Professional Certificate in AI-Enhanced Health Coaching Support Systems

Evaluation Metrics for AI Health Interventions

Accuracy – Concept: Overall correctness of AI predictions. Related terms: precision, recall, specificity. Explanation: Proportion of true results (both positive and negative) among all evaluated cases. Example: An AI model correctly classifies 85 of 100 patient records, yielding 85% accuracy. Practical application: Baseline performance check for diagnostic chatbots. Challenges: Can be misleading in imbalanced datasets where majority class dominates.

Area Under the Curve (AUC) – Concept: Aggregate measure of performance across all classification thresholds. Related terms: ROC curve, c-statistic. Explanation: The probability that a randomly chosen positive instance scores higher than a randomly chosen negative one. Example: An AI-driven risk stratifier with AUC = 0.92 Shows excellent discrimination. Practical application: Comparing multiple models for cardiovascular risk prediction. Challenges: AUC may hide poor performance at clinically relevant thresholds.

Balanced Accuracy – Concept: Average of sensitivity and specificity. Related terms: accuracy, class imbalance. Explanation: Mitigates bias toward majority class by giving equal weight to both classes. Example: A model with 70% sensitivity and 90% specificity yields balanced accuracy of 80%. Practical application: Evaluating AI tools for rare disease detection. Challenges: Still sensitive to extreme imbalance; may not reflect clinical utility.

Calibration – Concept: Agreement between predicted probabilities and observed outcomes. Related terms: reliability, Brier score. Explanation: A well-calibrated model predicts 10% risk and, over many cases, roughly 10% actually experience the event. Example: A hypertension-prediction AI outputs 0.2 Risk for 100 patients; 20 develop hypertension. Practical application: Informing shared decision-making in lifestyle coaching. Challenges: Calibration can drift over time as population health changes.

Confusion Matrix – Concept: Tabular summary of prediction outcomes. Related terms: true positive, false negative. Explanation: Displays counts of TP, FP, FN, TN, enabling calculation of many metrics. Example: 50 TP, 10 FP, 5 FN, 35 TN for a depression-screening AI. Practical application: Quick diagnostic audit of a health-coach recommendation engine. Challenges: Interpretation becomes complex with multi-class outputs.

Cost-Effectiveness Analysis (CEA) – Concept: Economic evaluation comparing costs to health outcomes. Related terms: QALY, incremental cost-effectiveness ratio (ICER). Explanation: Determines whether an AI intervention provides sufficient health benefit per monetary unit spent. Example: AI-enhanced diet counseling costs $200 per patient and yields 0.03 Additional QALYs, ICER = $6,667/QALY. Practical application: Budgeting decisions for hospital AI rollouts. Challenges: Assigning monetary values to intangible benefits like patient empowerment.

Cross-Validation – Concept: Technique for assessing model generalizability. Related terms: k-fold, hold-out set. Explanation: Data are split into k subsets; each subset serves once as test while the remaining k-1 train

the model. Example: 5-Fold cross-validation reports mean AUC = 0.88 For a lifestyle-adherence predictor. Practical application: Robust performance estimation before deployment. Challenges: Computationally intensive for large neural networks; may still overestimate performance if data leakage occurs.

Cumulative Gain – Concept: Measure of how many positive outcomes are captured as the list is traversed. Related terms: lift chart, gain curve. Explanation: Plots proportion of true positives captured versus proportion of population screened. Example: Top 20% of AI-ranked patients contain 60% of high-risk cases, indicating a gain of 3. Practical application: Prioritizing outreach in chronic disease management. Challenges: Requires reliable ground truth and may be unstable with small sample sizes.

Decision Curve Analysis (DCA) – Concept: Evaluates clinical net benefit across threshold probabilities. Related terms: net benefit, threshold probability. Explanation: Compares the value of using a model versus treating all or none, incorporating patient preferences. Example: An AI risk model shows higher net benefit than standard care for thresholds between 10% and 30%. Practical application: Selecting AI tools for personalized coaching intensity. Challenges: Requires accurate estimation of harms and benefits, which can be subjective.

