

Algoritmo decisionale per l'interpretazione dell'overall survival immatura in onco-ematologia: sviluppo e applicazione

Guido Gini¹, Laura Gurgoglione, Riccardo Ceccarelli, Antonella Poloni

Clinica di Ematologia, Azienda Ospedaliero-Universitaria delle Marche, Università Politecnica delle Marche, Ancona - Italy

ABSTRACT

A decision algorithm for interpreting immature overall survival in onco-hematology: development and application

Introduction: In indolent haematological malignancies, interim overall survival (OS) analyses are frequently limited by few events and immature follow-up. Relying exclusively on hazard ratio (HR) and log-rank test may provide insufficient information for clinical and HTA decision-making. We propose a hierarchical decision algorithm integrating frequentist (HR, RMST, milestone survival) and multi-model Bayesian inference to classify the OS signal into four operational categories (A–D), enabling distinction between non-conclusive OS and absence of benefit.

Methods: The algorithm was developed using reconstructed pseudo-IPD from the AUGMENT trial (R^2 vs R-placebo, R/R FL/MZL), which has long-term OS analysis with known outcome. The framework integrates HR, RMST, milestone survival, and multi-model Bayesian inference (exponential, Weibull, Gompertz; neutral and sceptical priors) into a three-node hierarchical algorithm with four output classifications (A–D). It was applied to immature OS data from the inMIND trial (tafasitamab- R^2 vs R^2 , R/R FL).

Results: On AUGMENT, the algorithm classified immature OS as class B (HR 0.57; 95% CI 0.30–1.08; Δ RMST24 +0.334 months; $P(HR < 1)$ 93.2–94.5% with sceptical prior). The 5-year OS analysis confirmed a significant benefit (HR 0.59; 95% CI 0.37–0.95; $p=0.0285$), supporting classification consistency. Applied to inMIND (HR 0.59; 95% CI 0.31–1.13), the algorithm produced class B: OS not yet conclusive but robust benefit.

Conclusion: This proof-of-concept framework, developed on a case with known outcome (AUGMENT) and applied to a case with immature OS (inMIND), produced classifications consistent with the respective signals, supporting utility for structured interpretation of interim OS data in HTA contexts.

Keywords: Bayesian analysis, Decision algorithm, Follicular lymphoma, HTA, Immature overall survival, RMST

Introduzione

L'overall survival (OS) rappresenta uno degli endpoint di maggiore rilevanza nelle valutazioni cliniche, regolatorie e di health technology assessment (HTA) (1). Tuttavia, la sua interpretazione nelle analisi ad interim può risultare problematica, soprattutto nei contesti clinici caratterizzati da decorso indolente, basso numero di eventi e disponibilità di linee terapeutiche successive.

Nel linfoma follicolare recidivato/refrattario, tali criticità sono particolarmente evidenti. La storia naturale della

malattia e la possibilità di trattamenti multipli dopo progressione rendono spesso necessaria una prolungata osservazione prima che l'OS raggiunga un grado di maturità sufficiente per una valutazione conclusiva. In questo scenario, l'assenza di significatività statistica in un'analisi precoce non equivale necessariamente ad assenza di beneficio, ma può riflettere più semplicemente un'informazione ancora incompleta.

L'approccio convenzionale, basato principalmente su hazard ratio (HR), intervalli di confidenza e test log-rank, presenta limiti noti in questa fase. L'HR è una misura relativa dell'effetto, talvolta di difficile interpretazione clinica; inoltre, in presenza di pochi eventi o di una possibile non proporzionalità dei rischi, la sua capacità descrittiva può risultare ridotta (2,3).

Questo problema ha conseguenze dirette nei processi di HTA. Un'analisi retrospettiva delle valutazioni NICE ha mostrato che una quota rilevante dei farmaci oncologici viene sottoposta a decisioni di rimborsabilità in presenza di dati sull'OS ancora immaturi (4,5). Parallelamente, una revisione

Received: April 11, 2026

Accepted: June 24, 2026

Published online: July 6, 2026

This article includes supplementary material

Indirizzo per la corrispondenza:

Guido Gini

email: guido.gini@ospedaleiriuniti.marche.it



dei processi HTA in Inghilterra, Francia e Australia ha evidenziato che i metodi di estrapolazione proposti dai produttori per l'OS vengono accettati dalle agenzie solo in una minoranza dei casi, con l'imaturità del dato e la dipendenza dalle assunzioni parametriche tra le critiche più frequenti (6). Questi dati documentano un gap metodologico reale: le agenzie devono decidere sull'OS immatura, ma spesso gli strumenti disponibili non sono adeguati a gestire l'incertezza strutturale che ne deriva.

Per questo motivo è crescente l'interesse verso metriche complementari: il restricted mean survival time (RMST) per quantificare il beneficio assoluto, la milestone survival per descrivere la separazione tra i bracci e gli approcci bayesiani per esprimere i risultati in termini di probabilità di beneficio, superando la dicotomia tra significativo e non significativo.

Il presente lavoro propone, come proof of concept, un framework integrato, basato su HR, RMST, milestone survival e inferenza bayesiana multi-modello, formalizzato come algoritmo decisionale gerarchico a tre nodi, capace di classificare il segnale OS non solo in termini di significatività statistica ma anche di plausibilità e robustezza del beneficio. Il caso studio è il linfoma follicolare recidivato/refrattario; ulteriori affinamenti e validazioni su casistiche più ampie e setting diversificati sono necessari prima di una sua applicazione sistematica. La necessità di strumenti strutturati per interpretare l'OS ad interim e framework decisionali basati su evidenza surrogata è stata già evidenziata in letteratura (7,8).

