# Classificatori logistici per la valutazione del rischio sanitario

Gabriele Cevenini

*Dipartimento di Chirurgia e Bioingegneria - Università degli Studi di Siena*

## Modello di regressione logistica

Consideriamo un modello predittivo basato sulla regressione logistica (RL) per la stima della probabilità di rischio sanitario, per esempio relativa a pazienti critici ricoverati in clinica.

Dato il vettore dei dati dei pazienti $\boldsymbol{x}$, nello spazio delle variabili cliniche disponibili, il modello di RL fornisce una probabilità condizionata $P(M \mid \boldsymbol{x})$ della classe $M$ dei pazienti malati interpretabile come probabilità prognostica del rischio di malattia.

L'uso del modello a fini diagnostici implica la scelta di una soglia di probabilità, detta probabilità di decisione $P_d$, sopra/sotto la quale i pazienti sono classificati a rischio o non a rischio di malattia. In altre parole, i pazienti vengono assegnati alla classe di rischio quando $P(M \mid \boldsymbol{x}) > P_d$. La scelta di $P_d$ dipende dalle prestazioni del classificatore e dal costo clinico attribuibile ad una decisione sbagliata.

La RL è una variante della regressione statistica lineare. È usata per prevedere la probabilità di una variabile dipendente dicotomica (detta anche *outcome*), utilizzando tipicamente una combinazione di variabili indipendenti sia quantitative continue sia qualitative categoriche. La RL è stata anche estesa a condizioni in cui la variabile dipendente ha più di due categorie (regressione logistica multinomiale), ma ciò non è oggetto della presente trattazione.

Per quanto concerne l'*outcome* binario relativo ai dati dei pazienti raccolti retrospettivamente, è codificato con il valore 1 nella condizione di malattia accertata (probabilità dell'evento certo, cioè rischio massimo) e col valore 0 nel caso abbiano avuto un decorso clinico normale, cioè in assenza di patologie o con guarigione (probabilità dell'evento impossibile, cioè rischio nullo). La relazione tra la variabile dipendente e le variabili indipendenti è di tipo log-lineare. Specificatamente, le probabilità a posteriori possono essere espresse come:

$$P(M \mid \boldsymbol{x}) = \frac{e^{\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}\right)}}{1 + e^{\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}\right)}}$$

$$P(N \mid \boldsymbol{x}) = 1 - P(M \mid \boldsymbol{x}) = \frac{1}{1 + e^{\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}\right)}}$$

La costante $\alpha$ e il vettore dei coefficienti $\boldsymbol{\beta}$ sono i parametri del modello che devono essere stimati del campione di addestramento (*training data*) usando la tecnica classica detta della massima verosimiglianza. T indica l'operatore di trasposizione vettoriale.

Si noti come il numero di coefficienti nel vettore $\boldsymbol{\beta}$ debba essere pari al numero di variabili predittive $\boldsymbol{x}$ utilizzate nel modello.

Il progetto del modello è effettuato cercando di massimizzare le prestazioni predittive (minimizzazione dell'errore tra valori veri e predetti), controllando il potere di generalizzazione. Il potere predittivo è solitamente definito attraverso la discriminazione, la calibrazione e l'accuratezza. Di ciò si tratterà più dettagliatamente nel seguito.

In aggiunta alla capacità predittiva, altri importanti criteri che contribuiscono al successo clinico del modello prognostico in applicazioni pratiche sono la semplicità del modello, intesa come ridotto numero di variabili predittive, e la facile implementazione nell'ambiente clinico. Ciò comprende la personalizzazione del modello a particolari centri sanitari e/o condizioni locali, il facile aggiornamento alla disponibilità di nuovi dati, la facilità computazionale e la capacità di fornire informazioni cliniche supplementari.

### *Odds ratio*

Il rapporto tra la probabilità di malattia $P(M \mid \boldsymbol{x})$ stimata dal modello e quella di decorso clinico normale, $P(N \mid \boldsymbol{x})$, che può essere calcolata come $P(N \mid \boldsymbol{x}) = 1 - P(M \mid \boldsymbol{x})$, è noto come rischio incrociato (*odds ratio*, OR) e vale:

$$\begin{aligned} \mathrm{OR} &= \frac{P(M \mid \boldsymbol{x})}{P(N \mid \boldsymbol{x})} = \frac{P(M \mid \boldsymbol{x})}{1 - P(M \mid \boldsymbol{x})} = e^{\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}\right)} \\ &= e^{\alpha} e^{\beta_1 x_1} e^{\beta_2 x_2} \ldots e^{\beta_s x_s} \end{aligned}$$

$\beta_1, \beta_2, \ldots \beta_s$ sono i coefficienti delle $s$ variabili indipendenti, $x_1, x_2, \ldots x_s$, presenti nel modello.

