e le prestazioni riportate non sono trasferibili a I2 senza evidenza specifica nel contesto.

La ricerca non ha prodotto studi indipendenti in cui il modello ESM, applicato al triage di Pronto Soccorso, dimostri prestazioni diagnostiche sufficienti a sostenere in sicurezza l'omissione dei protocolli clinici (inferenza I2).

7.5.4 Passo 4. Perimetro di I2

Sulla base dell'evidenza documentale disponibile al 23 agosto 2026, l'inferenza I2 **non risulta sostenuta**.

L'analisi delle singole condizioni di validità per il modello appreso si esprime nei seguenti stati:

- **C1, validità interna. Non soddisfatta.** L'evidenza pertinente al contesto di I2 (Ostermayer et al., 2024) contraddice l'assunzione che l'assenza di allerta, alla soglia considerata, sia informativa rispetto all'assenza della condizione.
- **C6 (Rappresentatività), C7 (Stabilità temporale) e C8 (Parità ambiente): Non accertate.** Trattandosi di un modello proprietario con specifica chiusa, la coerenza della coorte di sviluppo, la stabilità rispetto alla deriva dei dati (drift) e la replicabilità esatta della pipeline in tempo reale non sono verificabili senza la documentazione tecnica del fornitore e l'ispezione dei dati locali.
- **C9 (Identità versione) e C10 (Corrispondenza soglia): Non accertabili.** La determinazione se la versione del modello in produzione e la soglia configurata localmente corrispondano a quelle validate in letteratura richiede audit operativi interni alla struttura specifica.

- **C2 (Calibrazione), C3 (Contenuto), C4 (Validità consequenziale) e C5 (Uso reale): Non accertabili** senza l'analisi dei dataset locali grezzi e l'osservazione diretta del comportamento degli operatori.

Nessuna delle dieci condizioni risulta soddisfatta per I2 nel contesto considerato: una risulta non soddisfatta sulla base di evidenza diretta contraria, le restanti nove non accertate o non accertabili nelle condizioni date.

Nota da riportare nell'enunciato. L'evidenza disponibile riguarda una versione (ESM v1.0) e una soglia (pari a 6) ben determinate. Qualora la struttura utilizzi una versione differente (es. ESM v2) o abbia configurato una soglia locale diversa, essa si colloca interamente fuori dal perimetro delle validazioni pubblicate, in una condizione in cui non esiste alcuna evidenza scientifica esterna (né a favore né contro). In tale scenario l'inferenza I2 non è sostenuta né contraddetta dall'evidenza pubblicata, e le condizioni C9 e C10 restano determinabili solo con informazioni in possesso della struttura.

7.5.5 Cosa questa applicazione mostra sul metodo

Tre osservazioni, da tenere brevi.

Il metodo produce un perimetro anche su un modello opaco. Nessun accesso al codice, alla coorte di sviluppo o ai dati. Il perimetro che ne risulta è più ristretto e con più condizioni indeterminate, ma è un perimetro, e le condizioni indeterminate sono elencate insieme a ciò che le renderebbe accertabili.

Questo corregge un'affermazione della sezione 7.3: l'accesso non è la condizione di applicabilità del metodo, è ciò che determina quante condizioni possono essere accertate.

La libreria si accresce. C10 non era stata formulata per NEWS2, perché NEWS2 ha soglie fisse nella specifica. È emersa da questo caso, ed è riutilizzabile per qualunque sistema di allerta con soglia configurabile. È il primo esempio documentato del meccanismo per cui la libreria accumula valore fra un caso e il successivo.

Le condizioni ausiliarie fanno il lavoro pesante. Nei due casi analizzati, ciò che restringe il perimetro non è la qualità dello strumento: è lo scarto fra le condizioni assunte dalle validazioni disponibili e quelle dell'uso concreto. Contesto, versione, soglia, popolazione, definizione dell'esito. Sono tutte informazioni che esistono ma che nessuno confronta sistematicamente, e nessuna di esse compare in una metrica aggregata di accuratezza.

