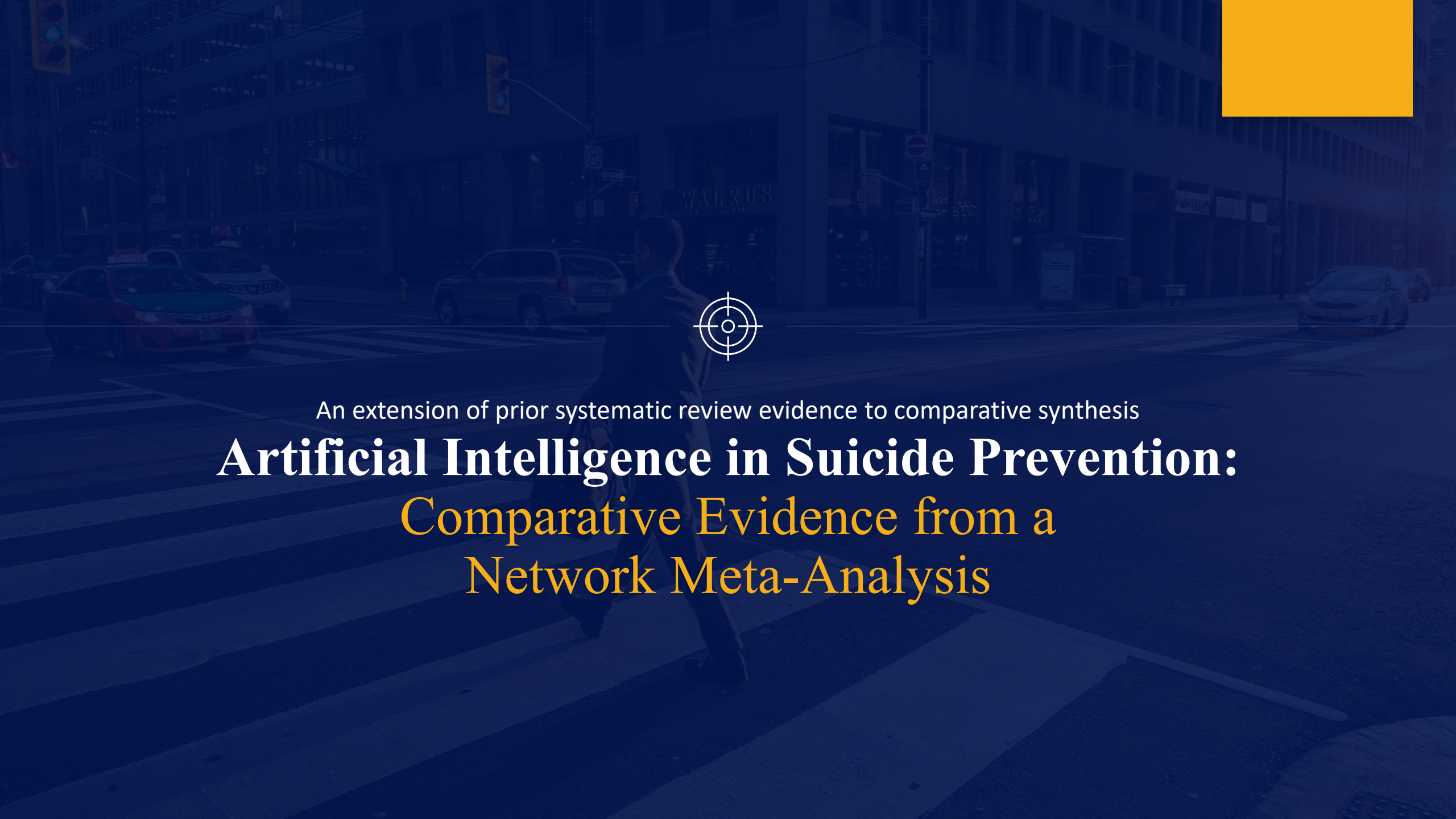