Diagnostic Odds Ratio (DOR) – Concept: Single indicator of test effectiveness. Related terms: sensitivity, specificity. Explanation: Ratio of the odds of positivity in diseased versus non-diseased groups (TP/FN ÷ FP/TN). Example: DOR = 12 suggests the AI test is 12 times more likely to correctly identify disease than miss it. Practical application: Summarizing performance of AI-based triage systems. Challenges: Non-intuitive magnitude; does not convey direction of errors.

Discrimination – Concept: Ability of a model to separate those with and without the outcome. Related terms: AUC, c-statistic. Explanation: Higher discrimination means greater separation between predicted risk distributions. Example: A model with AUC = 0.95 Discriminates well between patients who will and will not develop diabetes. Practical application: Selecting models for risk-based coaching. Challenges: High discrimination does not guarantee good calibration.

Effect Size – Concept: Magnitude of difference attributable to an intervention. Related terms: Cohen's d, hazard ratio. Explanation: Quantifies practical significance beyond statistical significance. Example: AI-guided exercise program yields Cohen's d = 0.6 Improvement in VO_2max versus standard care. Practical application: Communicating benefits to stakeholders. Challenges: Depends on variability of the outcome; may be inflated in small samples.

F1 Score – Concept: Harmonic mean of precision and recall. Related terms: precision, recall. Explanation: Balances false positives and false negatives, useful for imbalanced classes. Example: Precision = 0.8, Recall = 0.6 → F1 = 0.69. Practical application: Evaluating AI alerts for medication non-adherence. Challenges: Does not consider true negatives; may be less relevant when specificity is critical.

False Discovery Rate (FDR) – Concept: Proportion of false positives among all positive calls. Related terms: type I error, positive predictive value. Explanation: Low FDR indicates that most flagged cases are true. Example: An AI screening tool with 5% FDR means 95% of alerts are genuine. Practical application: Reducing unnecessary follow-up in tele-health platforms. Challenges: Depends on prevalence; can be high in

low-prevalence populations.

False Positive Rate (FPR) – Concept: Proportion of negatives incorrectly labeled as positive. Related terms: 1-specificity, type I error. Explanation: FPR = FP/(FP+TN). Example: 15 False alerts out of 200 healthy users → FPR = 7.5%. Practical application: Assessing alarm fatigue in AI-driven monitoring devices. Challenges: High FPR can erode trust among clinicians.

Gini Coefficient – Concept: Measure of inequality in model predictions. Related terms: AUC, Lorenz curve. Explanation: Gini = 2 × AUC – 1; higher values indicate better discrimination. Example: AUC = 0.80 → Gini = 0.60. Practical application: Quick comparison of multiple AI models for health risk scoring. Challenges: Same limitations as AUC; not informative about calibration.

Hazard Ratio (HR) – Concept: Relative risk over time between two groups. Related terms: survival analysis, Cox model. Explanation: HR > 1 indicates higher hazard in the treatment group; HR Incremental Cost-Effectiveness Ratio (ICER) – Concept: Additional cost per additional unit of effect. Related terms: QALY, CEA. Explanation: ICER = $(Cost_1 – Cost_0)/(Effect_1 – Effect_0)$. Example: AI-enhanced nutrition counseling costs $500 more and yields 0.05 Extra QALYs → ICER = $10,000/QALY. Practical application: Informing reimbursement decisions. Challenges: Sensitive to small differences in effect; ethical concerns over valuing life years.

Interpretability – Concept: Degree to which a human can understand model decisions. Related terms: explainable AI, black-box. Explanation: Transparent models enable clinicians to trust and act on AI recommendations. Example: A decision-tree model shows weight-loss recommendation driven by BMI > 30. Practical application: Integrating AI insights into patient counseling sessions. Challenges: Trade-off between interpretability and predictive performance.

Kaplan-Meier Estimate – Concept: Non-parametric survival curve estimator. Related terms: censoring, hazard ratio. Explanation: Plots probability of remaining event-free over time. Example: AI-guided cardiac rehab shows 80% event-free survival at 12 months versus 70% for standard care. Practical application: Visualizing outcomes of AI-supported interventions. Challenges: Does not adjust for covariates; requires sufficient follow-up.