Materiali e Metodi

Disegno dello studio metodologico

È stata condotta un'analisi metodologica applicativa con una struttura a due fasi. Nella prima fase, l'algoritmo è stato sviluppato e verificato su pseudo-IPD ricostruiti dello studio AUGMENT (9), uno studio randomizzato di fase III di lenalidomide più rituximab (R^2) versus rituximab più placebo in pazienti con linfoma follicolare o della zona marginale recidivato/refrattario. AUGMENT è stato selezionato come dataset di sviluppo perché dispone di un'analisi di OS a lungo termine con outcome noto (9), che ha permesso di verificare la coerenza della classificazione algoritmica con l'esito finale osservato. Nella seconda fase, l'algoritmo è stato applicato, a titolo illustrativo, ai dati immaturi dello studio inMIND (10) (tafasitamab-lenalidomide-rituximab versus lenalidomide-rituximab, linfoma follicolare R/R), nel quale al cutoff pubblicato l'OS risultava ancora immatura con 38 decessi totali e HR osservato pari a 0,59 (IC 95% 0,31-1,13).

Ricostruzione dei dati individuali

Per entrambi gli studi, in assenza di accesso ai dati individuali originali, sono stati ricostruiti pseudo-individual patient data (pseudo-IPD) a partire dalle curve di Kaplan-Meier pubblicate e dai numeri a rischio riportati. La digitalizzazione delle curve è stata effettuata mediante WebPlotDigitizer versione 4.5 (Rohatgi A; <https://automeris.io/WebPlotDigitizer>, accessed April 2026) (11). La ricostruzione è stata utilizzata esclusivamente a fini esplorativi e metodologici. Per limitare l'inferenza al tratto osservato, l'orizzonte analitico principale

è stato fissato a 24 mesi per entrambi i dataset. Le lievi differenze tra gli HR ricostruiti e quelli pubblicati sono attribuibili all'approssimazione introdotta dalla procedura di ricostruzione degli pseudo-IPD.

Struttura del framework analitico

Il framework proposto è stato articolato in due livelli distinti ma complementari: modellizzazione della sopravvivenza e inferenza statistica.

Nel primo livello, sono stati considerati tre modelli parametrici di sopravvivenza comunemente impiegati in ambito HTA ed economico-valutativo: exponential, Weibull e Gompertz. La scelta dei modelli candidati è stata guidata dai principi del NICE DSU Technical Support Document 14 di Latimer (12), secondo cui la selezione del modello deve basarsi su bontà di adattamento, coerenza con i dati osservati e plausibilità clinica della forma della curva.

Nel secondo livello, sui modelli selezionati sono state applicate sia un'analisi frequentista che un'inferenza bayesiana, con l'obiettivo di quantificare l'effetto di trattamento in termini sia relativi che assoluti. I risultati dei due livelli sono stati integrati in un algoritmo decisionale gerarchico (descritto nella sezione successiva), che formalizza operativamente il framework traducendolo in una logica classificatoria applicabile a qualsiasi dataset di OS con le caratteristiche di applicabilità descritte nel testo.

Analisi frequentista

L'effetto relativo del trattamento è stato stimato mediante modello di Cox, riportando HR e intervallo di confidenza al 95%. Come misura assoluta del beneficio, è stato inoltre stimato il restricted mean survival time (RMST) a 12, 18 e 24 mesi. Le milestone survival sono state calcolate agli stessi timepoint per descrivere la separazione assoluta tra i bracci nel tempo.

Analisi bayesiana

Per ciascuno dei tre modelli parametrici è stata applicata un'inferenza bayesiana, stimando la distribuzione posteriore dell'HR e della differenza di RMST a 24 mesi. Sono stati considerati due scenari di prior sul log-hazard ratio: una prior neutra, centrata su assenza di effetto e con dispersione più ampia, e una prior scettica, anch'essa centrata su assenza di effetto ma con dispersione più contenuta. Il modello è stato stimato con campionamento MCMC (4 catene: 2.000 iterazioni ciascuna; warm-up: 1.000 iterazioni) mediante un'implementazione equivalente in Python (PyMC3/NumPyro), in quanto R/rstan non era disponibile nell'ambiente computazionale utilizzato. La convergenza è stata verificata mediante ispezione delle trace plot, $R\text{-hat} < 1,01$ e numerosità campionaria effettiva ($ESS > 400$). Per ciascun modello e per ciascuno scenario di prior sono stati riportati:

- HR mediano a posteriori con intervallo di credibilità al 95% (95% CrI);
- probabilità di un qualsiasi beneficio in OS, espressa come $P(HR < 1)$;

- probabilità di una riduzione del rischio di morte almeno del 10% e del 25%, espresse come $P(HR < 0,9)$ e $P(HR < 0,75)$;
- differenza mediana di RMST a 24 mesi con 95% CrI (orizzonte primario per le stime bayesiane; i timepoint a 12 e 18 mesi sono riportati solo come stime frequentiste descrittive);
- probabilità di una differenza positiva di RMST a 24 mesi, espressa come $P(\Delta RMST24 > 0)$.