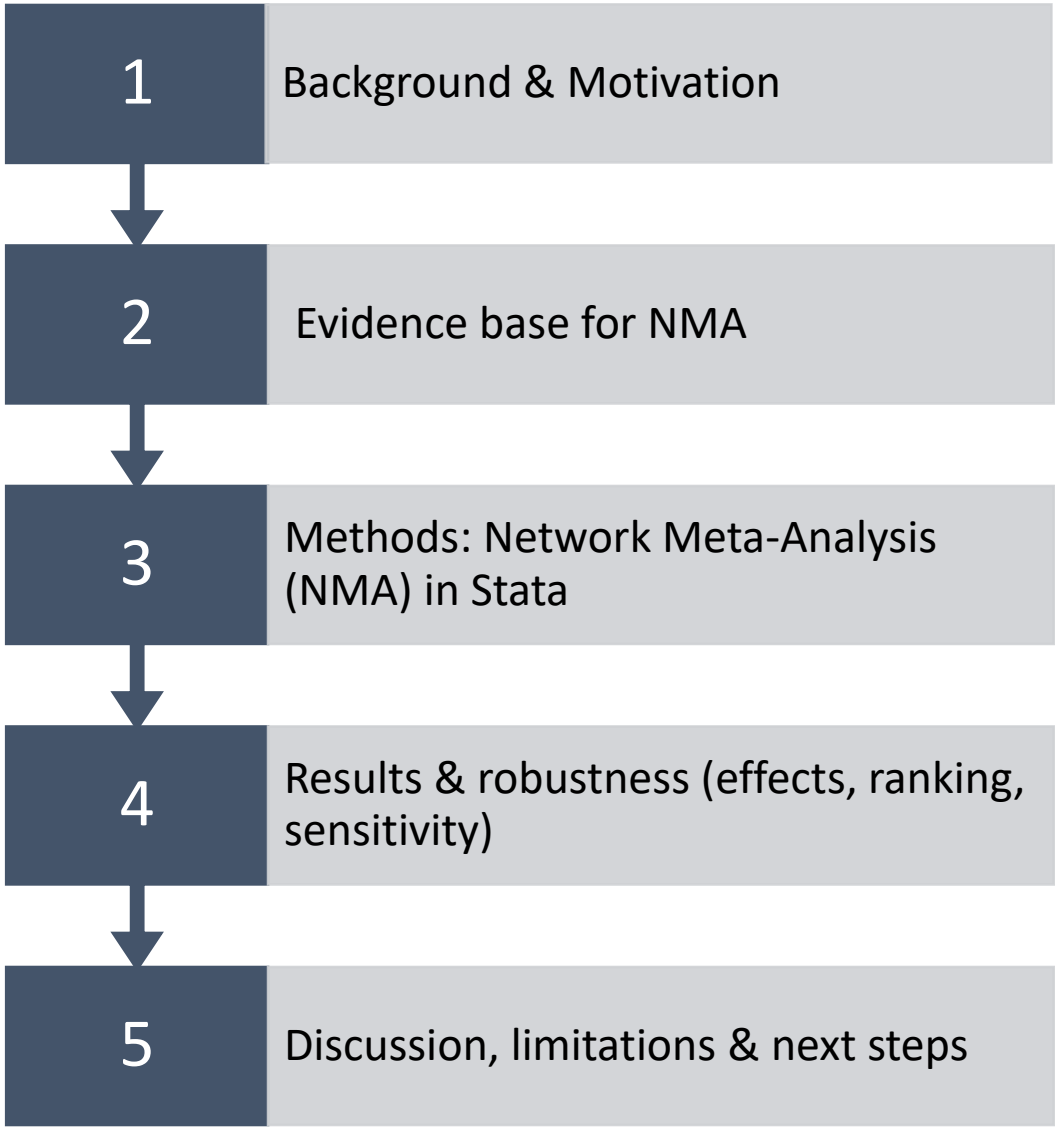


