8-23-2024

“Dead or Alive?” Assessment of the Binary End-of-Event Outcome Indicator for the NEMSIS Public Research Dataset

Mary E. Helander
meheland@syr.edu

Recommended Citation

Helander, Mary E. “‘Dead or Alive?’ Assessment of the Binary End-of-Event Outcome Indicator for the NEMSIS Public Research Dataset.” Prehospital Emergency Care, 2024, pp. 1–15, https://doi.org/10.1080/10903127.2024.2389551

"Dead or Alive?" Assessment of the Binary End-of-Event Outcome Indicator for the NEMSIS Public Research Dataset

Mary E. Helander, MS PHD NR-EMT

Syracuse University, Syracuse, New York, United States of America: Maxwell School of Citizenship and Public Affairs, Department of Social Science and Falk College, Department of Public Health

Author's Contact Information: meheland@syr.edu or maryehelander@gmail.com ,
Mailing address: 413 Maxwell Hall, Syracuse University, Syracuse NY USA 13244

https://www.maxwell.syr.edu/academics/social-science-phd-program/people/doctoral-students/emmy-helander

Author's Original Manuscript (AOM)
January 18, 2024

The **Version of Record** (VOR) of this manuscript has been published and is available Prehospital Emergency Care (Taylor & Francis), August 23, 2024, DOI: 10.1080/10903127.2024.2389551

Citation: Helander, Mary E. "'Dead or Alive?' Assessment of the Binary End-of-Event Outcome Indicator for the NEMSIS Public Research Dataset." *Prehospital Emergency Care*, 2024, pp. 1–15, https://doi.org/10.1080/10903127.2024.2389551.

“Dead or Alive?” Assessment of the Binary End-of-Event Outcome Indicator for the NEMSIS Public Research Dataset

OBJECTIVE: The broad absence of definitive patient outcomes in the NEMSIS public release data hinders research that seeks to understand the impact of pre-hospital care, operations, and overall patterns of population health – including geospatial and demographic differences. This study evaluated the recently proposed binary end-of-event outcome indicator to provide additional validity of the method, to evangelize its employment for more studies to analyze survival impact following an emergency medical event, and to identify appropriate use and interpretation given imperfection in predicted outcomes. **METHODS:** A recently published binary end-of-event outcome indicator was applied to datasets for each year from 2017 to 2022. Produced indicators were adjusted to address the method's inconsistencies. An array of established performance metrics from the binary classification in the machine learning literature were applied and interpreted. **RESULTS:** Over-fitting was detected for year 2018, as well as a degradation in performance when applying the method for datasets from year to year. Extended metrics revealed the method's weakness in accurately indicating the minority class: e.g., after adjustments for conflicting labels, “Dead” prediction accuracy was 77.7% for 2018 and 61.8% over the six-year NEMSIS sub-sample, verses 98.8% overall. **CONCLUSIONS:** After reproducing and then replicating a previously proposed method for predicting NEMSIS binary end-of-event outcomes, this study shows that it produces reasonably good “Dead” or “Alive” indicators. Reporting True Positive Rate (“Dead” prediction accuracy) and True Negative Rate (“Alive” prediction accuracy) is recommended whenever the method is used in NEMSIS analyses. For certain analyses, outcomes at the individual-level may be more appropriately quantified as probabilities using methods such as logistic regression, instead of predicted binary indicators. In the field, more attention to PCR completion of NEMSIS elements eOutcome.01 and eOutcome.02, whenever possible, can significantly enhance the public research datasets. [288 words]

Keywords: binary classification, health outcome prediction, NEMSIS, population health patterns,

INTRODUCTION

The National Emergency Medical Services Information System (NEMSIS) project provides voluminous public research data consisting of nearly a quarter billion events (1). To date, over one thousand scholarly articles have leveraged the dataset for retrospective studies which analyze EMS operations and pre-hospital care (2), but only a recent handful have considered survival/mortality outcomes (3,4,5,6). Unfortunately, just a tiny fraction (<1%) of NEMSIS patient records include a definitive end-of-event status indicating whether an individual survived a medical emergency, i.e. “lived” or “died” (3). Confounding this is a *rare event problem* (7): on average, fatalities occur in only 1.7% of patient care events (3). See Table 1.

(Insert Table 1 here)

By their nature, medical emergencies in need of 9-1-1 assistance are assumed to involve potentially life-threatening health situations (8,9). The broad absence of definitive patient outcomes hinders assessments that seek to understand the survival impact of EMS care, operations, and overall patterns of population health – including geospatial and demographic differences. Increased knowledge of patient outcomes would enhance NEMSIS's research utility so that more studies are able to quantify the impact of a pre-hospital intervention, situation, or population characteristic on likelihood of survival to hospital discharge – for example, manual verses mechanical chest compression (10); the effect of intubation (11) or bystander cardiopulmonary resuscitation (CPR) (12); race, ethnicity, age, sex, and geographic differences (5,13,14); advanced verses basic life support (15); mortality from traumatic penetrations including gunshot wounds (16); use of lights and sirens (17); and many others.

In a recent study, researchers developed and evaluated a method for predicting an EMS patient's end-of-event outcome – *presumptively dead* or *presumptively alive* –

using a combination of NEMSIS element and code combinations (3). It is referred to here as the *MLB prediction method* for its authors Miller, Lincoln, and Brown, or *MLB* for short. Their study was pivotal because it meant the NEMSIS public research dataset could include a predicted binary end-of-event outcome for each patient care record, increasing the dataset's utility for evaluating the most consequential aspect of EMS: mortality and survival patterns subsequent to pre-hospital care.

To date, despite a seemingly important contribution and the open availability of STATA code, only three published studies have leveraged the MLB method in their analysis. These include: a study characterizing patterns of pre-hospital care for cardiac arrests related to trauma (4); comparison of cardiac arrest outcomes in rural, suburban, and urban settings (5); and an assessment of dispatch protocols on survival rates for cardiac arrest (6).

The purpose of the present study is to evangelize MLB's significance to support future studies leveraging NEMSIS data, while providing a broader assessment of the MLB method's validity. This paper also provides recommendations for appropriate use, interpretation of outcomes predicted by the MLB method, and future study limitations given the inherent but quantifiable imperfection of predicted binary end-of-event outcome indicators for the NEMSIS dataset.

METHODS

Data Source and Instance Selection

This study analyzed the public research dataset for six consecutive years, from 2017 to 2022, from the NEMSIS project (1). NEMSIS is a national project funded by NHTSA's Office of EMS and hosted by the University of Utah. The project centralizes, standardizes, maintains, and publishes de-identified event records, originating from

patient care reports (PCRs) compiled by EMS responders from agencies across U.S. states and territories.

The sub-sample used in the present study, summarized in Table 1, was extracted following the MLB inclusion logic; see Table 2 and Figure 1. That is, only 9-1-1 dispatched EMS ground responses with patient contact were considered. Excluded, for example, were helicopter and airplane transports, cancelled dispatches, responses where no patient was found at the scene, etc. Further, the isolated instances were retained only if they included a definitive emergency room or hospital disposition – i.e., "lived" or "died" – following the same MLB logic.

(Insert Table 2 here)

(Insert Figure 1 here)

Variables Relevant to Predicted Outcomes

Variables used to specify binary outcomes for each patient event were derived from the same NEMSIS elements and code combinations used by the MLB method. The elements, codes, and descriptions used in MLB's novel two table prediction logic (3), for determining *presumptively dead* and *presumptively alive*, are recapitulated, and condensed here in the Table 3. For example, when EMS responders note a cardiac arrest patient with obvious signs of death, e.g. decapitation, dependent lividity, or rigor mortis (8), and do not attempt CPR for that reason, then the patient's care record coded with the value of **3016005** for NEMSIS element **eArrest.16** is evidence supporting the predicted status of *presumptively dead*, corresponding to row eleven in Table 3.