Algoritmo decisionale

Il framework è stato strutturato secondo una logica gerarchica a tre nodi condizionali, applicabile in sequenza a qualsiasi dataset OS con le caratteristiche di applicabilità descritte nel testo (Fig. 1).

Nodo 1 - Conclusività formale (HR frequentista): se l'IC 95% dell'HR esclude 1, l'OS è formalmente dimostrata e l'algoritmo si arresta con classificazione A. Se l'HR è sfavorevole o l'IC attraversa 1 in modo simmetrico senza tendenza direzionale, il segnale è non informativo (classificazione D) e si raccomanda di attendere il follow-up successivo. Se l'HR è favorevole ma l'IC attraversa 1, si procede al Nodo 2.

Nodo 2 - Rilevanza clinica contestuale (RMST e milestone survival): a differenza degli altri nodi, questo filtro non si basa su un criterio meccanico ma su un giudizio esperto pre-specificato, contestualizzato alla storia naturale del setting in esame. Il filtro è positivo se $\Delta RMST24 > 0$ e la differenza di milestone survival a 24 mesi è giudicata clinicamente rilevante; in caso contrario, la classificazione è C e si raccomanda di attendere la maturazione del dato. A titolo esemplificativo, nel setting del linfoma follicolare R/R adottato in questo

lavoro, caratterizzato da OS attesa lunga e popolazione politrattata, sono stati pre-specificati come soglie di rilevanza clinica un $\Delta RMST24 > 0$ e una differenza di milestone OS a 24 mesi ≥ 3 punti percentuali. Tali soglie sono contestuali e non trasferibili meccanicamente ad altri setting; qualsiasi applicazione dell'algoritmo richiede la pre-specificazione di criteri analoghi prima dell'analisi dei dati, così da contenere la variabilità decisionale introdotta dal giudizio esperto.

Nodo 3 - Robustezza probabilistica (inferenza bayesiana multi-modello): il segnale è considerato probabilisticamente solido se tutti e tre i seguenti criteri sono soddisfatti su tutti i modelli parametrici testati: (i) $P(HR < 1) > 80\%$ con prior scettica; (ii) $P(HR < 0,9) > 60\%$ con prior scettica; (iii) range di $P(HR < 1)$ tra modelli < 5 punti percentuali (coerenza cross-model). Le soglie adottate non sono standard universalmente validati, ma punti di ancoraggio operativi e pragmatici, pre-specificati per questo framework e coerenti con il range osservato nella letteratura sui trial bayesiani in oncologia (70-99%, mediana 95%) (14). Esse devono essere interpretate come soglie orientative, suscettibili di raffinamento in studi futuri di calibrazione su casistiche più ampie e setting diversificati. Se tutti e tre i criteri sono soddisfatti, la classificazione è B: OS non ancora conclusiva ma beneficio plausibile e robusto. Se uno o più criteri non sono soddisfatti, la classificazione è C: segnale incerto, ulteriore follow-up necessario.

Le quattro classificazioni possibili dell'algoritmo sono: A (OS dimostrata, evidenza diretta); B (OS non conclusiva, beneficio plausibile e robusto, supporto indiretto, attesa maturazione); C (segnale incerto, insufficiente per conclusioni, rivalutare a follow-up successivo); D (segnale non informativo, attendere maturazione prima di applicare il framework).

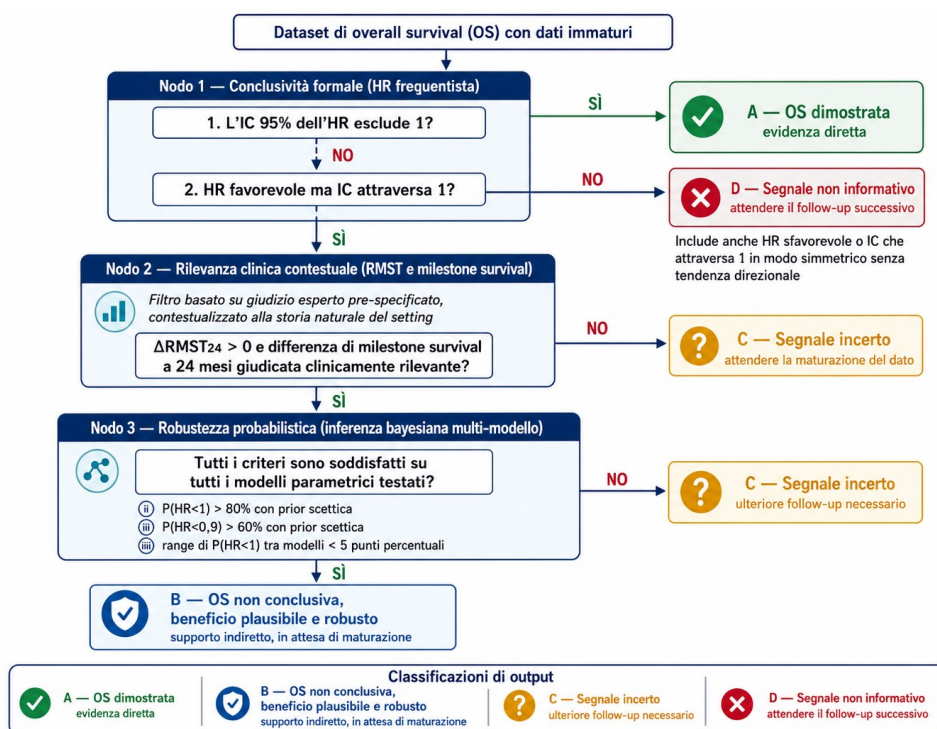


FIGURA 1 - Algoritmo decisionale per l'interpretazione dell'overall survival in presenza di dati immaturi.