An extension of prior systematic review evidence to comparative synthesis

Artificial Intelligence in Suicide Prevention: Comparative Evidence from a Network Meta-Analysis

An extension of prior systematic review evidence to comparative synthesis

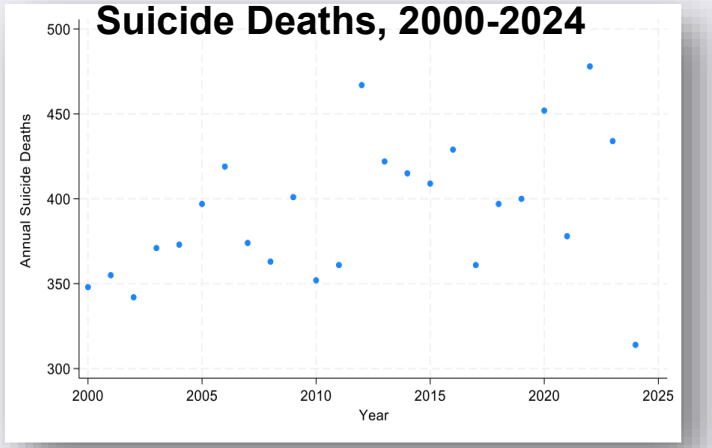
Contents

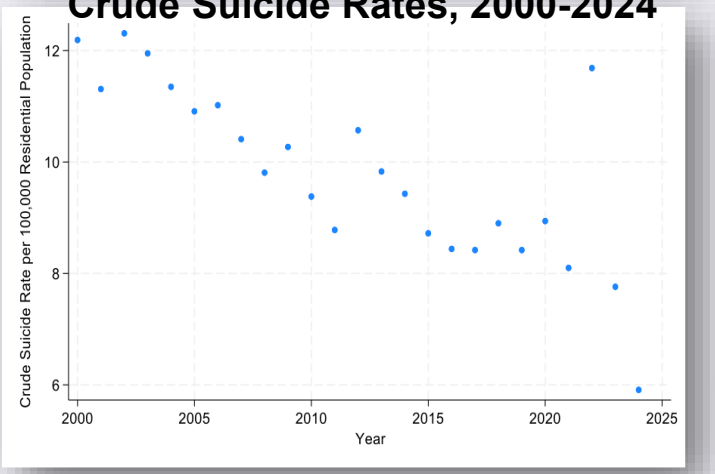


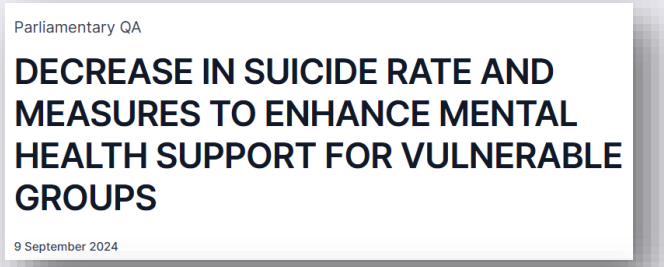
Background: Suicide Risk and the Case for Prediction Models

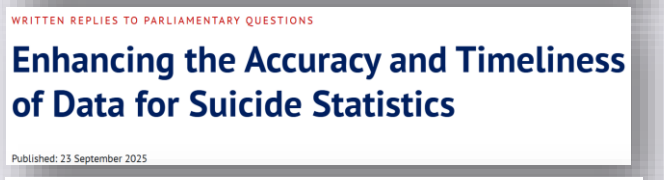
Singapore: persistent suicide burden → motivation for AI-enabled early risk identification

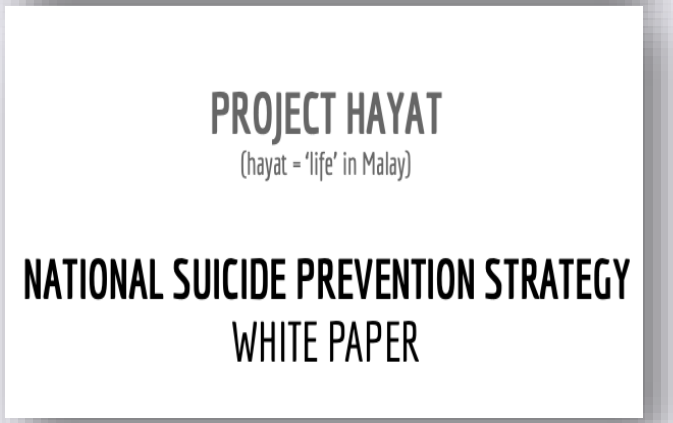
AI focus: Evidence-based benchmarking of suicide-risk prediction models to support Singapore's Smart prevention workflows.











Sources: WHO GHO (accessed 26 Sep 2025); MOH (2024); MHA (2025); Project Hayat (2024).

Falling suicide rates in Singapore, but not the actual numbers

Motivation: Why Comparative Evidence Across Algorithms Matters

Challenging problem

Persistent burden motivates earlier, data-driven identification of individuals at elevated risk.