(Insert Table 3 here)

Data Preparation and Analysis Methods

The present study began by recognizing MLB as a custom-tailored binary classifier where the output is one of two classes, "Dead" or "Alive." Binary classification is a well-known task in data science and supervised machine learning with a broad array of metrics for evaluating performance (18,19,20).

Here, the approach first replicated and generalized the MLB logic (from STATA to Python) to isolate the NEMSIS sub-sample for years 2017 through 2022. The MLB logic for outcome prediction, i.e. *presumptively dead* or *presumptively alive*, was then applied to all six NEMSIS years. Confusion matrix counts, Cohen's Kappa Coefficient, and Overall Accuracy were computed for the 2018 dataset sub-sample. Counts, Cohen's Kappa, and Overall Accuracy were compared against the original MLB assessment (3) to verify that year 2018 results were correctly reproduced.

Next, MLB predicted outcomes were adjusted to remove "conflicts." That is, a PCR instance predicted as (simultaneously) both *presumptively dead* and *presumptively alive* was considered a "conflicted" instance. Prediction conflicts were possible from the MLB method's two table logic. The adjustment method dropped the "correctly predicted" indicator from each "conflicted" instance and kept the "incorrectly predicted" indicator, following the rationale that dual labelling was ambiguous and therefore equivalent to being incorrect.

The result of the adjustments was: each instance was unambiguously labelled either *presumptively dead* or *presumptively alive*, but not both. The resulting confusion matrix (20,21) comprised of True Positive (TP, or truly "Dead"), True Negative (TN, or truly "Alive"), False Negative (FN, or falsely labeled "Alive" when truly "Dead"), and False Positive (FP, or falsely labeled "Dead" when truly "Alive") included all instances

in the entire sub-sample and was comprised of counts from the four mutually exclusive sets.

Figure 2 provides a general depiction of the confusion matrix for "Dead" or "Alive" prediction. Note that this matrix is a special case of a *contingency table*: a well-known construct used in epidemiology and biostatistics for organizing data and for testing hypotheses that compare groups, interventions, or situations via risk and odds ratios and their confidence intervals (22).

(Insert Figure 2 here)

Finally, counts and an extended array of assessment metrics from the binary classification literature (18, 19, 20, 21, 23) were computed as a function of definitive and predicted outcome vectors. Figure 3 describes the assessment metrics selected for this study. The assessment results were organized by year and totals across the six-year NEMSIS sub-sample, and were assessed for general prediction quality, possible sore spots, over-fitting, and other general trends. Lastly, recommendations, mindful of prediction imperfections, were formulated for use of the MLB method in future retrospective studies involving the NEMSIS public release research datasets.

(Insert Figure 3 here)

RESULTS

NEMSIS Sub-Sample, 2017-2022

Table 4 summarizes the extracted NEMSIS sub-sample for years 2017 through 2022, showing a total of 686,075 instances -- PCRs for ground transport, patient contact, and definitive outcome by the MLB logic. Instances increased from year to year, but this also coincides with progression of NEMSIS v3 standard adoption by U.S. states and territories (1).

(Insert Table 4 here)

MLB Reproduction

Table 4, column two, shows the extracted NEMSIS sub-sample for 2018 that matched the developed MLB isolated data (with published correction)(3) for their equivalent to the confusion matrix, *before* adjustment for conflicts: 748 True Positives (TP, or truly "Dead"), 34,247 True Negatives (TN, or truly "Alive"), 143 False Negatives (FN, or falsely labeled "Alive" when truly "Dead"), and 152 False Positives (FP, or falsely labeled "Dead" when truly "Alive").

For the 2018 NEMSIS sub-sample year, Cohen's Kappa (COH =.831) and Overall Accuracy (ACC =99.2%) matched the original MLB's assessment metrics exactly. In summary, the present study correctly reproduced the MLB results for 2018.

Conflict Resolution Adjustment

In the 2018 sub-sample, the MLB method predicted both *presumptively dead* and *presumptively dead* for 173 instances which were in truth "Alive" at end-of-event. Similarly, MLB predicted both *presumptively dead* and *presumptively dead* for 72 instances that they were in truth "Dead." That is, MLB produced 215 conflicts for the 2018 NEMSIS sub-sample; see Table 1 rows for Conflicting Labels.

Adjustments resulted in increased counts for False Positive and False Negative categories accordingly. For example, for 2018, False Positives increased by 173 instances from 152 to 325 and False Negatives by 72 instances, i.e. from 143 to 215. See Tables 4 and 5, corresponding rows for Conflicting Labels, False Negatives, and False Positives.

Note that, for the 2018 NEMSIS sub-sample year, Cohen's Kappa and Overall Accuracy both decreased with this adjustment, which was expected since False Negatives and Positives are both increased while there was no change to True Positives and Negatives. That is, Cohen's Kappa decreased from .831 to .727 and Overall Accuracy from 99.2% to 98.5%. See Tables 4 and 5, rows for Cohen's Kappa and Overall Accuracy.

(Insert Table 5 here)

Extended Assessment of MLB

Results of the MLB prediction method to the six years of NEMSIS datasets, 2017 to 2022, are summarized in Table 4 and, after the adjustment correcting for conflicts, Table 5. Not surprisingly, all performance metrics worsened after the adjusting correction was applied. Still, metric values were consistent collectively at respectable levels signally reasonably good performance across years.

DISCUSSION

Interpretation of Cohen's Kappa and Accuracy for the Extended Sub-Sample

Cohen's Kappa Coefficient (COH)

Cohen's Kappa is a correlation-type of measure designed to help assess how closely one data set resembles another (24). When first proposed in 1960, the aim was to improve on approaches that were criticized for being too purely *percent agreement* (25). As such, it allowed for comparison that was statistical in nature, and thus more forgiving of random differences while considering non-discrete datasets.

Use of Cohen's Kappa to compare two deterministic binary vectors, as is the case in binary classification, was not its original intention. Even so, it has become somewhat widely used as a metric of comparison. Its value ranges from minus one to plus one, as per usual correlation metrics – closer to one (+, -) indicating similarity and zero indicating no similarity. Apparently, absolute values of Cohen's Kappa that are greater than .41 are considered acceptable for concluding reasonable similarity (24).

The value of Cohen's Kappa assessed on outcomes generated by MLB applied to all years was .638 (Table 5) and ranged from .597 for year 2022 to .727 for year 2018. Overall, this indicates reasonably good prediction performance. That year 2018 is much larger than other years, and given the development of MLB's prediction criteria, suggests there is over-fitting (26, 27) by the method. That is, the prediction rules may be overly customized for the year 2018. This is analogous to the problem in machine learning tasks whereby predictions are better for data used in model training than for other data (19).

Overall Accuracy (ACC)

Table 5 shows that MLB produces consistently high Overall Accuracy, in the range .983 to .993 across years and .988 for all years. The year-to-year pattern of these measurements is different than that of Cohen's Kappa, which peaked in 2018 and exhibited much more year to year variation. While Overall Accuracy may be considered a de facto measure of performance, it is known to be misleading in classification assessment when there is class imbalance (28).