Pertanto, sotto l'ipotesi di variabili linearmente incorrelate, $e^{\beta_i}$ rappresenta il contributo all'OR della i-esima variabile unitaria, in pratica se la variabile vale 1 e tutte le altre sono nulle, compresa la costante $\alpha$. $\mathrm{OR}_i$ rappresenta quindi una stima del rischio relativo associabile a ciascuna variabile indipendente.

Nel caso di variabili binarie $\mathrm{OR}_i$ definisce quante volte è più alto il rischio di contrarre la malattia in presenza del fattore di rischio definito dalla variabile $i$. Nel caso la variabile categorica abbia più di due categorie, può essere resa *dummy* prendendo una categoria di riferimento e l'interpretazione dell'OR diventa analoga, ma può essere distinta per ciascuna categoria rispetto al riferimento. Si noti che il fattore $i$ rappresenta effettivamente un aumento del rischio solo se OR risulta superiore a 1, altrimenti ha un effetto di riduzione del rischio (fattore protettivo). Naturalmente, trattandosi di una stima campionaria, il valore dell'OR deve essere valutato assieme a un intervallo di confidenza, tipicamente fissato al 95% di significatività statistica. In tal caso, quando tale intervallo di confidenza ricomprenderà il valore 1, è impossibile giudicare quel fattore come pericolo o protezione e la stima puntuale dell'OR perderà completamente di significato.

Se invece la variabile è quantitativa, $\mathrm{OR}_i$ definisce quante volte aumenta il rischio in presenza di un incremento unitario della variabile $i$. Può allora apparire conveniente applicare il modello di RL a variabili quantitative standardizzate, cioè riportate a media nulla e deviazione standard unitaria. In tal caso, l'incremento unitario si riferisce all'incremento di una deviazione standard della variabile originaria. Per una migliore comprensione dell'OR e una maggiore utilità pratica nell'interpretazione dei risultati, può risultare spesso conveniente riportare le variabili quantitative a variabili qualitative dicotomiche, introducendo un valore di taglio, *cut-off*, o categoriche, con più categorie corrispondenti a classi di valori, come per esempio 3 fasce d'età, quali bambini, adulti e anziani. In quest'ultimo esempio, bisogna fissare gli estremi delle classi come ad esempio 12 anni tra bambini e adulti e 65 anni tra adulti e anziani. In base alle finalità dello studio può risultare opportuno scegliere la categoria dei bambini come riferimento e sarà quindi possibile valutare due valori OR che indicheranno il maggior rischio di ammalarsi di adulti e anziani rispetto ai bambini.

### *Variabili confondenti*

In qualche studio può accadere che esistano variabili riconosciute essere influenti sull'aumento del rischio, ma del cui OR non si abbia interesse. Per tener conto della loro incidenza sul campione, separandola da quella delle altre variabili oggetto dello studio, occorre pertanto includere nel modello di RL queste variabili, dette confondenti. L'OR di queste ultime può essere o considerato, per valutarne la loro influenza, o, più frequentemente, eliminato dallo studio. In quest'ultimo caso, l'effetto importante sarà quello di avere degli OR associati alle variabili di interesse clinico che rispecchiano di più l'effettivo rischio da esse direttamente, non corrotto, alterato o mascherato dall'indiretta presenza di variabili confondenti correlate.

### *Caratteristiche del modello*

La RL è una valida procedura statistica che comprende un'ampia varietà di forme parametriche di distribuzioni statistiche dei dati, ma che può essere applicata empiricamente qualsiasi sia la

natura dei dati. Per la verità, la RL non fa alcuna ipotesi sulla distribuzione delle variabili indipendenti che possono dunque assumere qualsiasi forma. Perciò, esse non devono necessariamente né essere distribuite secondo la curva normale o gaussiana, né avere una struttura e un'organizzazione lineare, né avere ugual matrice di covarianza tra le due classi di malattia e di decorso clinico normale.