Many model options

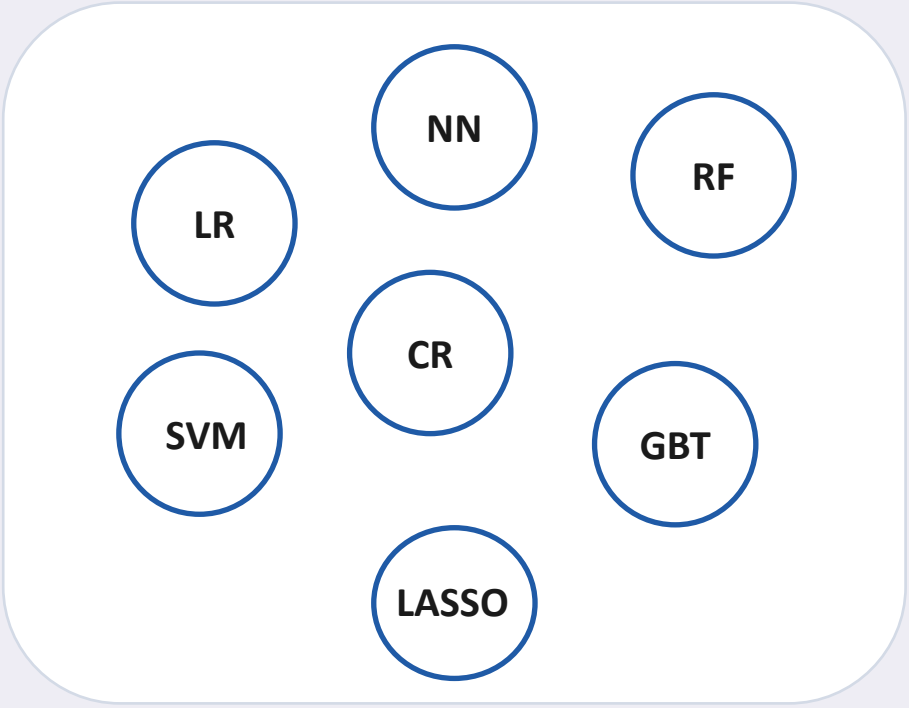
LR, RF, GBT, SVM, NN, LASSO, CR etc. are widely used in EHR and administrative data.

Sparse comparisons

Few head-to-head studies; most models are compared only against a single baseline.

Evidence-based selection

We benchmark algorithms using AUC-based NMA before local validation and integration into SMART workflows.



From Systematic Review to Comparative Synthesis (NMA)

Artificial intelligence and suicide prevention: A systematic review

Alban Lejeune[1*], Aziliz Le Glaz[1], Pierre-Antoine Perron[1], Johan Sebti[2], Enrique Baca-Garcia[3], Michel Walter[1,4], Christophe Lemey[1,4,5] and Sofian Berrouiguet[1,6]