Albero a tre nodi condizionali: Nodo 1 (HR frequentista - conclusività formale); Nodo 2 (RMST e milestone survival - rilevanza clinica contestuale, giudizio esperto pre-specificato); Nodo 3 (inferenza bayesiana multi-modello - robustezza probabilistica). Le quattro classificazioni di output sono: A (OS dimostrata); B (OS non conclusiva, beneficio plausibile e robusto); C (segnale incerto); D (segnale non informativo). Il caso studio analizzato (13) è classificato in classe B.

Risultati

Analisi frequentista

Su AUGMENT, l'analisi di OS sullo pseudo-IPD ricostruito ha fornito HR 0,57 (IC 95% 0,30-1,08), favorevole ma non conclusivo in termini inferenziali.

Restricted mean survival time e milestone survival

Le analisi assolute su AUGMENT hanno mostrato una separazione favorevole a R², crescente nel tempo. La differenza di RMST è risultata pari a +0,026 mesi a 12 mesi, a +0,104 mesi a 18 mesi e a +0,334 mesi a 24 mesi. Le milestone survival hanno mostrato una progressiva divergenza: sopravvivenza a 12 mesi 97,7% vs 96,6% ($\Delta + 1,0\%$), a 18 mesi 95,1% vs 93,0% ($\Delta + 2,1\%$), a 24 mesi 94,3% vs 89,5% ($\Delta + 4,8\%$). Gli intervalli bootstrap esplorativi per Δ RMST24 erano compresi tra circa -0,48 e +1,02 mesi.

Analisi bayesiana cross-model

L'analisi bayesiana su AUGMENT, applicata ai tre modelli parametrici (exponential, Weibull, Gompertz), ha prodotto risultati coerenti tra modelli, confermando che il segnale non dipende dalla specifica forma parametrica assunta.

Con prior neutra, l'HR mediano a posteriori era 0,593 (exponential), 0,607 (Weibull) e 0,580 (Gompertz), con P (HR < 1) tra il 94,3% e il 95,5%, P (HR < 0,9) tra il 90,3% e il 91,3% e P (HR < 0,75) tra il 76,4% e il 77,1%.

Con prior scettica, l'HR mediano era 0,649 (exponential), 0,670 (Weibull) e 0,673 (Gompertz), con P (HR < 1) tra il 93,2% e il 94,5%, P (HR < 0,9) tra l'86,7% e l'88,4% e P (HR < 0,75) tra il 64,3% e il 69,9%.

Differenza di RMST a 24 mesi

Le stime bayesiane di Δ RMST24 su AUGMENT sono risultate coerenti tra modelli. Con prior neutra: +0,555 mesi (exponential), +0,488 mesi (Weibull), +0,459 mesi (Gompertz); P (Δ RMST24 > 0) tra il 94,3% e il 95,5%. Con prior scettica: +0,450, +0,378 e +0,343 mesi; P (Δ RMST24 > 0) tra il 93,2% e il 94,5%.

Nel complesso, le analisi su AUGMENT illustrano come il framework possa supportare la distinzione tra OS non ancora conclusiva e assenza di beneficio in questo contesto specifico. Applicando l'algoritmo decisionale, la classificazione è B: Nodo 1 superato (HR 0,57, IC attraversa 1); Nodo 2 superato (Δ RMST24 +0,334 mesi, Δ S24 +4,8%, clinicamente interpretabile in questo setting); Nodo 3 superato [P (HR < 1) 93,2-94,5% con prior scettica, P (HR < 0,9) 86,7-88,4%, coerenza cross-model 1,3 punti percentuali]. La classificazione B, ottenuta con i dati immaturi, è risultata coerente con l'OS finale a 5 anni dello studio, che ha mostrato un beneficio statisticamente significativo (HR 0,59; IC 95% 0,37-0,95; p = 0,0285; OS a 5 anni 83,2% vs 77,3%) (9).

Applicazione illustrativa: studio inMIND

A titolo illustrativo, l'algoritmo è stato applicato ai dati immaturi dello studio inMIND (13), nel quale l'OS al cutoff pubblicato risultava ancora immatura (38 decessi totali; HR pubblicato 0,59, IC 95% 0,31-1,13). L'analisi su pseudo-IPD ricostruiti

ha fornito HR 0,57 (IC 95% 0,30-1,08): OS non conclusiva in senso formale, ma con segnale direzionalmente favorevole.

Al Nodo 1, l'HR è favorevole ma l'IC attraversa 1: si procede al Nodo 2. Le analisi assolute mostrano Δ RMST24 +0,76 mesi (frequentista) e una differenza milestone a 24 mesi di +7,3 punti percentuali (93,1% vs 85,7%): segnale positivo e clinicamente interpretabile in un setting con OS ancora acerba. Si procede al Nodo 3. L'inferenza bayesiana multi-modello con prior scettica mostra P (HR < 1) compresa tra l'85,9% e l'86,9%, P (HR < 0,9) tra il 73,9% e il 75,1% e coerenza cross-model di 1,3 punti percentuali: tutti e tre i criteri del Nodo 3 sono soddisfatti.