For “Dead” or “Alive” prediction from the NEMSIS datasets, the class imbalance is evidenced in Table 1. For example, for the year 2018 NEMSIS sub-sample there are almost thirty-six times the number of patients who lived than died

following an EMS event – and almost sixty times overall. This concern is addressed next, in examining the individual class accuracy scores presented as extended assessment metrics: True Positive Rate (TPR), which measures the accuracy of predicting "Dead" instances, and True Negative Rate (TNR), which measures the accuracy of predicting the "Alive" instances.

Interpretation of Extended Assessment Metrics for MLB

True Positive Rate (TPR)

The True Positive Rate, also known as *Recall* or *Sensitivity*, is essentially accuracy measured only for the "Dead" (positive) category. Values are fractions between zero and one, with higher values reflecting better accuracy. Table 5 reveals that MLB's "Dead" accuracy ranges from a low of .521 in 2022 to a high of .777 in 2018, with an average of .618 across all years. Note that True Positive Rate is significantly lower than the Overall Accuracy, which is a commonly encountered plight in binary classification when there is a minority class, such as the case with "Dead" and "Alive."

That True Positive Rate in year 2018 is much larger than other years, and given the development of MLB's prediction criteria, adds to the suggestion that there is over-fitting by the MLB method.

True Negative Rate (TNR)

The True Negative Rate, also known as *Specificity* or *Selectivity*, measures the accuracy of the "Alive" (negative) category. Values are fractions between zero and one, with higher values reflecting better accuracy. Table 5 reveals that MLB's "Alive" accuracy ranges from a low of .989 in 2017 to a high of .996 in 2019 and 2022, with an average

of .995 across all years. Overall accuracy is essentially a convex combination of True Negative Rate and True Positive Rate.

When the "Alive" category is the majority class, by the principle of random incidence (29) it is easier to predict accurately. It's value and the plenitude of instances in this class pull the value of Overall Accuracy higher. If, instead of using the MLB prediction method, all instances were labelled as *presumptively alive*, then True Negative Rate would be equal to one (i.e. 100% accuracy for the "Alive" category), the True Positive Rate would be zero (i.e. 0% accuracy for the "Dead" category), and the Overall Accuracy would be equal to the survival rate in Table 1; for example, 98.3% over all years. This demonstrates how an alternative and trivial prediction method can be very accurate even when ignoring the minority category when it involves a rare event. Clearly, this would not be helpful if the intent of providing "Dead" and "Alive" predicted outcomes is to evaluate mortality and survival EMS patterns.

Balanced Accuracy (BACC)

The Balanced Accuracy is a direct average of True Positive Rate and True Negative Rate, which reflect "Dead" and "Alive" category accuracy respectively. Thus, Balanced Accuracy is an accuracy metric with equal weights between the two categories even though "Dead" has many fewer instances than "Alive." This results in Balanced Accuracy being less optimistic than Overall Accuracy in its accuracy assessment, as seen in Table 5 where its values range from a low of .759 in 2022 to a high of .884 in 2018.

Precision (PRE)

Precision measures the fraction of predicted *presumptively dead* that were indeed dead. Precision measurements take on values between zero and one, with higher values reflecting better quality. Table 5 reveals that MLB's Precision ranges from a low of .593 in 2017 to a high of .714 in 2022, with an average of .672 across all years. The rare event nature of the "dead" (positive) class together with the inaccuracy of many "alive" (i.e. the False Negatives) pushes these measurements to lower values, as is the case here.

F1 Score

As illustrated in Figure 3, F1 is a function of Precision and True Positive Rate: it is the harmonic mean of these two metrics. The F1 Score became popular with information systems, where it was used to measure retrieval performance (20). F1 Scores can range from zero to one, with better performance indicated by higher values. This metric is helpful when category imbalance is present such as the case of "Dead" and "Alive."

Table 5 reveals that MLB's F1 ranges from a low of .602 in 2022 to a high of .735 in 2018, with a value if .644 measured across all years. That year 2018 has a significantly higher F1 is further evidence to suggest over-fitting.

Matthew's Coefficient (MAT)

Matthew's Coefficient is a correlation-type of measure, with possible values ranging from minus one to plus one: a value of one indicates a perfect match, minus one a perfect inverse match, and zero indicates no correlation. Matthew's Coefficient is generally "liked" in binary classification because high (absolute value) scores are believed to align with confusion matrix values where "Trues" are maximized and "Falses" are minimized (30). There is some published empirical evidence suggesting

that Matthew's Coefficient has advantages over several other assessment metrics, such as F1, Balanced Accuracy and Cohen's Kappa (30,31,32).

Table 5 reveals that MLB's Matthew's Coefficient ranges from a low of .604 in 2022 to a high of.728 in 2018 in 2018, mirroring the pattern of Cohen's Kappa within early identical values. Comparison of the formulae for computing Matthew's Coefficient and Cohen's Kappa, shown in Figure 3, gives some insight to why these values are close: they differ by a constant multiplier in the numerator and a slight variance in combination of same terms in the denominator. However, Matthew's Coefficient and Cohen's Kappa are not always completely aligned, as illustrated in work that compared them based on simulated datasets (32). That Matthew's and Cohen's Kappa Coefficients are both closer to one than zero, as well as like each other in value and pattern, is stronger evidence that the MLB predictions are consistently well-correlated with the definitive outcomes.

Hamming Loss (HL)

Hamming's Loss measures the fraction of predictions that are incorrect. As a fraction, its values can range from zero to one with lower values indicating better performance. Hamming's Loss can be interpreted as a measure of overall inaccuracy. As such, it is equivalent in value to one minus Overall Accuracy, which can be observed by inspecting Table 5 rows for the two metrics. For example, Overall Accuracy over all years is 98.8% while Hamming's Loss is 1.2%, which is 1-Overall Accuracy.

Hamming's Loss has its origins in computer science and was used for bit checking to assess information loss in digital communications (33). It was included in the MLB extended assessment to illustrate the connection – in this case, equivalence –

between assessment metrics for binary classifiers, and to bring attention to their historical context.

*Jaccard Similarity (**J**)*

Jaccard Similarity is also known as the Jaccard Index, Jaccard Metric, or the Jaccard Coefficient. The metric measures the ratio of the intersection and union of the predicted and definitive label sets.

Like Hamming's Loss, the metric has a history with other domains -- in this case, ecology (used to compare plant species) and engineering (facility location) (34). Jaccard Similarity is re-emerging as a metric in the machine learning field, for example to guide a search algorithm to achieve a maximized similarity (35).

Possible values for Jaccard range from zero to one, with larger values indicating better performance. For the MLB, assessment with Jaccard revealed in Table 5 shows it ranging from .431 in year 2022 to .581 in year 2018. Compared to other assessment metrics, Jaccard gives a less optimistic view of the MLB's performance. However, the pattern across years is consistent with several other metrics: showing a common pattern of achieving a low and high in the same years as the Balanced Accuracy, True Positive Rate, F1, Matthew's Coefficient, and Cohen's Kappa.

Summary of Assessment Results

In general, the results from an assemblage of assessment metrics applied to MLB – the **accuracy rates and similarity measurements** of Overall Accuracy, True Positive Rate, True Negative Rate, Hamming's Loss, and Jaccard; the **combination metrics** of Balanced Accuracy, Precision, and F1; the **correlation coefficients** Cohen's Kappa and Matthew's – showed agreement on reasonably good prediction quality.

