**Appendix**

**Table 1.** Literature Search Terms to Identify Early Warning System Studies Using Multivariable Regression or Machine Learning for Inclusion in a Systematic Review

|  |
| --- |
| **PubMed and CINAHL Plus** |
| Timeframe: 1/1/2012 - 9/15/2018Search terms: “early warning score OR early warning system AND deterioration OR predict transfer ICU” Search details: (early[All Fields] AND warning[All Fields] AND score[All Fields]) OR (early[All Fields] AND warning[All Fields] AND system[All Fields]) AND deterioration[All Fields] OR (predict[All Fields] AND ("transfer (psychology)"[MeSH Terms] OR ("transfer"[All Fields] AND "(psychology)"[All Fields]) OR "transfer (psychology)"[All Fields] OR "transfer"[All Fields]) AND ("intensive care units"[MeSH Terms] OR ("intensive"[All Fields] AND "care"[All Fields] AND "units"[All Fields]) OR "intensive care units"[All Fields] OR "icu"[All Fields])) AND ("2012/01/01"[PDAT] : "2018/09/15"[PDAT]) |

**Table 2.** Screening Inclusion and Exclusion Criteria

|  |  |  |
| --- | --- | --- |
| **Selection Criterion** | **Included** | **Excluded** |
| Research Population | Hospitalized adults ($\geq $18 years)  | Adults under observation statusObstetric patientsPost-surgical patientsPediatric patients |
| Setting | General Medical-Surgical wardsStep-Down wards  | Intensive care unitTransitional care unitEmergency roomLabor & deliveryOperating roomOncology wardPrimary care |
| Timeframe | January 1, 2012- September 15, 2018 | Before 2012 |
| Method | Quantitative  | Mixed methodQualitativeCase reports or commentaries |
| Model | EHRa-based Multivariable regressionMachine learning | Paper-basedAggregate-weighted EWSb only |
| Predictors | Vitals signsLaboratory valuesSeverity of illness scoresComorbidity scoresCode Status and other EHRa data | Monitor data (wave forms) |
| Outcome | Composite of ICUc transfer and mortality | RRTd activationSepsisCardiac arrest onlyMortality only |
| Model Performance | AUCe (required)SensitivitySpecificityPositive Predictive ValueRRTd workload (workup to detection ratio) | Risk ratiosOdds ratiosChi SquareANOVA or other comparison of groups |

Note.

a Electronic Health Record

b Early Warning System

c Intensive Care Unit

d Rapid Response Team

e Area Under the [Receiver Operator] Curve

**Table 3.** Measures of Model Performance

|  |  |  |
| --- | --- | --- |
| **Measure Name** | **Description** | **Formula** |
| Pre-test probability | Prevalence: % of those with the outcome among the sample | $$\frac{cases}{entire sample}$$ |
| Pseudo-R2 a | % of variation explained by the model | (not applicable) |
| Sensitivity | % true positive cases among all positive cases | $$\frac{true positives}{true positives+false negatives}$$ |
| Specificity | % true negative cases among all negative cases | $$\frac{true negatives}{true negatives+false positives}$$ |
| PPV | % true positive cases among all positive tests | $$\frac{true positives}{true positives+false positives}$$or$$\frac{sensitivity\*prevalence}{sensitivity\*prevalence+\left(1-specificity\right)\*(1-prevalence)}$$ |
| AUC/c-stat | True positive (TP) rate plotted against false positive (FP) rate | $$\frac{Number of concordant pairs}{Total number of pairs}+0.5\*\frac{Number of tied pairs}{Total number of pairs}$$ |
| Workup-to-Detection | Workload measure: Number needed to evaluate to find one positive case | $$\frac{true positives+false positives}{true positives}$$or$$\frac{1}{WDR}$$ |
| RRT evaluations per hospital per day | Workload measure: The total number of patients RRTs need to evaluate per day (round up to full integer)  | $$WDR\*\frac{ cases }{\frac{hospitals}{days}}$$ |

Note.

 aLogistic regression does not use R2 but Likelihood ratio R2, Cox and Snell R2, Nagelkerke R2 or others

**Table 4.** Predictive Model Characteristics and Model Performance of 6 Early Warning Systems Using Multivariable Regression or Machine Learning to Identify Clinical Deterioration Risk