La classificazione algoritmica di inMIND è B: OS non ancora conclusiva ma beneficio plausibile e robusto. L'applicazione dell'algoritmo a un caso distinto da quello su cui è stato sviluppato, con struttura clinica simile (FL R/R, OS immatura, braccio di controllo attivo), produce la medesima classificazione osservata su AUGMENT con i dati immaturi, confermando la coerenza interna del framework in questo specifico setting; la generalizzabilità ad altri contesti richiede una validazione dedicata.

Analisi di sensibilità sulle soglie del Nodo 3

Per verificare la dipendenza della classificazione dalle soglie probabilistiche adottate, sono state testate tre combinazioni alternative dei criteri (i) e (ii) del Nodo 3: [P (HR < 1) > 80%, P (HR < 0,9) > 60%] (combinazione di riferimento); [P (HR < 1) > 85%, P (HR < 0,9) > 65%]; e [P (HR < 1) > 90%, P (HR < 0,9) > 70%], mantenendo invariato il criterio di coerenza cross-model (< 5 punti percentuali). Sia AUGMENT [P (HR < 1) 93,2-94,5%; P (HR < 0,9) 86,7-88,4% con prior scettica] sia inMIND [P (HR < 1) 85,9-86,9%; P (HR < 0,9) 73,9-75,1% con prior scettica] hanno mantenuto la classificazione B nelle prime due combinazioni. Con la soglia più conservativa (> 90%, > 70%), inMIND transitava a classe C, mentre AUGMENT manteneva la classe B. Questo comportamento documenta che la classificazione, anche se robusta a variazioni moderate delle soglie, è appropriatamente sensibile a criteri marcatamente più stringenti ed è coerente con la maggiore immaturità relativa del segnale inMIND rispetto ad AUGMENT. Una sintesi dei risultati ottenuti con i diversi approcci analitici su AUGMENT è riportata in Tabella 1. I dettagli completi sono riportati come materiale supplementare (Tab. S1).

Discussione

Il presente lavoro presenta una proof of concept di un algoritmo decisionale gerarchico, sviluppato su un caso con outcome noto (AUGMENT) e applicato a un caso con OS ancora immatura (inMIND), con l'obiettivo di supportare la distinzione tra OS non ancora conclusiva e assenza di beneficio, una distinzione con implicazioni dirette nelle valutazioni cliniche, regolatorie e di health technology assessment. Il framework non è inteso come strumento validato, ma come proposta metodologica che richiede affinamento e verifica su casistiche più ampie.

L'elemento più rilevante del presente lavoro è rappresentato dalla convergenza interpretativa tra misure relative, misure assolute di beneficio e modelli parametrici differenti.

TABELLA 1 - Sintesi dei risultati ottenuti con i diversi approcci analitici (AUGMENT) Le righe della Tabella corrispondono ai tre nodi dell’algoritmo decisionale (Fig. 1): Nodo 1 (effetto relativo), Nodo 2 (rilevanza clinica contestuale: RMST e milestone survival), Nodo 3 (probabilità di beneficio e robustezza modellistica). I dati completi per tutti i timepoint e tutti i modelli parametrici sono disponibili come materiale supplementare. Dati da AUGMENT (dataset di sviluppo). P (HR < 1): range cross-model con prior neutra 94,3-95,5% e prior scettica 93,2-94,5%. RMST24: stima frequentista +0,334 mesi; stime bayesiane +0,459-0,555 mesi (prior neutra) e +0,343-0,450 mesi (prior scettica). OS finale a 5 anni di AUGMENT: HR 0,59 (IC 95% 0,37-0,95; p=0,0285); OS 83,2% vs 77,3% (9).

Dominio	Metrica	Risultato chiave	Interpretazione
Nodo 1 - Effetto relativo	HR Cox	0,57 (IC 95% 0,30-1,08)	Favorevole ma non conclusivo (AUGMENT)
Nodo 2 - Beneficio assoluto	RMST24	+0,334 mesi (frequentista)	Beneficio assoluto positivo e crescente nel tempo
Nodo 2 - Separazione temporale	Milestone OS24	94,3% vs 89,5% (+4,8%)	Divergenza crescente fino a 24 mesi
Nodo 3 - Probabilità di beneficio	P (HR < 1)	93,2-95,5% (range cross-model)	Alta compatibilità con un beneficio in OS
Nodo 3 - Robustezza modellistica	3 modelli parametrici	Risultati coerenti tra exponential, Weibull, Gompertz	Il segnale non dipende dalla forma della curva

Su AUGMENT, metriche con assunzioni diverse, HR, RMST, milestone survival, probabilità bayesiane su tre modelli, producono tutte una lettura coerente. In particolare, P (HR < 1) è risultata tra il 94,3% e il 95,5% con prior neutra e tra il 93,2% e il 94,5% con prior scettica, con coerenza cross-model inferiore a 2 punti percentuali. La medesima coerenza si osserva applicando l’algoritmo a inMIND, suggerendo che il segnale risieda nei dati e non nelle assunzioni specifiche del modello.