Malgrado la sua eccellente flessibilità ad approssimare le proprietà statistiche dei dati, il modello di RL presenta gli stessi inconvenienti dei classici modelli di regressione multivariata, cioè principalmente risente della necessaria assunzione che le variabili predittive siano tra loro linearmente indipendenti (assenza di multicollinearità), pena la perdita di significato dei coefficienti di regressione e quindi anche dell'OR. Vedremo in seguito come il problema possa essere minimizzato usando tecniche *stepwise* per la scelta del sottoinsieme ottimo di variabili statisticamente significative.

Quando le due popolazioni di pazienti malati e sani sono distribuite normalmente con ugual matrice di covarianza (dette omoscedastiche[1]), la funzione discriminante logistica è esattamente identica alla consueta funzione discriminante lineare, ma i parametri si stimano in modo differente nelle due tecniche.

L'analisi discriminante di tipo logistico è generalmente preferibile a quella lineare in campioni di piccola dimensione, le cui distribuzioni di provenienza non sono chiaramente verificabili essere normali e omoscedastiche.

Un modello di RL non è facilmente aggiornabile a nuovi dati disponibili di training, tuttavia un completo riaddestramento periodico non crea eccessivi problemi. La sua implementazione clinica necessita dell'uso di un computer che può fornire anche molte altre informazioni; per esempio, l'uso del modello in simulazione può permettere di riprodurre artificialmente molte differenti situazioni cliniche, utili a fini terapeutici.

[1]La parola omoscedastico è un'eccezione della lingua italiana in cui 'sce' si pronuncia 'sche'

## Generalizzazione

La generalizzazione è di cruciale importanza in un modello predittivo progettato su un campione costituito da un insieme di dati rappresentativi della popolazione oggetto dello studio (*training set*). Essa può essere definita come la capacità del modello di mantenere le stesse prestazioni predittive su dati appartenenti alla medesima popolazione, ma non usati per identificare i parametri del modello.

La generalizzazione è valutata attraverso le performance del modello su un insieme di dati noti (*testing set*), diverso dal *training set*. Il modello generalizza bene quando gli errori predittivi nel *training set* e nel *testing set* non differiscono significativamente.

Il progetto del modello va effettuato controllando efficacemente il processo di addestramento per aumentare il potere predittivo garantendo la generalizzazione. Teoricamente, l'ottimo è il modello più semplice possibile costruito sui dati di *training*, che mostra il più alto possibile potere predittivo su un qualsiasi insieme *testing* di dati nuovi, ugualmente rappresentativo della popolazione studiata. Modelli troppo complessi, con un numero eccessivo di parametri e/o variabili, tendono a dare una sovrastima della loro performance predittiva (*overfitting*), cioè tendono a mostrare un errore sui dati di training significativamente più basso che sui dati di *testing*. L'*overfitting* è una sorta di memorizzazione dei dati che preclude l'apprendimento robusto delle sole regole di classificazione. Esso deve pertanto essere evitato perché ovviamente causa una perdita di generalizzazione.

### *Cross-validation*

Esistono molti metodi per controllare l'*overfitting*, migliorare e verificare la generalizzazione. Qui usiamo tecniche di validazione incrociata (*cross-validation*), ponendo particolare attenzione a quella nota col nome *leave-one-out* (LOO), letteralmente 'lasciare fuori un caso'. Le tecniche di *cross-validation* permettono di verificare che i risultati del modello statistico possano essere generalizzati a un insieme di dati indipendente, ma appartenente alla stessa popolazione di quelli usati per costruire il modello. Talora si eseguono diverse sessioni di *training* corrispondenti a diverse partizioni dei dati per poi mediare i risultati su tutte le sessioni.

Nella più semplice tecnica di *cross-validation* si esegue una sola sessione di addestramento usano solo due sottoinsiemi di *training* e *testing sets* con uguale numerosità. In una conveniente variante si usa un *training set* di maggiore numerosità, in modo da addestrare il modello su un numero maggiore di casi, senza tuttavia restare con un *testing set* troppo esiguo. Un buon compromesso è quello di usare il 70% di casi per il *training* e il 30% per il *testing*.