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Prediction method/predictor variables** | **Reference standard used** | **Sensitivity,****Specificity,****Positive Predictive Value,****Negative Predictive Value** | **AUROC/c-statistic** | **Calibration metric** | **Workup to Detection Ratio** | **Relevant Findings** | **Strengths** | **Limitations** |
| Escobar et al., 2012 | Laboratory tests, vitals signs, shock index, age, sex, LAPS1, COPS1, admission diagnosis, admission type, code status, length of stay | Non-events were comparison | Not reportedNot reportedNot reportedNot reported | 0.78 for EDIP (Ranging from .68 to .84 across diagnostic strata)0.70 for MEWS (ranging from 0.54 to 0.79 across diagnostic strata) | Not discussed | Workup volume for MEWS threshold of >=6:EDIP: 14.5 false alarms for each ICU transfer MEWS: 34.4 false alarms for each transfer | EDIP outperformed manual MEWS scoring system in all models | Very large dataset, very complex risk adjustment, very precise variables and methods | Large integrated health system with fully integrated EHR, computational infrastructure limited, unable to determine if ward patient should have been a ICU admit |
| Alvarez, et al., 2013 | Laboratory data, vital signs, level of consciousness, STAT orders, STAT medications, MEWS, “high risk floor” | Non-events were comparison | Regression Model0.52MEWS: 0.42Regression Model0.94MEWS0.91Regression Model0.10MEWS 0.06Regression Model0.99MEWS 0.99 | Regression Model0.85 MEWS0.75 | Hosmer-Lemeshow p-value for calibration | Median number of alarms per day: 9Median number of RRT calls per day: 2 | The automated EHR model performed better than MEWS alone and reduced number of false positive alarms. The model was twice as sensitive as manual RRT activation (0.52 vs. 0.26) and trigger 5.7 hours sooner than RRT. | Provided important clinical comparison of RRT activation (human vigilance) and basic MEWS. Demonstrated that EWS accuracy can be improved by regression techniques. | Single center study with small cohort. Included all wards deaths as “unexpected” |
| Churpek et al., 2014 |  Patient demographics, vital signs, mental status, laboratory test values,  | Non-events were comparison | 0.16-0.89 depending on model risk score cutoff0.54 at model score of $\geq $170.52-0.99 depending on model risk score cutoff0.90 at model score of $\geq $17Not reportedNot reported | 0.77 for eCART (combined outcomes)0.70 for MEWS | Calculated predicted event probability. Did not discuss a calibration metric | Did not discuss workup or similar workload metric for selected risk score cutoff | eCART performed substantially better than MEWS, likely because model was more complex | Large dataset, complex set of covariates, very detailed analytic approach | Did not include comorbidity or severity of illness score, did not discuss workload generated by score |
| Churpek et al. 2016 | Age, Length of stay, number of prior ICU stays, vital signs, laboratory values | Non-events were comparison | Not reportedNot reportedNot reportedNot reported | Random forest: 0.80Gradient boosted machine: 0.79Bagged trees: 0.79Support vector machine: 0.79Neural network: 0.78Logistic regression (spline): 0.77K-nearest neighbor: 0.75Logistic regression (linear): 0.74Decision tree: 0.73MEWS: 0.70 | Hosmer-Lemeshow p-value for calibration and O/E plotting | At 75% sensitivity level Random Forrest model would screen 13% fewer than logistic linear model or more than 500,000 fewer screens out of a pool of 4.6 million observations | Machine learning algorithms were superior to traditional regression models and both RF and GBM had very good discrimination and calibration | Introduced novel “data science” machine learning methods that show superior performance to traditional supervised predictive analytics approaches (regression). Large sample size.  | Black box output (clinicians cannot understand why a patient scores high).Composite outcome does not seem to account for expected deaths. |
| Kipnis et al., 2016 | Laboratory test values, vital signs, comorbidity composite (COPS2), acute physiological instability index (LAPS2), Length of stay, age, sex, code status, time of day, season, admit category, hospital  | Non-events were comparison | 0.38-0.56 (across medical centers)0.88-0.95 (across medical centers)0.11-0.23 (across medical centers)0.97-0.99 (across medical centers) | 0.82 for AAM0.79 for eCART0.76 for NEWS | Hosmer-Lemeshow p-value for calibration | Developed workup-to-detection ratio and also defined an operational alarm cutoff so that model was calibrated against maximum of one alarm per 35-bed unit per day  | AAM performed better than eCART and NEWS, likely because model was more complex | Very large dataset, very complex risk adjustment, very precise variables and methods | Large integrated health system, computational infrastructure limited method (now working on more machine learning models) |
| Green et al., 2018 | Laboratory test values, vital signs, patient demographics | Non-events were comparison | 0.16 - 0.81 depending on model risk score cutoff0.50 at model score of $\geq $90.60 - 0.99 depending on model risk score cutoff0.90 at model score of $\geq $9Not reportedNot reported | eCART (random forest): 0.801 (0.799-0.802)NEWS: 0.72 (0.716-0.720)MEWS: 0.70 (0.696-0.700)Between the flags: 0.66 (0.661-0.664) | Not discussed. See Churpek et al., 2016 | Compared number of patients identified and number of false positives. eCART identified fewer false positives and more true positives than aggregate-weighted models | eCART model is more accurate and generates fewer evaluations than aggregate-weighted models by adding additional clinical covariates to its model. | Validated a previous study by Churpek et al. (2016) which introduced the machine learning eCART model. Large sample size.  | Same sample than prior setting with an additional 6 months of hospitalization data. Composite outcome does not seem to account for expected deaths. |