L’utilizzo di una prior scettica rappresenta uno stress test deliberato del segnale (13). Su AUGMENT, anche assumendo a priori uno scetticismo ragionevole sull’entità del beneficio, la probabilità posteriore di un qualsiasi vantaggio in OS rimane superiore al 93% in tutti e tre i modelli. La probabilità di una riduzione del rischio di almeno il 10% [P (HR < 0,9)] supera l’86% anche nello scenario più prudente. L’OS finale a 5 anni dello studio ha confermato il segnale con significatività statistica, dimostrandone la coerenza con la classificazione algoritmica ottenuta con i dati immaturi. Questi risultati suggeriscono che il segnale non sia un artefatto dipendente da assunzioni ottimistiche, ma una caratteristica dei dati che resiste a forme ragionevoli di scetticismo informato.

Il RMST ha fornito indicazioni convergenti sia nell’analisi frequentista (+0,334 mesi a 24 mesi) sia nelle stime bayesiane (+0,459-0,555 mesi con prior neutra), con probabilità che ΔRMST24 sia positivo superiore al 90% con prior neutra e all’85% con prior scettica (2,3). Le milestone survival completano il quadro: la differenza assoluta a 24 mesi su AUGMENT (+4,8 punti percentuali, 94,3% vs 89,5%) e su inMIND (+7,3 punti percentuali, 93,1% vs 85,7%) è clinicamente interpretabile in entrambi i setting, confermando la coerenza del segnale tra i due dataset.

L’interesse applicativo del framework risiede nella sua potenziale capacità di supportare, e non sostituire, il giudizio valutativo in contesti HTA dove la piena maturazione dell’OS non è disponibile al momento della decisione (4-6). In tali contesti, la distinzione qualitativa tra “OS non conclusiva” e

“assenza di beneficio” può informare la struttura del ragionamento valutativo, senza costituire di per sé una raccomandazione applicativa. Il framework proposto, fondato su metriche assolute e probabilità posteriori, offre un linguaggio interpretativo complementare rispetto alla sola dicotomia del test log-rank, coerente con i riferimenti metodologici del NICE DSU TSD 14 (14) e con le indicazioni EMA (1). La proposta di algoritmo decisionale a tre nodi si inserisce in un filone metodologico che raccomanda esplicitamente di sviluppare strumenti strutturati per l’interpretazione dell’OS ad interim (7) e trova supporto empirico nell’osservazione che i trial oncologici già interpretati come positivi tendono a mostrare probabilità bayesiane di beneficio marginale superiori al 90%, mentre quelli negativi si collocano frequentemente al di sotto (15). Va sottolineato esplicitamente che l’algoritmo non costituisce uno strumento decisionale e non sostituisce i modelli di costo-efficacia né i processi deliberativi delle agenzie HTA. Esso fornisce un supporto interpretativo strutturato, finalizzato a informare, e non a rimpiazzare, il giudizio valutativo. I dati attualmente disponibili non supportano una raccomandazione all’applicazione operativa dell’algoritmo, che deve essere considerato una proposta metodologica suscettibile di affinamento e validazione formale.

Il contributo più originale del presente lavoro è la formalizzazione in logica condizionale di un framework classificatorio operativo. Framework descrittivi che combinano metriche diverse sono già stati proposti (8); ciò che ancora appare limitato in letteratura è la loro traduzione in categorie direttamente utilizzabili nei processi decisionali. La granularità A-D non è disponibile con il solo test log-rank ed è precisamente ciò di cui i processi HTA hanno bisogno quando si trovano a valutare farmaci oncologici con OS ancora immatura (5,6).

La trasferibilità del framework a contesti diversi dal linfoma follicolare R/R richiede considerazioni specifiche. Nei setting di prima linea dei linfomi indolenti, la disponibilità di multiple linee di salvataggio successive attenua



marcatamente il segnale OS, diluendo i benefici di PFS attraverso l'effetto "rescue" delle terapie successive; in tali contesti un vantaggio di PFS può non tradursi in beneficio OS e il Nodo 2 richiederebbe una ricalibrazione delle soglie di rilevanza clinica oltre a un orizzonte osservazionale più lungo. Nei setting recidivati/refrattari, al contrario, la più breve storia naturale residua può consentire al segnale OS di emergere più precocemente e con un minore confondimento da parte delle linee successive. Nei linfomi aggressivi (per esempio, DLBCL) l'OS matura più rapidamente e le analisi ad interim possono avere maggiore potere conclusivo, mentre l'estensione ai tumori solidi richiederebbe verosimilmente un adattamento sia delle soglie del Nodo 2 sia della struttura dei prior del Nodo 3. L'applicabilità dell'algoritmo è pertanto, allo stato attuale, empiricamente supportata solo per il setting R/R del linfoma follicolare.

Il framework si distingue concettualmente dagli approcci basati su endpoint surrogati. La validazione di surrogati per l'OS, incluse le relazioni di surrogacy PFS-OS e gli endpoint compositi recentemente proposti nel linfoma follicolare, come la complete response a 30 mesi (CR30) (16), sempre più adottata come endpoint primario nei trial, affronta un problema distinto: predire l'OS a partire da un endpoint diverso. L'algoritmo qui proposto non sostituisce l'OS con un surrogato, ma struttura l'interpretazione del segnale OS osservato, per quanto immaturo. I due approcci sono pertanto complementari più che alternativi: la surrogacy informa il disegno dello studio e la scelta dell'endpoint primario, mentre il framework interpreta il dato di OS effettivamente disponibile al momento della decisione.