### *Validazione random*

Il metodo estrae più volte (sessioni) a sorte (random) insiemi di dati di *training* e *testing*. Per ognuna delle sessioni il modello è costruito sui dati di *training* e la sua capacità predittiva viene valutata sui dati di *testing*. I risultati sono poi mediati su tutte le sessioni.

Il vantaggio di questo metodo, rispetto al seguente, è che la proporzione di dati *training*/*testing* non è collegata al numero di possibili sessioni, che possono essere quante se ne desidera, lo svantaggio è che alcuni dati possono non essere mai utilizzati in fase di *testing* mentre altri possono essere utilizzati molte volte.

### *Rotazione*

La validazione a rotazione, detta anche *k-fold cross-validation*, è basata sulla suddivisione dell'originale campione di dati in vari sottocampioni complementari, solitamente di ugual numerosità. Uno solo di questi $k$ sottoinsiemi è usato come *testing set* e gli altri $k$-1 costituiscono il *training set*. Il processo è ripetuto $k$ volte (*folds*) a rotazione e i risultati delle $k$ sessioni di *testing* vengono mediati o combinati in modo da produrre una stima singola dell'accuratezza predittiva del modello. Nella pratica è consuetudine assumere $k$=10.

Il vantaggio del metodo di rotazione, rispetto a quello *random*, è che sono usati sempre tutti i dati e, per di più, una sola volta in fase di *testing*.

### *Leave-one-out*

Il metodo LOO è particolarmente utile in campo biomedico dove i dati disponibili sono solitamente pochi, perché permette di utilizzare efficientemente tutti i dati di *training* a disposizione sia per costruire il modello sia per testarne le sue capacità predittive, cioè valutarne la generalizzazione, anche se ciò va a scapito della semplicità e della velocità computazionali. In pratica, si tratta del metodo di rotazione dove le partizioni sono esattamente pari al numero $n$ dei dati disponibili e sono cioè costituite ciascuna da un solo caso.

In altre parole, avendo a disposizione $n$ dati retrospettivi, il LOO si basa sulla costruzione di tutte le $n$ possibili combinazioni di $n$-1 casi e sull'uso di queste combinazioni per $n$ sessioni di training. Gli $n$ casi lasciati fuori (uno per sessione) sono usati per calcolare l'errore. Rispetto ad altri metodi di *cross-validation*, tale procedura fornisce una valida stima dell'errore di predizione, con una distorsione piuttosto piccola. Siccome tuttavia, benché a turno, tutti i dati sono comunque usati molte volte in fase di *training*, la generalizzazione dovrebbe essere confermata anche su ulteriori dati, mai usati durante il processo di *training*.

### *Metodo stepwise di scelta delle variabili*

Un'altra ben nota fonte di perdita di generalizzazione è l'uso di troppe variabili predittive. Insiemi differenti di variabili possono descrivere ampie zone comuni e sottoinsiemi di esse potrebbero pertanto presentare lo stesso potere predittivo. In generale, più grande è il numero delle variabili, più alto è il numero di parametri del modello che devono essere stimati.

Per un assegnato insieme di dati *training*, l'identificazione dei parametri del modello peggiora tangibilmente all'aumentare del numero di variabili. La conseguenza immediata è una perdita significativa di generalizzazione. È perciò consigliabile scegliere il sottoinsieme minimo ottimo di variabili (rinominate spesso *features*).

Per la scelta ottima del tipo e del numero di *features*, è allora conveniente usare una tecnica numerica al computer a passi (*stepwise*), magari in combinazione con il metodo LOO suddetto.

Ad ogni passo del processo *stepwise*, una variabile viene immessa/rimossa nel/dal modello valutando se la riduzione dell'errore nella discriminazione è statisticamente significativo.

La procedura *stepwise* si arresta quando tutte le variabili fuori/dentro il modello non soddisfano più il criterio statistico di inclusione/rimozione.

Poiché l'introduzione/rimozione delle variabili nel modello ad ogni *step* avviene massimizzando il contributo statistico della variabile scelta, è evidente che tale variabile sarà quella che, tra tutte, contiene in sé il minimo/massimo potere discriminante in comune con quelle già presenti nel modello. Se ne evince come ciò permetta di limitare il problema tipico dei modelli di regressione della necessità di lineare indipendenza tra variabili, in quanto ogni variabile scelta a far parte del modello ha il minimo possibile di correlazione con le altre a fini discriminanti.