**Table 5.** Level of Scientific Evidence and Risk of Bias Assessment

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Level of Scientific Evidence based on Research Design (1: High - 7: Low)** | **Measurement Bias****Systematic differences in applying measurement** | **Detection Bias****Systematic differences in outcome measurement** | **Missing data Bias****Systematic differences in data sets** | **Threats to External Validity** | **Total Score Presence of Bias)** |
| Escobar et al., 2012 | 4 (not used in score) | 0: Used sophisticated adjustment techniques to account for confounding and validated model on hold-out dataset | 0.5: Outcome was clearly defined and clinical variations in patient presentation were included. Though nearly impossible to design programmatically, a small fraction of the conceptualized events may not have been appropriate ward admissions (error in judgment). Patient may have entered as full code but became an “appropriate” death event (palliative care) | 0: Discussed, models with imputed data and dropped observations were compared | 0.5. Large health system study in integrated care delivery network in NCAL. Plan members may be receiving better care at baseline. NCAL demographic and income/SES not generalizable to all settings (limited to similar metropolitan regions with similar make-up) | 1 |
| Alvarez, et al., 2013 | 4 (not used in score) | 1: Used sophisticated adjustment techniques to account for confounding and validated model on hold-out datasetUnexpected death definition included those made DNR or comfort care (an inaccurate measurement)Neuro status was based on natural language processing search of nursing notes (validity and reliability not discussed) | 0.5: Outcome was clearly defined and clinical variations in patient presentation were included. Though nearly impossible to design programmatically, a small fraction of the conceptualized events may have been misclassified | 1: Missing data not discussed, though likely a concern | 1: Single center studyDid not report demographics of cases and controls  | 3.5 |
| **Study** | **Level of Scientific Evidence based on Research Design (1: High - 7: Low)** | **Measurement Bias****Systematic differences in applying measurement** | **Detection Bias****Systematic differences in outcome measurement** | **Missing data Bias****Systematic differences in data sets** | **Threats to External Validity** | **Total Score Presence of Bias)** |
| Churpek et al., 2014 | 4 (not used in score) | 0.5: Used sophisticated adjustment techniques to account for confounding and validated model on hold-out datasetData were from two different health systems, potential for different documentation standards was not discussed | 1: Outcome was clearly defined and clinical variations in patient presentation were included. Though nearly impossible to design programmatically, a small fraction of the conceptualized events may have been misclassifiedConfirmation bias: Conflicts of interest: One researcher disclosed honoraria from a clinical alarm vendor | 0: Discussed, missing data were imputed | 0.5. Small-medium health system study (5 hospitals). Demographics reported in Table 1 | 2 |
| Churpek et al. 2016 | 4 (not used in score) | 0.5: Used sophisticated adjustment techniques to account for confounding and validated model on hold-out datasetData were from two different health systems, potential for different documentation standards was not discussed | 1: Outcome was clearly defined and clinical variations in patient presentation were included. Though nearly impossible to design programmatically, a small fraction of the conceptualized events may have been misclassified Confirmation bias: Conflicts of interest: Two researchers have a patent pending for a risk algorithm that may become commercially available. One researcher disclosed honoraria from a clinical alarm vendor | 0: Discussed, missing data were imputed | 0.5. Small-medium health system study (5 hospitals). Demographics not reported in text but same sample as 2014 paper | 2 |
| **Study** | **Level of Scientific Evidence based on Research Design (1: High - 7: Low)** | **Measurement Bias****Systematic differences in applying measurement** | **Detection Bias****Systematic differences in outcome measurement** | **Missing data Bias****Systematic differences in data sets** | **Threats to External Validity** | **Total Score Presence of Bias)** |
| Kipnis et al., 2016 | 4 (not used in score) | 0: Used sophisticated adjustment techniques to account for confounding and validated model on hold-out dataset | 0.5: Outcome was clearly defined and clinical variations in patient presentation were included. Though nearly impossible to design programmatically, a small fraction of patients may have had a first RRT call or code blue event without the outcome but subsequent deterioration. There were no data used for RRT activation or code blue. | 0: Discussed, missing data were imputed  | 0.5. Large health system study in integrated care delivery network in NCAL. Plan members may be receiving better care at baseline. NCAL demographic and income/SES not generalizable to all settings (limited to similar metropolitan regions with similar make-up) | 1 |
| Green et al., 2018 | 4 (not used in score) | 0.5: No validation in hold-out dataset. Data were from two different health systems, potential for different documentation standards was not discussed | 1: Outcome was clearly defined and clinical variations. Though nearly impossible to design programmatically, a small fraction of the conceptualized events may have been misclassified Confirmation bias: Conflicts of interest: Two researchers have a patent pending for a risk algorithm that may become commercially available. One researcher disclosed honoraria from a clinical alarm vendor | 0: Discussed, missing data were imputed  | 0.5. Small-medium health system study (5 hospitals). Demographics reported in Table 1 | 2 |