7.5.6 Forma predicativa di C10 e valutazione sui casi

Le condizioni enunciate in prosa descrivono che cosa va accertato. Perché una condizione diventi un elemento di libreria, e non la nota di un singolo caso, deve essere espressa in una forma che ne consenta la valutazione ripetuta su oggetti diversi senza riformularla. Si riporta qui la forma predicativa di C10 e la sua valutazione sui due casi trattati.

Il predicato opera su tre parametri, riferiti a una struttura, a un sistema in produzione e a un corpo di evidenza esterna.

- **ver_prod:** la versione del sistema effettivamente in esecuzione presso la struttura.

- **soglia_conf:** il valore di soglia impostato nella configurazione locale del sistema.
- **ver_val, soglia_val:** la coppia formata dalla versione e dal valore di soglia cui l'evidenza esterna invocata si riferisce.

Enunciato. C10 è soddisfatta quando ver_prod coincide con ver_val e soglia_conf coincide con soglia_val. È non soddisfatta quando almeno una delle due coppie risulta difforme. È non accertata quando almeno uno dei parametri necessari a decidere non è noto.

Proprietà della valutazione. La valutazione è a tre valori e non a due, e la falsità è decidibile con informazione parziale mentre la verità non lo è. È sufficiente che un solo parametro noto risulti difforme perché il predicato sia falso, quali che siano i parametri ignoti; perché sia vero occorre invece che tutti e quattro siano noti e concordanti. La conseguenza pratica è che l'assenza di informazione non è mai neutra a favore dell'inferenza.

Valutazione sui casi

- **Caso 1, NEWS2.** Soddisfatta. La specifica fissa sia le regole di attribuzione del punteggio sia le soglie di escalation, per cui soglia_conf e soglia_val coincidono per costruzione e ver_prod coincide con ver_val in quanto la specifica non ammette varianti configurabili. È questa la ragione per cui C10 non era stata formulata nel primo caso: il predicato è vero in modo degenerare. La valutazione diventa non accertata qualora la struttura adottasse soglie di escalation locali difformi da quelle della specifica, circostanza che dipende dal protocollo della struttura e non è nota.
- **Caso 2, Epic Sepsis Model in una struttura non specificata.** Non accertata. Sono noti i parametri dell'evidenza: ver_val corrisponde a ESM v1.0 e soglia_val al valore 6 raccomandato dal produttore. Non sono noti ver_prod né soglia_conf, che dipendono dalla struttura. Il predicato non è decidibile, e le due informazioni mancanti sono in possesso della struttura.
- **Caso 3, Epic Sepsis Model v2 con adattamento locale.** Non soddisfatta. Se la versione in produzione è ESM v2 con parametri ricalibrati su dati locali, ver_prod non appartiene all'insieme delle versioni cui l'evidenza esterna si riferisce, e il predicato è falso indipendentemente dal valore di soglia_conf, che resta ignoto. È il caso in cui la falsità è decidibile con informazione parziale.

Riusabilità. Il predicato è stato formulato una volta e valutato tre volte, assumendo un valore diverso in ciascun caso. La sua formulazione non contiene alcun riferimento alla sepsi, al deterioramento fisiologico o al dipartimento di emergenza: i parametri su cui opera sono proprietà della configurazione e dell'evidenza, non del dominio clinico. È applicabile pertanto a qualunque sistema di supporto alla decisione che produca allerte per confronto con una soglia configurabile.

Limite. La valutazione qui riportata è stata condotta manualmente. Il predicato è espresso in una forma che ne consente la valutazione automatica una volta che i parametri siano disponibili come dato strutturato, ma tale esecuzione non è stata realizzata e non è oggetto di questa prova di concetto.