Kappa Statistic – Concept: Agreement measure beyond chance. Related terms: inter-rater reliability, Cohen's kappa. Explanation: Values range from −1 (complete disagreement) to 1 (perfect agreement). Example: AI-annotated medical images achieve $\kappa = 0.82$ With radiologists, indicating strong agreement. Practical application: Validating AI labeling tools. Challenges: Affected by prevalence; may be low even with high accuracy in skewed datasets.

Lift – Concept: Improvement of model over random selection. Related terms: gain chart, cumulative gain. Explanation: Lift = (predicted positive rate)/(overall positive rate). Example: Top decile of AI risk scores captures 40% of cases while overall prevalence is 10% → lift = 4. Practical application: Targeting limited coaching resources to high-impact patients. Challenges: Lift diminishes as more of the population is screened.

Log-Loss (Cross-Entropy Loss) – Concept: Penalizes confident but incorrect predictions. Related terms:

binary cross-entropy, likelihood. Explanation: Lower values indicate better calibrated probability estimates. Example: Model with log-loss = 0.35 Outperforms one with 0.60 On validation data. Practical application: Training deep learning models for symptom triage. Challenges: Sensitive to outliers; may not reflect clinical relevance.

Mean Absolute Error (MAE) – Concept: Average magnitude of errors in continuous predictions. Related terms: RMSE, bias. Explanation: MAE = Σ|prediction – actual|/n. Example: AI predicts daily step count with MAE = 1,200 steps. Practical application: assessing precision of activity-tracking algorithms. Challenges: treats all errors equally; does not penalize large deviations as heavily as RMSE.

Mean Squared Error (MSE) – Concept: Average squared difference between predicted and actual values. Related terms: RMSE, variance. Explanation: MSE = Σ(prediction – actual)^2/n. Example: AI-estimated blood pressure has MSE = 25 mmHg^2, implying RMSE ≈ 5 mmHg. Practical application: Regression models for continuous health metrics. Challenges: Heavily penalizes outliers; not directly interpretable in original units.

Net Reclassification Improvement (NRI) – Concept: Quantifies improvement in risk category assignment. Related terms: integrated discrimination improvement (IDI), reclassification. Explanation: Sum of correctly moved individuals minus incorrectly moved ones across risk thresholds. Example: AI model reclassifies 15% of patients to more appropriate risk categories, yielding NRI = 0.15. Practical application: Justifying migration from legacy risk scores to AI-based tools. Challenges: Depends on chosen thresholds; can be inflated with many categories.

Negative Predictive Value (NPV) – Concept: Probability that a negative test result is truly negative. Related terms: specificity, prevalence. Explanation: NPV = TN/(TN+FN). Example: AI screening for hypertension yields NPV = 0.96, Meaning 96% of those flagged as low risk remain normotensive. Practical application: Safely ruling out disease in remote monitoring. Challenges: Declines as disease prevalence rises.

Odds Ratio (OR) – Concept: Odds of outcome in exposed versus unexposed groups. Related terms: logistic regression, relative risk. Explanation: OR > 1 indicates higher odds with exposure; OR Precision – Concept: Proportion of true positives among all positive predictions. Related terms: positive predictive value, F1 score. Explanation: Precision = TP/(TP+FP). Example: AI alerts for fall risk have precision = 0.78, Meaning 78% of alerts correspond to actual falls. Practical application: Reducing unnecessary interventions. Challenges: High precision may be achieved at expense of recall.

Positive Predictive Value (PPV) – Concept: Likelihood that a positive test reflects true condition. Related terms: precision, prevalence. Explanation: PPV = TP/(TP+FP). Example: AI-driven diabetes detection yields PPV = 0.85. Practical application: Confidence in AI-generated diagnoses. Challenges: Strongly influenced by disease prevalence; can be low in low-prevalence settings despite good sensitivity.