Note. Adopted from Higgins et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomized trial

**Table 6.** Sources of Clinical and Methodological Heterogeneity Across Selected Studies

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Setting** | **Putative impact on overall PPV** | **Mortality outcome definition** | **Putative impact on overall PPV** | **Event Rate** | **Putative impact on overall PPV** | **Selection of observations** | **Putative impact on overall PPV** |
| Escobar et al., 2012 | 14 Kaiser Permanente community hospitals | Assuming the severity of illness in community hospitals as the baseline, this setting has good generalizable properties, at least for similar demographics | Deathoutside the ICU among patients who were ‘‘full code’’ Excludes patients with DNR and comfort care | This definition attempts to account for patients who may be on an end-of-life trajectory, however not all DNRs experience an expected death. Impact on PPV is unknown. | 3.9% | Because the mean event rate was higher (5%), PPV will be lower but may be closer to the true average. | Several transfers to ICU in the same patient were permitted | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |
| Alvarez, et al., 2013 | 1 university hospitalMean severity of illness may be higher than in community hospitals | Presence of sicker patients may improve detectability of deterioration and a higher prevalence would boost PPV. | Unexpected death:1) an in-hospitaldeath that occurred on the medical ward; or 2)death that occurred in patients transferred to a medical orcardiac ICU team with an ICU length of stay <24 hours | This definition counts any death (including DNR and comfort care), and may inflate the numerator. This would increase PPV. | 7.8% | Higher event rate will increase PPV; the true average PPV may be lower. | Not discussedAt minimum, the first observed outcome | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |
| Churpek et al., 2014 | 1 university medical center, 2 teaching hospitals,2 community hospitalsMean severity of illness may be higher than community hospitals | Presence of sicker patients may improve detectability of deterioration and a higher prevalence would boost PPV. | Death on theward without activation of the cardiacarrest team | This method would exclude cardiac arrest patients who die on the ward (but counts all cardiac arrests)Impact on PPV is unknown. | 6.1% | Higher event rate will increase PPV; the true average PPV may be lower. | Not discussedAt minimum, the first observed outcome | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |
| Churpek et al. 2016 | 1 university medical center, 2 teaching hospitals,2 community hospitalsSee above | Presence of sicker patients may improve detectability of deterioration and a higher prevalence would boost PPV. | Death on the ward withoutattempted resuscitation | This method would exclude cardiac arrest patients who die on the ward (but counts all cardiac arrests)Impact on PPV is unknown. | 6.1% | Higher event rate will increase PPV, but the true average may be lower. | Not discussedAt minimum, the first observed outcome | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |
| Kipnis et al., 2016 | 21 Kaiser Permanente community hospitals | Assuming the severity of illness in community hospitals as the baseline, this setting has good generalizable properties, at least for similar demographics | Death outside the ICU in a patient whose care directivewas ‘‘full code”Excludes patients with DNR and comfort care | This definition attempts to account for patients who may be on an end-of-life trajectory, however not all DNRs experience an expected death. Impact on PPV is unknown. | 3.0% | This is the lowest observed event rate. Because the mean event rate across studies was higher (5%), PPV will be lower but may be closer to the true average. | Not discussedAt minimum, the first observed outcome | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |
| Green et al., 2018 | 1 university hospital, 2 teaching hospitals,2 community hospitals | Presence of sicker patients may improve detectability of deterioration and a higher prevalence would boost PPV. | Death on the ward occurring within 24 h of an observation | This definition counts any death (including DNR and comfort care), and may inflate the numerator. This would increase PPV. | 5.7% | Higher event rate will increase PPV, but the true average may be lower. | Not discussedAt minimum, the first observed outcome | Patients who transfer to ICU several times may reduce the model’s true predictive capabilities. |

**Table 7.** Comparison of EWS model performance (AUC) in Original vs External Patient Populations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Aggregate weighted EWS** | **External validation** | **Absolute performance drop** | **Advanced EWS** | **External validation** | **Absolute performance drop** |
| NEWS AUC in original Smith et al. (2013) paper: 0.87 | NEWS AUC in Kipnis et al. (2016): 0.76 | 11% | eCART AUC in Churpek et al. (2016): 0.80 | eCART AUC in Kipnis paper (2016): 0.79 | 1% |
|  | NEWS AUC in Green et al. (2018): 0.72 | 15% |  |  |  |