Una verifica particolarmente informativa del valore discriminante del framework consisterebbe nella sua applicazione a studi con noto disaccoppiamento tra PFS e OS. Lo studio PRIMA, che ha dimostrato un ampio e duraturo beneficio di PFS nel mantenimento con rituximab nel linfoma follicolare di prima linea senza alcun beneficio di OS al follow-up a lungo termine (HR OS 1,04), rappresenterebbe un test di specificità ideale: ci si attenderebbe che un framework dotato di reale capacità discriminante classifichi tale segnale come non robusto, evitando un falso positivo nonostante l'evidente vantaggio di PFS; tale ipotesi richiede tuttavia una verifica empirica e non è stata testata nel presente lavoro. Studi di calibrazione futuri dovrebbero pertanto includere sistematicamente sia casi positivi sia casi negativi con outcome noto, tra cui, oltre a PRIMA, studi quali GALLIUM, FOLL05 e ROSEWOOD, quest'ultimo caratterizzato da discrepanze tra OS ad interim e finale, per stimarne sensibilità e specificità della classificazione.

Il presente lavoro ha limiti che devono essere esplicitamente riconosciuti. Il principale è l'utilizzo di pseudo-IPD ricostruiti: la lieve discrepanza osservata tra l'HR pubblicato (0,59) e quello stimato sullo pseudo-IPD (0,60) dimostra che tale approssimazione è quantitativamente trascurabile in questo caso, ma non è generalizzabile. La performance dell'algoritmo potrebbe inoltre differire quando applicato a dati individuali originali (IPD): la ricostruzione da curve di Kaplan-Meier introduce un errore non eliminabile, la cui entità può variare in funzione della qualità grafica delle curve pubblicate e della densità dei numeri a

rischio riportati. La trasferibilità delle classificazioni a contesti basati su IPD reali non può pertanto essere assunta senza una verifica dedicata. In secondo luogo, l'analisi bayesiana è condotta su un orizzonte di 24 mesi: l'estensione a orizzonti più lunghi richiederebbe scelte di estrapolazione che esulano dall'obiettivo metodologico del presente lavoro. In terzo luogo, le soglie decisionali del Nodo 3 sono punti di ancoraggio operativi e pragmatici, non standard clinicamente validati: devono essere interpretate come tali e sono suscettibili di revisione sulla base di future calibrazioni su casistiche più ampie e setting diversificati. In quarto luogo, la verifica di consistenza su AUGMENT, anche se incoraggiante, non costituisce una validazione formale: AUGMENT non era dimensionato per rilevare un beneficio in OS e l'applicazione dell'algoritmo è retrospettiva su IPD ricostruiti. Infine, entrambi i casi studio, AUGMENT e inMIND, sono contesti con segnale favorevole; la trasferibilità dell'algoritmo a scenari con risultati neutri, negativi o con crossing delle curve richiede una verifica applicativa ulteriore. Le classificazioni dell'algoritmo e le relative implicazioni interpretative per il contesto HTA sono riassunte in Tabella 2.

Conclusioni

Nei setting caratterizzati da OS immatura, la sola analisi convenzionale risulta spesso insufficiente per una piena interpretazione del segnale disponibile. Il presente lavoro propone un algoritmo decisionale gerarchico a tre nodi che classifica il segnale OS in quattro categorie operative (A-D) come supporto interpretativo nei processi clinici, regolatori e valutativi.

Su AUGMENT, la classificazione B ottenuta con i dati immaturi è risultata coerente con la successiva maturazione dell'OS finale a 5 anni (HR 0,59; $p = 0,0285$), fornendo una prima verifica di consistenza interna del framework. Applicato a inMIND, l'algoritmo produce la medesima classificazione B in un contesto clinicamente analogo. Questi risultati supportano la fattibilità dell'approccio come proof of concept, ma non ne costituiscono una validazione formale. Studi futuri dovranno testare il framework su casistiche più ampie e setting diversificati, affinare le soglie decisionali del Nodo 3 mediante calibrazione empirica e valutarne la riproducibilità con dati IPD originali.

Acknowledgements

Digitisation software: Kaplan-Meier survival curves from the AUGMENT and inMIND trials were digitised using WebPlotDigitizer version 4.5 (Rohatgi A; [Online](#), accessed April 2026) (14) for the purpose of pseudo-individual patient data reconstruction. Responsibility for the accuracy and interpretation of the reconstructed data remains entirely with the Authors.

Disclosures

Conflict of interest: All Authors declare no financial or non-financial conflict of interest relevant to this manuscript.

Financial support: This research received no specific funding from any agency in the public, commercial, or not-for-profit sectors.

TABELLA 2 - Classificazione dell’algoritmo e implicazioni decisionali Le implicazioni decisionali sono da intendersi come indicazioni orientative basate sulla classificazione algoritmica, non come raccomandazioni vincolanti. La decisione finale spetta alle autorità regolatorie e alle agenzie HTA competenti, che integrano il giudizio analitico con ulteriori considerazioni cliniche, economiche e di contesto.