## Discriminazione

La disciplina che tratta della discriminazione è nota come *pattern recognition* e si riferisce a metodi per separare e classificare efficacemente dati entro distinte classi d'interesse. Un modello ha un'elevata capacità discriminante quando riesce a distinguere bene pazienti a rischio da pazienti a normale decorso clinico. Esistono molti criteri per valutare la performance discriminante di un modello predittivo, tra i più noti c'è quello basato sulla valutazione dell'area della curva ROC.

### *Area ROC*

Un modo usuale per valutare il potere discriminante di modelli diagnostici progettati per riconoscere solo un evento sfavorevole dalla normalità, è di calcolarne la sensibilità (SE) e la specificità (SP) che descrivono la frazione di pazienti classificati correttamente, rispettivamente a rischio e non a rischio. SE e SP dipendono della scelta di una soglia di probabilità (probabilità di decisione) $P_d$, con la quale viene confrontata la probabilità prevista dal modello.

Una valutazione complessiva, indipendente da $P_d$, delle capacità discriminanti del modello, è data dall'area sottesa dalla curva ROC (*Receiver Operating Characteristics*[2]), denominata AUC (*Area Under ROC Curve*).

La figura 1 mostra una tipica curva ROC. Essa da' una rappresentazione grafica della relazione tra la frazione di casi veri positivi (*True-Positive Fraction* TPF=SE), cioè la frazione di casi riconosciuti correttamente come a rischio dal modello, relativa a tutti i casi a rischio, e la frazione di falsi positivi (*False-Positive Fraction* FPF=1-SP), cioè la frazione di casi riconosciuti erroneamente come a rischio dal modello, relativa a tutti i casi non a rischio. Ciò per tutte le possibili scelte di $P_d$. AUC rappresenta la probabilità del modello di fornire probabilità di rischio più alte delle probabilità a decorso clinico normale, in pazienti che realmente vanno incontro ad un evento sanitario sfavorevole.

Un criterio usato frequentemente è quello di scegliere $P_d$ in corrispondenza al punto della curva ROC dove si hanno uguali valori di SE e SP (vedi croce in Fig. 1), così da ottenere lo stesso errore nel classificare pazienti normali e a rischio. Naturalmente possono essere scelte diverse coppie di SE e SP, a seconda del costo clinico attribuibile a una decisione sbagliata.

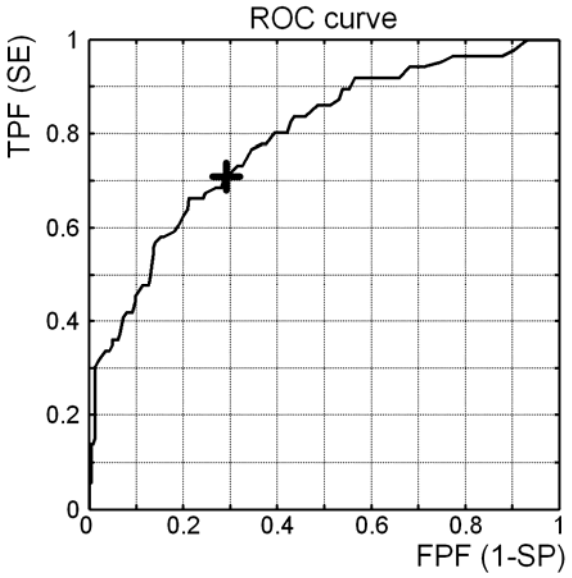


**Figura 1.** Curva ROC. La croce corrisponde al punto di ugual sensibilità (SE) e specificità (SP). Nell'esempio è all'incirca del 70%. L'area sottesa dalla curva ROC è un'ottimo indice quantitativo di bontà delle prestazioni del classificatore.

[2]Le curve ROC furono introdotte durante la II guerra mondiale per valutare la capacità dei tecnici militari addetti ai radar, nella corretta identificazione sul monitor di aerei nemici.

AUC può essere calcolato in modo semplice da dati campionari, usando il metodo empirico che implica l'integrazione numerica della curva ROC con la tecnica dei trapezi.