Probabilistic Forecast – Concept: Prediction expressed as a probability distribution. Related terms: prediction interval, calibration. Explanation: Provides a range of likely outcomes with associated likelihoods. Example: AI predicts 30% chance of medication non-adherence within 30 days. Practical application: Tailoring motivational messaging intensity. Challenges: Requires robust uncertainty quantification; over-confident forecasts erode trust.

Propensity Score Matching (PSM) – Concept: Technique to balance covariates between treatment groups. Related terms: causal inference, confounding. Explanation: Matches individuals with similar probability of receiving the intervention, based on observed characteristics. Example: AI-supported coaching participants matched to non-participants via PSM to assess impact on HbA1c. Practical application: Observational evaluation of AI health programs. Challenges: Only controls for observed confounders; hidden bias may remain.

Recall (Sensitivity) – Concept: Ability to identify true positives. Related terms: true positive rate, miss rate. Explanation: Recall = TP/(TP+FN). Example: AI symptom checker captures 92% of influenza cases (recall = 0.92). Practical application: Ensuring critical conditions are not missed. Challenges: Increasing recall often lowers precision, leading to more false alarms.

Receiver Operating Characteristic (ROC) Curve – Concept: Plot of true-positive rate versus false-positive rate across thresholds. Related terms: AUC, threshold analysis. Explanation: Visual tool to assess discrimination and select operating point. Example: ROC curve shows optimal cutoff at 0.65 Probability for maximizing Youden's index. Practical application: Choosing decision thresholds for AI triage bots. Challenges: ROC can be overly optimistic in highly imbalanced data; precision-recall curves may be more informative.

Relative Risk Reduction (RRR) – Concept: Proportion by which risk is reduced in the treatment group. Related terms: absolute risk reduction (ARR), number needed to treat (NNT). Explanation: RRR = (Incidence_control – Incidence_treatment)/Incidence_control. Example: AI-guided weight-loss program lowers obesity incidence from 20% to 12% → RRR = 40%. Practical application: Communicating benefits to patients. Challenges: Can be misleading without absolute risk context.

Root Mean Squared Error (RMSE) – Concept: Square root of MSE; reflects typical prediction error magnitude. Related terms: MAE, standard deviation. Explanation: RMSE = √MSE, expressed in original units. Example: RMSE = 4 mmHg for AI-estimated blood pressure. Practical application: Evaluating regression models for vital-sign prediction.

Sample Size Calculation – Concept: Determination of the number of participants needed for adequate statistical power. Related terms: effect size, power analysis. Explanation: Incorporates anticipated effect, variability, significance level, and desired power. Example: Detecting a 0.5 % HbA1c reduction with 80% power requires 350 participants per arm. Practical application: Designing trials of AI-enabled health coaching. Challenges: Under-estimation leads to inconclusive results; over-estimation inflates cost.

Sensitivity Analysis – Concept: Testing robustness of results to changes in assumptions or parameters. Related terms: scenario analysis, uncertainty quantification. Explanation: Systematically vary inputs (e.G., Cost, adherence) to observe impact on outcomes. Example: Varying AI adoption rate from 30% to 70% changes cost-effectiveness by ±15%. Practical application: Informing policy decisions under uncertainty. Challenges: Can become complex with many interacting variables.

Specificity – Concept: Ability to correctly identify true negatives. Related terms: true negative rate, false positive rate. Explanation: Specificity = TN/(TN+FP). Example: AI screening for skin cancer achieves specificity of 0.94. Practical application: Minimizing unnecessary biopsies. Challenges: High specificity may reduce

sensitivity, risking missed diagnoses.

Standardized Mortality Ratio (SMR) – Concept: Observed deaths divided by expected deaths based on a reference population. Related terms: risk adjustment, benchmarking. Explanation: SMR > 1 indicates higher mortality than expected. Example: AI-augmented postoperative monitoring shows SMR = 0.85, Suggesting reduced mortality. Practical application: Evaluating safety impact of AI tools. Challenges: Requires accurate baseline rates; confounding can bias interpretation.