Classe	Significato	Interpretazione	Implicazione HTA plausibile
A	OS formalmente dimostrata	IC 95% dell’HR esclude 1; effetto statisticamente conclusivo	Supporto diretto al claim di sopravvivenza; evidenza primaria disponibile
B	OS non conclusiva ma beneficio plausibile e robusto	HR favorevole, RMST positivo, probabilità bayesiane > 80% anche con prior scettica, coerenza cross-model	Può informare il ragionamento valutativo a supporto di decisioni condizionate, in attesa di maturazione dei dati
C	OS non conclusiva, segnale insufficiente o incerto	HR favorevole ma RMST trascurabile o probabilità bayesiane < 80% o incoerenza cross-model	Non supporta claim robusti su OS; rivalutazione necessaria al follow-up successivo
D	Dato non informativo	HR sfavorevole o IC simmetrico senza tendenza direzionale; framework non applicabile	Attendere maturazione del dato; non usare OS nel value claim in questa fase

Authors’ contributions: GG: Conceptualization; Methodology; Formal analysis; Writing - original draft; Writing - review & editing; LG: Conceptualization; Methodology; Formal analysis; Writing - original draft; Writing - review & editing; RC: Conceptualization; Methodology; Formal analysis; Writing - original draft; Writing - review & editing. AP: Conceptualization; Methodology; Formal analysis; Writing - original draft; Writing - review & editing.

Data availability statement: The data presented in this study are not publicly available. Individual patient data were not available; pseudo-IPD were reconstructed from published Kaplan-Meier curves and at-risk tables from the AUGMENT and inMIND trials using WebPlotDigitizer version 4.5. Restrictions apply to the availability of the original trial data, which belong to the respective trial sponsors.

Bibliografia

1. European Medicines Agency. *Guideline on the evaluation of anticancer medicinal products in man*. EMA/CHMP/205/95 Rev.5. EMA; 2022.

2. Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(1):152. [CrossRef PubMed](#)

3. Pak K, Uno H, Kim DH, et al. Interpretability of Cancer Clinical Trial Results Using Restricted Mean Survival Time as an Alternative to the Hazard Ratio. *JAMA Oncol*. 2017;3(12):1692-1696. [CrossRef PubMed](#)

4. Kang J, Cairns JA, Latimer NR, Duffield S, Grieve R. An assessment of the maturity of cancer survival data used in economic models for the National Institute for Health and Care Excellence's single technology appraisals. *Value Health*. 2025;28(11):1705-1713. [CrossRef](#)

5. Tai T-A, Latimer NR, Benedict Á, et al. Prevalence of immature survival data for anti-cancer drugs at time of approval and its impact on decision making: a retrospective analysis of health technology assessments. *Value Health*. 2021;24(8):1167-1175.

6. Trouiller J-B, Quenéchdu A, Toumi M, et al. The acceptance of overall survival extrapolation methods in solid tumor treatments by HTA agencies in England, France, and Australia between 2017 and 2022. *Med Decis Making*. 2025. online ahead of print. [CrossRef](#)

7. Friends of Cancer Research. Enhancing study designs and interpretation of interim overall survival data in oncology trials. White paper. Washington, DC: Friends of Cancer Research; 2024. Available from: [Online](#)

8. Ciani O, Grigore B, Taylor RS. Development of a framework and decision tool for the evaluation of health technologies based on surrogate endpoint evidence. *Health Econ*. 2022;31(S1):44-72. [CrossRef](#)

9. Leonard JP, Trněný M, Zhang H, et al; AUGMENT Trial Investigators. Lenalidomide plus rituximab for relapsed/refractory indolent non-Hodgkin lymphoma: 5-year follow-up and subgroup analyses from the phase III AUGMENT trial. *J Clin Oncol*. 2026;44(16):1483-1489. [CrossRef PubMed](#)

10. Sehn LH, Hübner K, Luminari S, et al; inMIND Study team. Tafasitamab, lenalidomide, and rituximab in relapsed or refractory follicular lymphoma (inMIND): a global, phase 3, randomised controlled trial. *Lancet*. 2026;407(10524):133-146. [CrossRef PubMed](#)

11. Rohatgi A. WebPlotDigitizer. Version 4.5. Austin, Texas, USA; 2021. Online. Accessed April 2026. Available from: [Online](#) (Accessed April 2026).

12. Latimer NR. Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. NICE DSU Technical Support Document 14. 2011, updated 2013. Sheffield: School of Health and Related Research, University of Sheffield. Available from: [Online](#)

13. Smit AJT, Van der Maas NG, Van der Holt B, et al. The added value of Bayesian interim analysis in randomized phase II and phase III clinical trials in hemato-oncology. *Haematologica*. 2025. [CrossRef](#)

14. Barret L, Liaigre L, Hlavaty A, et al. Priors and decision thresholds in phase 2 and phase 3 randomized controlled trials evaluating drug efficacy using Bayesian methods: a systematic review. *J Clin Epidemiol*. 2025;179:111651; online ahead of print. [CrossRef PubMed](#)

15. Sherry AD, Msaouel P, Kupferman GS, et al. Evidenced-based prior for estimating the treatment effect of phase III randomized trials in oncology. *JCO Precis Oncol*. 2024;8(8):e2400363. [CrossRef PubMed](#)

16. Shi Q, Flowers CR, Hiddemann W, et al. Thirty-month complete response as a surrogate end point in first-line follicular lymphoma therapy: an individual patient-level analysis of multiple randomized trials. *J Clin Oncol*. 2017;35(5):552-560. [CrossRef PubMed](#)