L'intervallo di confidenza di AUC si può calcolare con la tecnica computerizzata di ricampionamento nota col nome di *bootstrap*, usata per valutare l'incertezza di molti stimatori statistici complessi. La tecnica crea un certo numero $n_b$ di campioni casuali rimpiazzando dati presi da un campione originale di partenza. Un particolare dato dell'insieme originale può comparire più volte nei campioni ricostruiti con il *boostrap* mantenendo però la stessa numerosità campionaria.

Abbiamo visto come i modelli predittivi per la valutazione del rischio di un evento sanitario sfavorevole producono in uscita una probabilità di rischio che può essere utilizzata per discriminare tra due singole condizioni: evento sfavorevole (p.es. insorgenza di una malattia) ed evento propizio (stato di buona salute o guarigione). Ciò è possibile semplicemente definendo una probabilità c*ut-off* o soglia di decisione. L'approccio è lo stesso che si applica allo studio delle prestazioni di un test di laboratorio per scopi diagnostici che, come noto, fornisce un risultato negativo (nel nostro caso associabile alla predizione modellistica dello stato di salute) o positivo (associabile alla predizione di malattia del modello).

Si rimanda alla dispensa "Modelli predittivi di rischio: Sensibilità, Specificità e Area ROC" per un utile approfondimento sul tema, che comprende anche molti esempi tratti da ricerche sperimentali in campo biomedico.

## Calibrazione

L'accordo tra le probabilità vera e predetta dal modello è nota come calibrazione o bontà di adattamento (*goodness of fit*). Un modello ben calibrato fornisce valori di probabilità che possono realisticamente essere associati al rischio di sviluppare un evento sanitario sfavorevole.

Nelle condizioni di outcome dicotomico, le reali probabilità di rischio non sono intrinsecamente note. Infatti, dati retrospettivi forniscono solo outcome dicotomici, cioè pazienti con e senza evento sfavorevole. Nonostante tutto, può spesso essere utile stimare l'accadere di questi eventi usando una scala di valori continui, poiché i medici di solito preferiscono una stima probabilistica della prognosi meglio accettata di una semplice decisione duale.

Un test sulla bontà di adattamento, comunemente associato alla regressione logistica, è il test di Hosmer-Lemeshow (HL) e serve per verificare la calibrazione del modello di rischio progettato su dati retrospettivi con due outcome. Esso è basato sulla statistiche chi-quadrato la quale confronta frequenze osservate con frequenze attese di outcome, distribuendole in dieci gruppi di uguale numerosità.

Esistono due diverse formulazioni del test HL, a seconda del modo con cui si divide l'intervallo di probabilità. Quest'ultimo viene in ogni caso diviso in un certo numero di gruppi, di solito non più di 10. Se la numerosità campionaria è molto bassa (in pratica meno di 50 casi) si dovrebbe formare un numero inferiore di gruppi.

La prima formulazione del test HL, nota come statistica $\hat{H}$, raggruppa l'intervallo di probabilità in decili, conta i casi osservati per ogni decile e confronta le 10 frequenze ottenute con 10 frequenze attese tutte uguali (pari a un decimo del numero di casi totali).

La seconda formulazione (statistica $\hat{C}$), viceversa, ordina i casi osservati per probabilità di modello crescenti, li divide in 10 gruppi con la stessa frequenza e calcola le frequenze attese attraverso la media delle probabilità di ciascun gruppo. Anche se la statistica $\hat{C}$ del test è stata provata essere fortemente dipendente dalla dimensione del campione e dal criterio di raggruppamento è comunque di norma preferibile la seconda formulazione perché ha il vantaggio di evitare gruppi vuoti.

## Accuratezza

Come indice globale della performance predittiva del modello di rischio viene spesso assunto l'errore quadratico medio MSE (*Mean Squared Error*) tra le probabilità predetta e la probabilità osservata che naturalmente coincide con l'outcome binario. Il MSE tiene conto sia della discriminazione sia della calibrazione ed è noto anche come accuratezza. Esso vale:

$$\mathrm{MSE} = \sum_{i=1}^{n} \frac{\left[O_i - P(M \mid \boldsymbol{x}_i)\right]^2}{n}$$

$O$ è l'outcome uguale a 1 e a 0 per pazienti malati e sani, rispettivamente.

È stato dimostrato che il MSE può essere scomposto in tre addendi che tengono conto separatamente della calibrazione, della discriminazione e della numerosità campionaria.