The extended assessment metrics and year to year trends revealed a subtle deficiency in the MLB performance in terms of minority accuracy – i.e. the accuracy of correctly predicting instances as “Dead” is only 61.8% across the six-year data set, in comparison to Overall Accuracy of 98.8% after (99.2% before) the conflict adjustment. Diminished accuracy for one category is of concern whenever the number of instances is far outnumbered by the other category. MLB's prediction does a much better at indicating truly alive than truly dead instances, i.e. 99.5% verses 61.8%, and therefore this is a valid concern. Veritably, “Dead” should be the more important class to predict accurately in any study based on NEMSIS data were outcomes and patterns of interest relate to potential fatalities.

While values and computation for individual metrics are different, Balanced Accuracy, True Positive Rate, F1, Matthew's, and Cohen's Kappa were all significantly higher in 2018 than for other years. As mentioned earlier, that original MLB development was anchored in 2018. However, that prediction performance is not as stellar for all other years suggests over-fitting by the method. The implication is that the measured accuracy reported for year 2018 (3) cannot be guaranteed for prediction use in other years. In machine learning, there are standard approaches to resolve over-fitting such as tuning of model hyper-parameters (e.g., learning rate or epochs (20), removing or penalizing (*regularization*) some variables (36), or changing model architectures (e.g., from a random forest to a logistic regression (19), or from a deep to a shallow neural network (37)). Resolving MLB's over-fitting likely means a deeper investigation of the two-tabled criteria with respect to years beyond 2018, with the objective of achieving equivalent assessments from year to year. This is left for future research.

The array of metrics considered in the extended assessment of MLB gives a broader perspective and adds to the credibility of the MLB prediction method by showing consistent and reasonably good performance. The advantage of considering a broader set of binary classification assessment metrics is that one or just a few can potentially show MLB to be overly optimistic or pessimistic, for example the Jaccard Similarity. There is also the possibility of missing a trend that indicates a deficiency such as minority class inaccuracy.

Other binary classification metrics from the literature, but not considered here, include: False Negative Rate (FNR), which is equivalent to one minus True Positive Rate; False Positive Rate (FPR), which is equivalent to one minus True Negative Rate; Average Precision Score (APS), which combines Precision and True Positive Rate; Receiver Operating Characteristic/Area Under the Curve (ROC/AUC) (38), which is equivalent to Balanced Accuracy in the discrete case; and Zero One Loss, which is equivalent to the Hamming Loss when normalized. Other less commonly used binary classification metrics, also not considered here, include Brier score, error rate, geometric mean, bookmaker score, informed-ness, and marked-ness (21).

Finally, it is noted that "Dead" is the usual minority class in the NEMSIS dataset because most people, fortunately, survive their medical emergencies in the United States. However, for out-of-hospital cardiac arrests (OHCA), "Alive" is the minority class. For example, in the NEMSIS sub-sample for 2018 there were 732 (67%) definitive instances of "Dead" and 358 (33%) definitive instances of "Alive" (3).

Recommendations for Applying MLB in NEMSIS Analyses

When using MLB to infer binary "Dead" or "Alive" outcome indicators for NEMSIS, the author recommends reporting the True Positive ("Dead") and True Negative

("Alive") rates from the appropriate NEMSIS sub-sample. Reporting these, for example in study limitations, provides quantified information about uncertainty – that is, the accuracy for prediction of *presumptively dead* (True Positive Rate) or *presumptively alive* (True Negative Rate) may be interpreted as the probability of a predicted binary outcome indicator being correct. For example, for out-of-hospital cardiac arrests (OHCA), for 2018 TP=694, FN=121, FP=38, and TN=237 (3) which results in True Positive Rate=.948 and True Negative Rate=.662. In other words, the probability of a predicted outcome "Dead" ("Alive") being accurate is 94.8% (66.2%). As a probability, it can also be used to estimate a confidence interval for the number in "Dead" or "Alive" categories, as the parameter of a binomial distribution (39,40).

LIMITATIONS

The analyses, results, and conclusions in this study rely on the accuracy and completeness of the NEMSIS public research dataset and, ultimately, the patient care reports completed by EMS providers. More frequent and unambiguous documentation of **eOutcome.01** and **eOutcome.02** elements – i.e., patient disposition from the emergency department and, if admitted, from the hospital, respectively – would significantly enhance the public research dataset by increasing the size of the sub-sample for ``learning" criteria and for assessing quality of the predicted indicators. Currently, however, there is a noteworthy practical challenge for this data improvement: lessened visibility by EMS providers to patient status *after* transfer-of-care to the emergency department.

CONCLUSION

The proposed MLB method for predicting NEMSIS binary end-of-event outcomes

produces reasonably good ``Dead'' or ``Alive'' binary outcome indicators, even after an adjustment for conflicts that recategorized them as incorrect predictions. After reproducing the MLB method, and then replicating it for several more NEMSIS dataset years, this study provided an extended assessment that adds to the validity of the MLB prediction method with an aim of inspiring more NEMSIS analysis to use it to analyze survival and mortality patterns related to EMS situations. These potential studies would add to the understanding of population health and of EMS' contribution to health and wellness in the United States.

It is recommended that researchers using MLB in their analysis should clearly state the True Positive Rate and True Negative Rate for the NEMSIS sub-sample with definitive outcomes that correspond to their study extract. This makes transparent that predicted outcomes are imperfect and provides the best-known quantifications describing the imperfections. In the field, EMS practitioner completion of PCR documentation related to NEMSIS eOutcome.01, .02, eArrest.01, .03, .12, .16, .17, .18, eDisposition.12, .19, .21, eScene.08, and eSituation.13 elements would significantly enhance the public research dataset.

Minority class accuracy and evidence of over-fitting of MLB suggest there is room for improving the prediction quality of the method. A next study examines alternatives to the MLB method by seeking probabilistic indicators of the end-of-event outcome instead of the deterministic binary indicator using Logit and Probit regression models.

FUNDING

The author was funded by Syracuse University, Maxwell School of Citizenship and Public Affairs.

HUMAN SUBJECTS REVIEW

This project was reviewed and approved by the Syracuse University Office of Research Integrity and Protection and determined to be exempt.

DECLARATION OF CONFLICTING INTEREST

The author declares that there are no conflicting interests.

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

The author declares that artificial intelligence (AI) was not used for manuscript writing.

DATA STATEMENT

This project analyses the public research dataset acquired from NEMSIS.org, used with permission and under all guidelines specified by the NEMSIS TAC data use agreement.

REFERENCES

(1) NEMSIS. What is NEMSIS?, 2023. URL https://nemsis.org/what-is-nemsis/. Accessed December 30, 2023.

(2) NEMSIS. Articles and publications using NEMSIS data, 2023. URL https://nemsis.org/using-ems-data/articles-and-publications/ Accessed December 30, 2023.

(3) Melissa L. Miller, Erin W. Lincoln, and Lawrence H. Brown. Development of a binary end-of-event outcome indicator for the NEMSIS public release research dataset. Prehospital Emergency Care, 25(4):504–511, 2021. doi: https://doi.org/10.1080/10903127.2020.1794435.

(4) Alexander J. Ordoobadi, Gregory A. Peters, Sean MacAllister, Geoffrey A. Anderson, Ashish R. Panchal, and Rebecca E. Cash. Prehospital care for traumatic cardiac arrest in the US: A cross-sectional analysis and call for a national guideline. Resuscitation, 179:97–104, 2022. doi: https://doi.org/10.1016/j.resuscitation.2022.08.005.

(5) Gregory A. Peters, Alexander J. Ordoobadi, Ashish R. Panchal, and Rebecca E. Cash. Differences in out-of-hospital cardiac arrest management and outcomes across urban, suburban, and rural settings. Prehospital Emergency Care, 27(2):162–169, 2023. doi: https://doi.org/10.1080/10903127.2021.2018076.