Survival Analysis – Concept: Statistical methods for time-to-event data. Related terms: Cox proportional hazards, Kaplan-Meier. Explanation: Models account for censoring and estimate hazard functions. Example: AI-generated adherence scores predict time to relapse using Cox regression. Practical application: Planning duration of coaching interventions. Challenges: Proportional hazards assumption may be violated; requires sufficient follow-up.

Time-Dependent ROC – Concept: ROC analysis that incorporates the timing of events. Related terms: survival ROC, c-index. Explanation: Evaluates discriminative ability at specific time horizons. Example: AI model shows AUC = 0.78 For predicting 6-month cardiovascular events. Practical application: Selecting models for short-term risk prediction. Challenges: Requires accurate event times; computationally intensive.

True Positive Rate (TPR) – Concept: Same as sensitivity; proportion of actual positives correctly identified. Related terms: recall, ROC curve. Explanation: TPR = TP/(TP+FN). Example: TPR = 0.90 For AI detection of atrial fibrillation episodes. Practical application: Ensuring critical alerts are captured. Challenges: High TPR may increase false positives if threshold is low.

True Negative Rate (TNR) – Concept: Same as specificity; proportion of actual negatives correctly identified. Related terms: specificity, ROC curve. Explanation: TNR = TN/(TN+FP). Example: TNR = 0.96 For AI-based sleep-apnea screening. Practical application: Avoiding unnecessary referrals. Challenges: May be sacrificed when attempting to raise sensitivity.

Uncertainty Quantification – Concept: Assessment of confidence in model predictions. Related terms: prediction intervals, Bayesian methods. Explanation: Provides ranges (e.G., 95% CI) rather than point estimates. Example: AI predicts weight loss of 5 kg ± 2 kg. Practical application: Informing patients about expected variability of outcomes. Challenges: Requires sophisticated modeling; may be computationally demanding.

Validation Cohort – Concept: Independent dataset used to assess model performance. Related terms: external validation, generalizability. Explanation: Ensures that results are not specific to training data. Example: Model trained on US patients validated on European cohort, maintaining AUC = 0.84. Practical application: Confirming readiness for deployment across regions. Challenges: Data heterogeneity can cause performance drop; access to external data may be limited.

Variable Importance – Concept: Ranking of predictors based on contribution to model performance. Related terms: feature importance, SHAP values. Explanation: Identifies which inputs most influence predictions. Example: Age, BMI, and activity level emerge as top variables in AI risk model. Practical application: Focusing coaching efforts on modifiable high-impact factors. Challenges: Importance can be model-specific;

correlated variables may obscure true effects.

Variance Inflation Factor (VIF) – Concept: Diagnostic for multicollinearity among predictors. Related terms: collinearity, regression diagnostics. Explanation: VIF > 10 indicates problematic redundancy. Example: VIF = 12 for cholesterol and LDL in AI risk equation. Practical application: Refining model inputs to improve stability. Challenges: Removing variables may reduce predictive power; requires domain expertise.

Weighted Accuracy – Concept: Accuracy that accounts for class importance. Related terms: cost-sensitive learning, balanced accuracy. Explanation: Assigns higher weight to minority or high-risk class errors. Example: Weighting false negatives twice as heavily yields weighted accuracy of 88%. Practical application: Prioritizing detection of life-threatening conditions. Challenges: Determining appropriate weights; may overfit to weighted class.

Yield – Concept: Proportion of screened individuals who receive a positive outcome (e.G., Diagnosis). Related terms: screening efficiency, positive predictive value. Explanation: High yield indicates effective targeting. Example: AI triage yields 25% diagnosed hypertension among those screened, versus 10% with random screening. Practical application: Optimizing resource allocation for community health drives. Challenges: Yield can be inflated by selecting high-prevalence subpopulations.

Z-Score – Concept: Standardized score indicating how many standard deviations an observation is from the mean. Related terms: normalization, statistical significance. Explanation: Z = (value – mean)/SD. Example: AI-predicted stress level of 75 yields Z = 1.5, Indicating above-average stress. Practical application: Flagging outliers for targeted coaching. Challenges: Assumes normal distribution; may misrepresent skewed health data.