8. Riferimenti

1. **Royal College of Physicians.** *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS.* Updated report of a working party. London: RCP, 2017. ISBN: 978-1-85717-693-3.
2. **Kemp K, Alakare J, Harjola VP, Strandberg T, Tolonen J, Lehtonen L, Castrén M.** National Early Warning Score 2 (NEWS2) and 3-level triage scale as risk predictors in frail older adults in the emergency department. *BMC Emerg Med.* 2020 Oct 28;20(1):83. DOI: 10.1186/s12873-020-00379-y. PMID: 33115446; PMCID: PMC7594283.
3. **Candel BGJ, Nissen SK, Nickel CH, Raven W, Thijssen W, Gaakeer MI, Lassen AT, Brabrand M, Steyerberg EW, de Jonge E, de Groot B.** Development and External Validation of the International Early Warning Score for Improved Age- and Sex-Adjusted In-Hospital Mortality Prediction in the Emergency Department. *Crit Care Med.* 2023 Jul 1;51(7):881-891. DOI: 10.1097/CCM.0000000000005842. PMID: 36951452; PMCID: PMC10262984.
4. **Covino M, Cacciamani Fanelli PM, Bonadia N, et al.** Early Warning Scores in Emergency Department Patients Aged 80 Years or Older. *JAMA Network Open.* 2026;9(3):e261532. DOI: 10.1001/jamanetworkopen.2026.1532.
5. **Warmerdam M, Stolwijk F, Boogert A, Sharma M, Tetteroo L, Lucke J, Mooijaart S, Ansems A, Cuevas LE, Rijpsma D, de Groot B.** Initial disease severity and quality of care of emergency department sepsis patients who are older or younger than 70 years of age. *PLoS One.* 2017 Sep 20;12(9):e0185214. DOI: 10.1371/journal.pone.0185214. PMID: 28945774; PMCID: PMC5612649.
6. **Kennedy M, Helfand BKI, Gou RY, Gartaganis SL, Webb M, Moccia JM, Bruursema SN, Dokic B, McCulloch B, Ring H, Margolin JD, Zhang E, Anderson R, Babine RL, Hshieh T, Wong AH, Taylor RA, Davenport K, Teresi B, Fong TG, Inouye SK.** Delirium in Older Patients With COVID-19 Presenting to the Emergency Department. *JAMA Netw Open.* 2020 Nov 2;3(11):e2029540. DOI: 10.1001/jamanetworkopen.2020.29540. PMID: 33211114; PMCID: PMC7677760.
7. **Wong A, Otles E, Ehrenfeld JM, et al.** External Validation of a Widely Implemented Proprietary Sepsis Prediction Model. *JAMA Intern Med.* 2021;181(8):1065-1070. DOI: 10.1001/jamainternmed.2021.2626. PMID: 34152373; PMCID: PMC8218124.
8. **Ostermayer DG, Braunheim B, Mehta AM, Ward J, Andrabi S, Sirajuddin AMS.** *External validation of the Epic sepsis predictive model in 2 county emergency departments.* *JAMIA Open.* 2024 Nov 13;7(4):ooae133. DOI: 10.1093/jamiaopen/ooae133. PMID: 39545248; PMCID: PMC11560849.
9. **Bambach K, Buck K.** Atypical Presentations of Critical Illness in Older Adults. *Clerkship Directors in Emergency Medicine (CDEM) / Society for Academic Emergency Medicine (SAEM) Curriculum.* Updated August 16, 2021. Disponibile su: SAEM CDEM.

10. Geriatric Emergency Department Collaborative (GEDC). *Sepsis in the Older Adult in the Emergency Department*. GEDC Educational Resources, 2020. Disponibile su: GEDC Sepsis Resource.

11. Mhd Yunin ASM, Tan TL. *National Early Warning Score 2 (NEWS2) as a prognostic tool for adult patients in emergency department: A retrospective observational study*. *PLoS One*. 2025 Jun 16;20(6):e0326058. DOI: 10.1371/journal.pone.0326058. PMID: 40523004.