(6) Alexander Colgan, Morgan B. Swanson, Azeemuddin Ahmed, Kari Harland, and Nicholas M. Mohr. Documented use of emergency medical dispatch protocols is associated with improved survival in out of hospital cardiac arrest. Prehospital Emergency Care, pages 1–8, 2023. doi: https://doi.org/10.1080/10903127.2023.2239363.

(7) Ido Erev, Ira Glozman, and Ralph Hertwig. What impacts the impact of rare events. Journal of Risk and Uncertainty, 36(2):153–177, 2008. doi: https://doi.org/10.1007/s11166-008-9035-z.

(8) Andrew N. Pollack. Emergency Care and Transportation of the Sick and Injured. Jones and Bartlett Learning, Burlington, MA, 11th edition, 2017.

(9) David Schuster and Dan Nathan-Roberts. Situation awareness, sociotechnical systems, and automation in emergency medical services. In Joseph R. Keebler, Elizabeth Lazzara, and Paul Misasi, editors, Human Factors and Ergonomics of Prehospital Emergency Care. CRC Press, Taylor & Francis Group, Boca Raton, FL, 2017. doi: https://doi.org/10.1201/9781315280172.

(10) J.-A. Olsen, E. B. Lerner, D. Persse, F. Sterz, M. Lozano Jr, M. A. Brouwer, M. Westfall, P. M. van Grunsven, D. T. Travis, U. R. Herken, C. Brunborg, and L. Wik. Chest compression duration influences outcome between integrated load-distributing band and manual CPR during cardiac arrest. Acta Anaesthesiologica Scandinavica, 60(2):222–229, 2016. doi: https://doi.org/10.1111/aas.12605.

(11) Ryan Huebinger, Hei K. Chan, N. C. Mann, Benjamin Fisher, Benjamin Karfunkle, and Bentley Bobrow. Out-of-hospital intubation trends through the coronavirus disease 2019 pandemic. Annals of Emergency Medicine, 82(6):763–765, 2023. doi: https://doi.org/10.1016/j.annemergmed.2023.07.013.

(12) Jake Toy, Nichole Bosson, Shira Schlesinger, and Marianne Gausche-Hill. Racial and ethnic disparities in the provision of bystander CPR after witnessed out-of-hospital cardiac arrest in the united states. Resuscitation, 190:109901–109901, 2023. doi: https://doi.org/10.1016/j.resuscitation.2023.109901.

(13) Dalton C. Brunson, Kate A. Miller, Loretta W. Matheson, and Eli Carrillo. Race and ethnicity and prehospital use of opioid or ketamine analgesia in acute traumatic injury. JAMA Network Open, 6(10):e2338070–e2338070, 10 2023. doi: https://doi.org/10.1001/jamanetworkopen.2023.38070.

(14) Aditya C. Shekhar, Christopher Mercer, Robert Ball, and Ira Blumen. Persistent racial/ethnic disparities in out-of-hospital cardiac arrest. Annals of Emergency Medicine, 78:314 – 316, 2021. doi: https://doi.org/10.1016/j.annemergmed.2021.04.020.

(15) Prachi Sanghavi, Anupam B. Jena, Joseph P. Newhouse, , and Alan M. Zaslavsky. Outcomes of basic versus advanced life support for out-of-hospital medical emergencies. Annals of Internal Medicine, 163(9):681–690, 2015. doi: https://doi.org/10.7326/M15-0557.

(16) Ryan Huebinger, Hei K. Chn, Justin Reed, N. C. Mann, Benjamin Fisher, and Lesley Osborn. National trends in prehospital penetrating trauma in 2020 and 2021. The American Journal of Emergency Medicine, 72:183–187, 10 2023. doi: https://doi.org/10.1016/j.ajem.2023.07.022.

(17) Brooke L. Watanabe, Gregory S. Patterson, James M. Kempema, Orlando Magallanes, and Lawrence H. Brown. Is use of warning lights and sirens associated with increased risk of ambulance crashes? A contemporary analysis using National EMS Information System (NEMSIS) data. Annals of Emergency Medicine, 74(1):101 – 109, 2019. doi: https://doi.org/10.1016/j.annemergmed.2018.09.032.

(18) Charu C. Aggarwal. Data Classification: Algorithms and Applications, volume 35. CRC Press, Taylor & Francis Group, Boca Raton, 1st edition, 2015. doi: https://doi.org/10.1201/b17320.

(19) Trevor Hastie, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2nd edition, 2009. doi: https://doi.org/10.1007/978-0-387-84858-7.

(20) Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge, MA, 1st edition, 2012.

(21) Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. A review of evaluation metrics in machine learning algorithms. In Radek Silhavy and Petr Silhavy, editors,

Artificial Intelligence Application in Networks and Systems, pages 15–25, Cham, 2023. Springer International Publishing.

(22) Melody S. Goodman. Biostatistics for Clinical and Public Health Research. Routledge, Milton Park, Abingdon, Oxon; New York, NY, 2018. doi: https://doi.org/10.4324/9781315155661.

(23) scikit-learn Developers. scikit-learn User Guide, API Reference: Metrics, 2023. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics. Accessed on December 30, 2023.

(24) Mary L. McHugh. Interrater reliability: The Kappa statistic. Biochemia Medica, 22(3):276–282, 2012. doi: https://doi.org/10.11613/BM.2012.031.

(25) Jacob Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960. doi: https://doi.org/10.1177/001316446002000104.

(26) Douglas M. Hawkins. The problem of overfitting. Journal of Chemical Information and Computer Sciences, 44(1):1–12, 2004. doi: https://doi.org/10.1021/ci0342472.

(27) Ewout W. Steyerberg. Overfitting and optimism in prediction models. In Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, pages 95–112. Springer International Publishing, Cham, 2019. doi: https://doi.org/10.1007/978-3-030-16399-0_5.

(28) Lawrence S. D. Mosley. A Balanced Approach to the Multi-Class Imbalance Problem. PhD thesis, Iowa State University, 2013.

(29) Richard C. Larson and Amedeo R. Odoni. Urban Operations Research. Prentice-Hall, Englewood Cliffs, N.J, 1981.

(30) Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1):6–6, 2020. doi: https://doi.org/10.1186/s12864-019-6413-7.

(31) Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14(1):13–13, 2021. doi: https://doi.org/10.1186/s13040-021-00244-z.

(32) Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier Score in binary classification assessment. IEEE Access, 9:78368–78381, 2021. doi: https://doi.org/10.1109/ACCESS.2021.3084050.

(33) R. D. Heidenreich and R. W. Hamming. Numerical evaluation of electron image phase contrast. Bell System Technical Journal, 44(2):207–233, 1965. doi: https://doi.org/10.1002/j.1538-7305.1965.tb01658.x.

(34) G. A. Watson. An algorithm for the single facility location problem using the Jaccard metric. Society for Industrial and Applied Mathematics.SIAM Journal on Scientific and Statistical Computing, 4(4):748–9, 12 1983. doi: https://doi.org/10.1137/0904052.

(35) Zifu Wang, Xuefei Ning, and Matthew B. Blaschko. Jaccard metric losses: Optimizing the Jaccard index with soft labels. Technical report, Cornell University Library, arXiv.org, 2023.

(36) Xue Ying. An overview of overfitting and its solutions. In Journal of Physics: Conference Series, volume 1168, page 022022. IOP Publishing, 2019. doi: https://doi.org/10.1088/1742-6596/1168/2/022022.

(37) Charu C. Aggarwal. Neural Networks and Deep Learning: A Textbook. Springer International Publishing, Cham, 2nd edition, 2023. doi: https://doi.org/10.1007/978-3-031-29642-0.

(38) Tom Fawcett. An introduction to ROC analysis. Pattern Recognition Letters, 27(8):861–874, 2006. doi: https://doi.org/10.1016/j.patrec.2005.10.010.

(39) Victor Mooto Nawa. A weighted score confidence interval for a binomial proportion. Japanese Journal of Statistics and Data Science, 5:133–147, 2022. doi: https://doi.org/10.1007/s42081-022-00146-2.

(40) Per G. Andersson. The Wald confidence interval for a binomial p as an illuminating "bad" example. The American statistician, 77(4):443–448, 2023. doi: https://doi.org/10.1080/00031305.2023.2183257.

TABLES

Table 1. Summary of NEMSIS patient care reports (PCR events) from years 2017, 2018, 2019, 2020, 2021, and 2022 that involved ground transport and patient contact, and where a definitive end of event (``Died'' or ``Lived'') is indicated (3).

	NEMSIS DATASET YEAR						
	2017	**2018**	**2019**	**2020**	**2021**	**2022**	**ALL**
EMS Activations (1)	7,907,829	22,532,890	34,203,087	43,488,767	48,982,990	53,179,492	210,295,055
PCRs*	4,728,800	13,299,079	19,567,334	24,818,400	27,542,665	30,478,236	120,434,514
Definitive Outcomes**	36,877	35,535	87,497	118,617	137,367	280,172	696,065
Died	591	963	957	1,484	2,836	4,733	11,564
Lived	26,296	34,572	86,540	117,133	134,531	275,439	674,511
Definitive Outcome %	0.78%	0.27%	0.45%	0.48%	0.50%	0.92%	0.58%
Mortality Rate***	2.20%	2.70%	1.10%	1.30%	2.10%	1.70%	1.70%
Survival Rate***	97.80%	97.30%	98.90%	98.70%	97.90%	98.30%	98.30%
*From 9-1-1 calls, EMS ground transport responses with patient contact (3).							
**``Dead'' or ``Alive'' determined from NEMIS data elements eOutcome.01 and eOutcome.02 (3).							
***Definitive sub-sample of PCRs.							

Table 2. Summary of instance inclusion/exclusion logic in the data preparation step of the MLB prediction method (1,3).

NEMSIS Element		**Code Value**	**(Supported Indicator) Description**
eDisposition.12	≠	4212001	(Exclusion Logic) NPTT - assist, agency.
eDisposition.12	≠	4212005	(Exclusion Logic) NPTT - assist, unit.
eDisposition.12	≠	4212007	(Exclusion Logic) NPTT - canceled prior to arrival at scene.
eDisposition.12	≠	4212009	(Exclusion Logic) NPTT - canceled on scene with no patient contact.
eDisposition.12	≠	4212011	(Exclusion Logic) NPTT - canceled on scene with no patient found.
eDisposition.12	≠	4212039	(Exclusion Logic) NPTT - standby-no services or support provided.
eDisposition.12	≠	4212041	(Exclusion Logic) NPTT - standby-public safety, fire, or EMS operational support provided.
eDisposition.12	≠	4212043	(Exclusion Logic) NPTT - transport non-patient, such as organs, etc.
ePayment.50	≠	2650011	(Exclusion Logic) Transport by fixed wing (airplane).
ePayment.50	≠	2650015	(Exclusion Logic) Paramedic intercept.
ePayment.50	≠	2650017	(Exclusion Logic) Transport by rotary wing (helicopter).
eResponse.05	=	2205001	(Inclusion Logic) 911 response (scene); ground transport.
eResponse.07	=	2207003	(Inclusion Logic) Ground transport.
eOutcome.01	in	[01-09, 21, 43-70]	(Inclusion Logic; D-Alive) The known disposition from the ED, i.e. admitted, transferred, discharged or left AMA.
eOutcome.01	=	20	(Inclusion Logic; D-Dead) The known disposition from the ED, i.e. deceased.
eOutcome.02	in	[01-09, 21, 43-70]	(Inclusion Logic; D-Alive) The known disposition from the hospital, if admitted, i.e. transferred, discharged or left AMA.
eOutcome.02	=	20	(Inclusion Logic; D-Dead) The known disposition from the hospital, if admitted, i.e. deceased.
Table Abbreviations:			
AMA ≡ Against medical advice			
ED ≡ Emergency department			
NPTT ≡ Non-patient transport or transfer			
Exclusion Logic ≡ Used to exclude instances}			
Inclusion Logic ≡ Used to exclude instances			
D-Alive ≡ Definitive outcome, "Alive"			
D-Dead ≡ Definitive outcome, "Dead"			

Table 3. Summary of NEMSIS elements and codes used by the MLB prediction method (1,3).