12. Emergency Department Study (Northern England). National Early Warning Score 2 (NEWS2) as a predictor of mortality in the Emergency Department. *Emergency Medicine Journal (Preprint/Post-submission)*. Disponibile su: Huddersfield Pure Repository.

13. Northern General Hospital ED Study (Sheffield). Accuracy of NEWS2 for predicting the need for time-critical treatment in the Emergency Department. *Emerg Med J*. 2024;42(6):396-402. DOI: 10.1136/emermed-2023-213795.

Appendice. Registro delle decisioni metodologiche

Questo registro documenta in modo trasparente e tracciabile le scelte discrezionali e le assunzioni metodologiche adottate nella costruzione del presente documento, al fine di garantire la totale riproducibilità dell'analisi da parte di terzi e strutturare una solida difesa metodologica in fase di audit o revisione.

- **Scelta dell'oggetto di analisi (NEWS2):**
 - *Decisione:* Selezione del National Early Warning Score 2 (NEWS2) del Royal College of Physicians come oggetto della prima Prova di Concetto (PoC).
 - *Razionale:* Trattandosi di un algoritmo a regole fisse, pubblico e interamente deterministico, NEWS2 permette di isolare la struttura formale del metodo LOGICA eliminando le incertezze probabilistiche e i bias tipici dei modelli di Machine Learning “scatola nera”. Funge da benchmark metodologico (white-box) ideale per testare il framework di valutazione in un ambiente controllato.
- **Definizione della popolazione e della soglia di età (Età ≥ 75 anni):**
 - *Decisione:* Focus sulla popolazione geriatrica definita da un'età anagrafica pari o superiore a 75 anni.
 - *Razionale:* È la soglia geriatrica standard ampiamente consolidata in letteratura clinica per l'identificazione del paziente ad alto rischio di fragilità e presentazioni atipiche di malattia critica. Consente inoltre la perfetta sovrapposizione e coerenza scientifica con i dati dello studio di validazione di riferimento (Kemp et al., 2020).
- **Criterio di definizione della fragilità (Clinical Frailty Scale ≥ 4):**
 - *Decisione:* Adozione della Clinical Frailty Scale (CFS) di Rockwood con soglia di inclusione ≥ 4.

- *Razionale:* La CFS è lo standard internazionale validato per la misura della fragilità in regime d'emergenza. La soglia ≥ 4 identifica i pazienti classificati come "vulnerabili" o lievemente fragili, intercettando la fase di transizione clinica in cui le risposte fisiologiche tipiche iniziano ad attenuarsi o a presentarsi in modo atipico.
- **Isolamento dell'Inferenza I1 (Differimento della rivalutazione al triage):**
 - *Decisione:* Focalizzazione esclusiva sulla decisione di differimento della rivalutazione clinica in Pronto Soccorso sulla base del punteggio NEWS2 sotto soglia.
 - *Razionale:* È l'inferenza in cui lo scarto fra prestazione aggregata e condizioni d'uso è più ampio, e in cui l'esito dell'analisi non era prevedibile prima di esaminare l'evidenza. Concentrarsi su I1 permette di mostrare che il perimetro non discende dall'accuratezza statistica generale, ma dalle condizioni della decisione presa a valle dell'output.
- **Inclusione ed esclusione delle fonti di letteratura:**
 - *Decisione:* Selezione mirata di tre grandi studi epidemiologici di validazione (Kemp et al. su 1.711 pazienti over 75 fragili; Candel et al. su 95.553 accessi; Covino et al. su 50.645 pazienti over 80) e del caso d'uso Epic Sepsis Model.
 - *Razionale:* Sono stati inclusi esclusivamente studi ad alta numerosità campionaria dotati di analisi stratificate per età e/o fragilità. È stata esclusa la letteratura generica non focalizzata sul Pronto Soccorso (es. reparti di degenza ordinaria) poiché non rappresentativa delle condizioni di carico e stress operativo assunte dal contesto di triage di I1.