NEMSIS Element		Code Value	(Supported Indicator) Description
eArrest.01	=	3001001	(presumptively alive) No indication of a cardiac arrest at any time during this EMS event.
eArrest.01	≠	3001003	(presumptively alive) No indication of a cardiac arrest at any time during this EMS event prior to EMS arrival.
eArrest.01	≠	3001005	(presumptively alive) No indication of a cardiac arrest at any time during this EMS event after EMS arrival.
eArrest.03	=	3003007	(presumptively dead) Indication of attempt to resuscitate the patient who is in cardiac arrest was: not attempted-considered futile.
eArrest.03	=	3003009	(presumptively dead) Indication of attempt to resuscitate the patient who is in cardiac arrest was: not attempted-DNR orders.
eArrest.03	≠	3003001	(presumptively alive) No Indication of attempt to resuscitate the patient who is in cardiac arrest using Defibrillation.
eArrest.03	≠	3003003	(presumptively alive) No Indication of attempt to resuscitate the patient who is in cardiac arrest using Ventilation.
eArrest.03	≠	3003005	(presumptively alive) No Indication of attempt to resuscitate the patient who is in cardiac arrest using Chest Compressions.
eArrest.12	=	3012001	(presumptively dead) Indication of whether or not there was any ROSC was: NO.
eArrest.16	=	3016001	(presumptively dead) Reason that CPR or the resuscitation efforts were discontinued was: DNR.
eArrest.16	=	3016005	(presumptively dead) Reason that CPR or the resuscitation efforts were discontinued was: obvious signs of death.
eArrest.16	=	3016011	(presumptively alive) Reason that CPR or the resuscitation efforts were discontinued was: ROSC, i.e. pulse or BP noted.
eArrest.17	=	9901001	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: Agonal/Idioventricular.
eArrest.17	=	9901003	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: Asystole.
eArrest.17	=	9901007	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Atrial Fibrillation.
eArrest.17	=	9901009	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Atrial Flutter.
eArrest.17	=	9901011	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: AV Block-1st Degree.
eArrest.17	=	9901021	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Left Bundle Branch Block.
eArrest.17	=	9901035	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: PEA.
eArrest.17	=	9901041	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Right Bundle Branch Block.
eArrest.17	=	9901043	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Sinus Arrhythmia.
eArrest.17	=	9901047	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Sinus Rhythm.
eArrest.17	=	9901049	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Sinus Tachycardia.
eArrest.17	=	9901059	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Supraventricular Tachycardia.
eArrest.17	=	9901065	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: unknown AED shockable rhythm.
eArrest.17	=	9901067	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: Ventricular Fibrillation.
eArrest.17	=	9901069	(presumptively alive) Patient's cardiac rhythm upon delivery/transfer to destination was: Ventricular Tachycardia(with pulse).
eArrest.17	=	9901071	(presumptively dead) Patient's cardiac rhythm upon delivery/transfer to destination was: Ventricular Tachycardia(pulseless).
eArrest.18	=	3018003	(presumptively dead) Patient's outcome at the end of the EMS event: expired in the field.
eArrest.18	=	3018007	(presumptively alive) Patient's outcome at the end of the EMS event: ROSC in the field.
eDisposition.12	=	4212013	(presumptively dead) Disposition treatment/transport indicates patient dead at scene-no resuscitation attempted(with transport).
eDisposition.12	=	4212015	(presumptively dead) Disposition treatment/transport indicates patient dead at scene-no resuscitation attempted(w/o transport).
eDisposition.12	=	4212019	(presumptively dead) Disposition treatment/transport indicates patient dead at scene-resuscitation attempted(w/o transport).
eDisposition.12	=	4212021	(presumptively alive) Disposition treatment/transport of patient: Patient Evaluated; No Treatment/Transport Required.
eDisposition.12	=	4212025	(presumptively alive) Disposition treatment/transport of patient: Patient Refused Evaluation/Care(w/o Transport).
eDisposition.12	=	4212027	(presumptively alive) Disposition treatment/transport of patient: Patient Treated - released AMA.
eDisposition.12	=	4212029	(presumptively alive) Disposition treatment/transport of patient: Patient Treated; Released(per protocol).
eDisposition.12	=	4212033	(presumptively alive) Disposition treatment/transport of patient: patient treated; transported.
eDisposition.12	=	4212035	(presumptively alive) Disposition treatment/transport of patient: Patient Treated; Transported by Law Enforcement.
eDisposition.12	=	4212037	(presumptively alive) Disposition treatment/transport of patient: Patient Treated; Transported by Private Vehicle.
eDisposition.19	=	4219003	(presumptively alive) Acuity of patient's condition after EMS care was: emergent(yellow).
eDisposition.19	=	4219005	(presumptively alive) Acuity of patient's condition after EMS care was: lower acuity(green).
eDisposition.19	=	4219007	(presumptively dead) Acuity of patient's condition after EMS care is dead w/o resuscitation efforts(black).
eDisposition.21	=	4221009	(presumptively dead) Type of destination the patient was delivered or transferred to is a morgue or mortuary.
eProcedure.03	≠	426220008	(presumptively alive) Procedure performed on the patient was NOT: External Ventricular Defibrillation.
eProcedure.03	≠	429283006	(presumptively alive) Procedure performed on the patient was NOT: Mechanically Assisted Chest Compression.
eProcedure.03	≠	450661000124102	(presumptively alive) Procedure performed on the patient was NOT: Defibrillation using AED.
eProcedure.03	≠	89666000	(presumptively alive) Procedure performed on the patient was NOT: CPR.
eScene.08	=	2708009	(presumptively dead) Triage classification for an MCI patient is black
eSituation.13	=	2813007	(presumptively dead) Acuity of patient's condition upon EMS arrival at the scene is dead w/o resuscitation efforts(black).
eTimes.11 - eVitals.01	in	[1,3]	(presumptively alive) Patient's vital was taken 1-3 minutes after the responding unit arrived with patient at the destination.
eVitals.06	in	[60,180]	(presumptively alive) Patient's systolic blood pressure is between 60 and 280 (i.e., viable).

Table Abbreviations:	ED ≡ Emergency department
AED ≡ Automated external (cardiac) defib	MCI ≡ Mass casualty incident
AMA ≡ Against medical advice}	NPTT ≡ Non-patient transport or transfer
AV ≡ Arteriovenous	PEA ≡ Pulse-less electrical activity
BP ≡ Blood pressure	presumptively alive ≡ MLB imputed outcome
CPR ≡ Cardiopulmonary resuscitation	presumptively dead ≡ MLB imputed outcome
DNR ≡ A do-not-resuscitate medical order	ROSC ≡ Return of spontaneous circulation

Table 4. Results of the MLB method applied to NEMSIS data-set years 2017, 2018, 2019, 2020, 2021, and 2022 before adjusting for conflicts.

Table 4. Results of the MLB method applied to NEMSIS data-set years 2017, 2018, 2019, 2020, 2021, and 2022 before adjusting for conflicts.

Year	2017	2018	2019	2020	2021	2022	ALL
Total Instances	26,887	35,535	87,497	118,617	137,367	280,172	686,075
True Positive (TP)	430	748	723	1,046	1,739	2,466	7,152
False Negative (FN)	134	143	155	306	732	1,716	3,186
True Negative (TN)	26,001	34,247	86,157	116,547	133,618	274,449	671,019
False Positive (FP)	101	152	224	359	592	473	1,901
Conflicting Labels (Truly Dead)	27	72	79	132	365	551	1,226
Conflicting Labels (Truly Alive)	194	173	159	227	321	517	1,591
Cohen's Kappa Coefficient (COH)	0.781	0.831	0.79	0.756	0.719	0.689	0.734
Overall Accuracy (ACC)	0.991	0.992	0.996	0.994	0.99	0.992	0.993

Table 5. Results of the MLB method applied to NEMSIS data-set years 2017, 2018, 2019, 2020, 2021, and 2022 after adjusting for conflicts and adding extended evaluation metrics.

Year	2017	2018	2019	2020	2021	2022	ALL
Total Instances	26,887	35,535	87,497	118,617	137,367	280,172	686,075
True Positive (TP)	430	748	723	1,046	1,739	2,466	7,152
False Negative (FN)	161	215	234	438	1,097	2,267	4,412
True Negative (TN)	26,001	34,247	86,157	116,547	133,618	274,449	671,019
False Positive (FP)	295	325	383	586	913	990	3,492
Conflicting Label (Truly ``Dead'')	0	0	0	0	0	0	0
Conflicting Label (Truly ``Alive'')	0	0	0	0	0	0	0
Cohen's Kappa Coefficient (COH)	0.645	0.727	0.697	0.667	0.626	0.597	0.638
Overall Accuracy (ACC)	0.983	0.985	0.993	0.991	0.985	0.988	0.988
True Positive Rate (TPR, Recall, Sensiti	0.728	0.777	0.755	0.705	0.613	0.521	0.618
True Negative Rate (TNR, Specificity, S	0.989	0.991	0.996	0.995	0.993	0.996	0.995
Balanced Accuracy (BACC)	0.858	0.884	0.876	0.85	0.803	0.759	0.807
Precision (PER)	0.593	0.697	0.654	0.641	0.656	0.714	0.672
F1 Score	0.653	0.735	0.701	0.671	0.634	0.602	0.644
Matthew's Coefficient (MCC)	0.648	0.728	0.699	0.668	0.627	0.604	0.639
Hamming Loss (HL)	0.017	0.015	0.007	0.009	0.015	0.012	0.012
Jaccard Similarity ($\cal{J}$)	0.485	0.581	0.54	0.505	0.464	0.431	0.475

FIGURES

Figure 1. Consort flow diagram for the analyzed NEMSIS sub-sample for years 2017 to 2022.

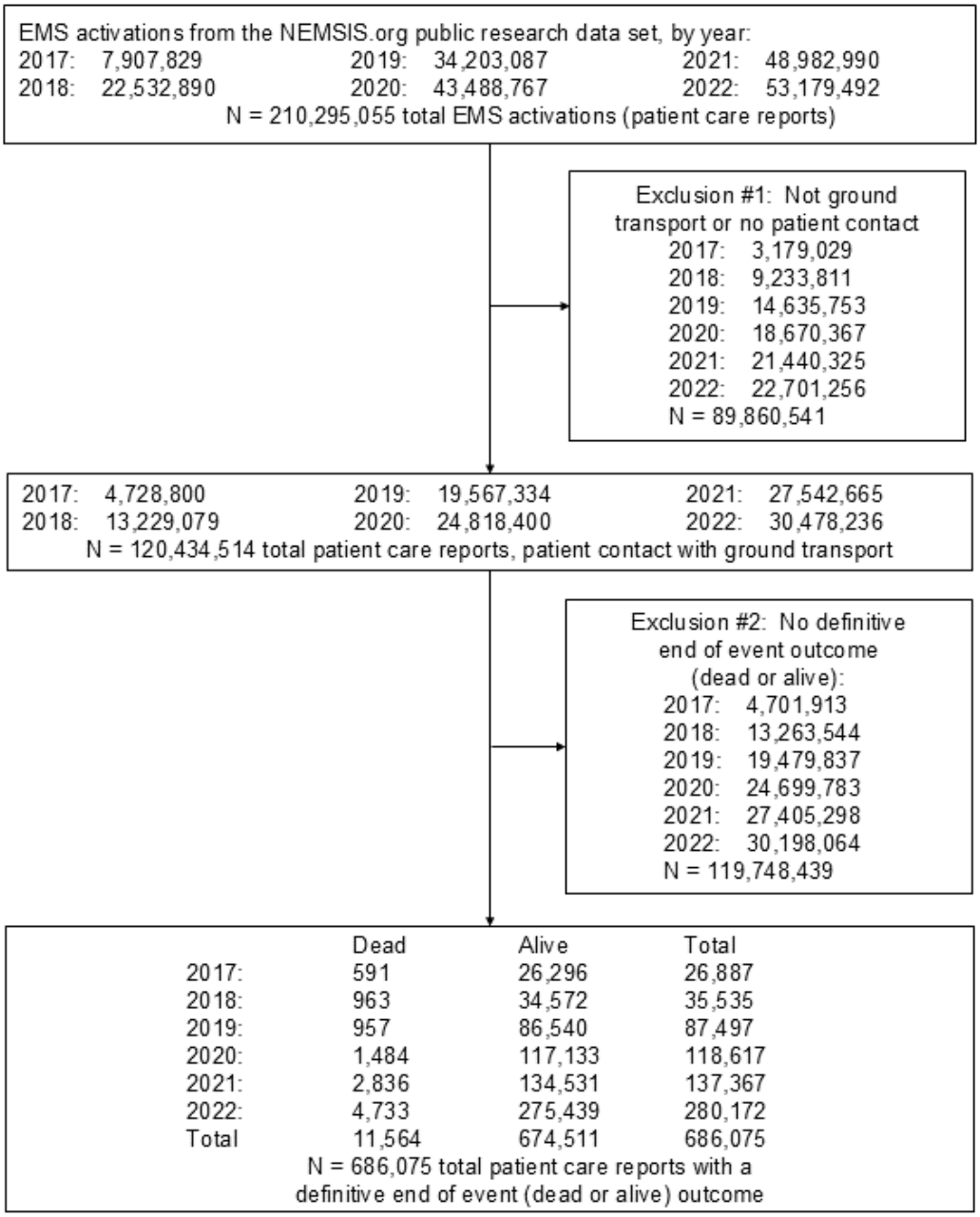

Figure 2. Generalization of the confusion matrix applied to ``Dead'' or ``Alive'' prediction; sklearn.metrics.confusion_matrix (23).

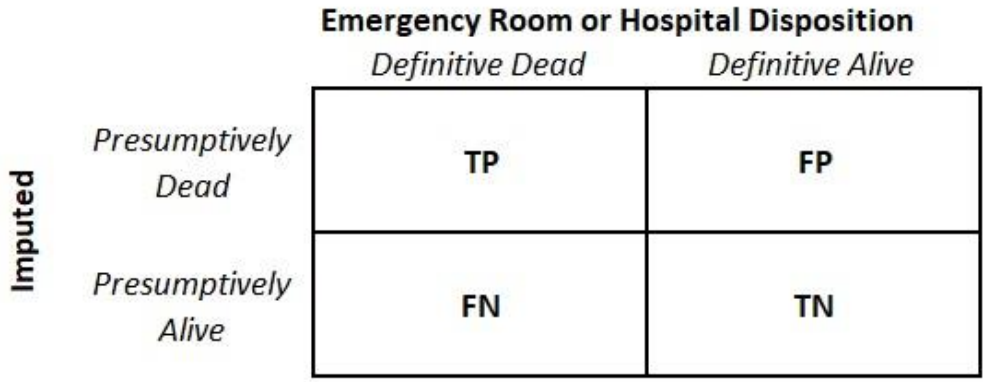

Figure 3. Metrics used in the present study to assess MLB predictions (18,19,20,21); computations via the scikit-learn Python development library (23).

Sub-Sample Mortality Rate (MR), Sub-Sample Survival Rate (SR) (Prevalence)	$MR = \frac{TP + FN}{TP + FN + TN + FP}$, $SR = \frac{TN + FP}{TP + FN + TN + FP}$
Overall Accuracy (ACC) *sklearn.metrics.accuracy_score*	$ACC = \frac{TP + TN}{TP + FN + TN + FP}$
True Positive Rate (TPR), "Dead" Category Accuracy (Sensitivity, Recall, Hit Rate, Detected Rate)	$TPR = \frac{TP}{TP + FN}$
True Negative Rate (TNR), "Alive" Category Accuracy (Specificity, Selectivity)	$TNR = \frac{TN}{TN + FP}$
Balanced Accuracy (BACC) *sklearn.metrics.balanced_accuracy_score*	$BACC = \frac{TPR + TNR}{2}$
Precision (PRE) (Positive Predictive Value, PPV), *sklearn.metrics.precision_score*	$PRE = \frac{TP}{TP + FP}$
F1 Score (F1) (F1 Measure), *sklearn.metrics.f1_score*	$F1 = \frac{2 \times PRE \times TPR}{PRE + TPR}$
Cohen's Kappa Coefficent (COH) *sklearn.metrics.cohen_kappa_score*	$COH = \frac{2 \times [(TP \times TN) - (FN \times FP)]}{[(TP + FP)(TN + FP)] + [(TP + FN)(TN + FN)]}$
Matthews Coefficient (MAT) *sklearn.metrics.matthews_corrcoef*	$MAT = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FN)}}$
Hamming Loss (HL) *sklearn.metrics.hamming_loss*	$HL = \frac{\sum_{i=1}^{M} (y_i \neq \hat{y}_i)}{M}$ $y_i, \hat{y}_i = \begin{cases} 1 & \text{if def, presumpt "Dead"} \\ 0 & \text{if def, presumpt "Alive"} \end{cases}$
Jaccard Similarity ($\mathcal{J}$) *sklearn.metrics.jaccard_score*	$\mathcal{J} = \frac{\sum_{i=1}^{M} (y_i \,\&\, \hat{y}_i)}{\sum_{i=1}^{M} (y_i \mid \hat{y}_i)}$ $y_i, \hat{y}_i = \begin{cases} 1 & \text{if def, presumpt "Dead"} \\ 0 & \text{if def, presumpt "Alive"} \end{cases}$

TP ≡ Number of True Positives; i.e. truely "Dead" and imputed "Dead"; FN ≡ Number of False Negatives; i.e. truely "Dead" but imputed "Alive";
TN ≡ Number of True Negatives; i.e. truely "Alive" and imputed "Alive"; FP ≡ Number of False Positives; i.e. truely "Alive" but imputed "Dead".
M ≡ Total number of instances in sub-sample.