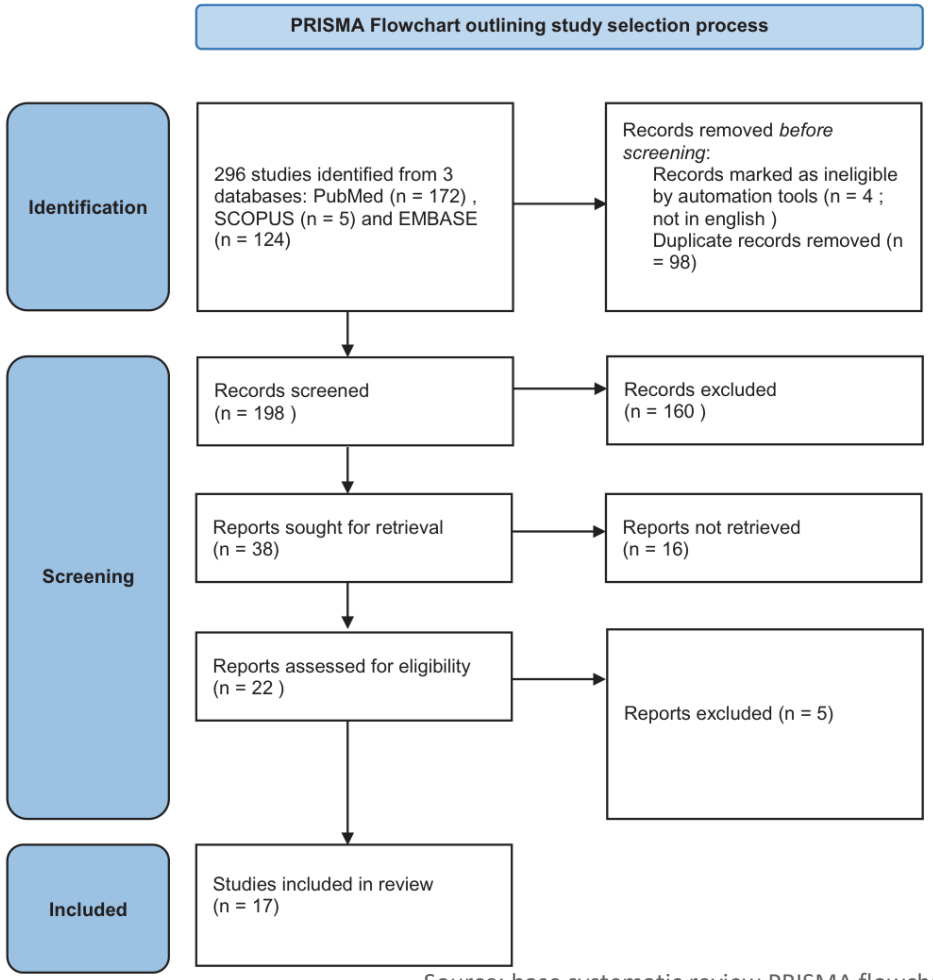
[1]URCI Mental Health Department, Brest Medical University Hospital, Brest, France; [2]Mental Health Department, French Polynesia Hospital, FFC3+H9G, Pirae, French Polynesia; [3]Departamento de Psiquiatria, IIS-Fundación Jiménez Díaz, Madrid, Spain; [4]EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France; [5]SPURBO, IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Brest, France and [6]LaTIM, INSERM, UMR 1101, Brest, France

- **Baseline evidence (Lejeune et al., 2022):** Synthesises AI/ML approaches for suicide prevention.
- **Key limitation:** Substantial heterogeneity across studies (populations, predictors, settings, validation strategies) limits direct comparability.
- **Comparative gap:** Sparse head-to-head comparisons prevent a coherent, unified cross-algorithm ranking and pooled relative-effect estimates.

Our contribution: *We translate the review into comparative evidence by extracting harmonised AUC inputs and conducting a reproducible network meta-analysis (NMA) in Stata.*

Evidence Identification and Study Selection (PRISMA)

- **Search & screening:** 296 records identified → **17 studies included** in the systematic review.
- **Review → NMA subset:** Only **10 studies reported extractable AUC** suitable for harmonisation and inclusion in an AUC-based NMA.
- **Implication:** Even within AUC-reporting studies, head-to-head evidence is sparse and the network is imbalanced across algorithms; a structured synthesis (NMA) is needed for cross-algorithm comparison.



Source: base systematic review PRISMA flowchart.

Figure 1. PRISMA flowchart outlining the study selection process.

Our approach: *Retain the review scope, extract harmonised AUC inputs, and apply NMA to jointly compare multiple algorithms by combining all available direct and indirect comparisons, yielding relative performance estimates (ΔAUC vs LR) with 95% confidence intervals (CIs).*

Included Studies and Algorithms (AUC Evidence Base)

Included studies for NMA (n = 10; same as base paper)

Base paper Table 1 (AUC by algorithm; 10 integrated studies)

Table 1. Performance in the prediction of suicide risk with the main algorithms, expressed in AUC, in studies in which this value was informed.

	NN	LR	LASSO	XGB/GBT	CR	SVM	RF	Cross validation
Sanderson et al. (2019) [22]	0.842	0.818		0.849				Yes
Sanderson et al. (2019) [24]	0.8352	0.8179						Yes
Sanderson et al. (2020) [23]		0.86		0.88				Yes
Zheng et al. (2020) [28]	0.769	0.604		0.702				Yes
Choi et al. (2018) [25]	0.683				0.688	0.687		Yes
Simon et al. (2019) [29]			0.85					Yes
Walsh et al. (2018) [27]		0.7					0.9	No
Ryu et al. (2019) [30]							0.947	Yes
Gradus et al. (2019) [31]							080–0.88	Yes
Miché et al. (2019) [32]		0.828	0.826				0.824	Yes

Abbreviations: AUC, area under the curve; BN, Bayesian network; CR, cox regression; LR, logistic regression; NN, neural network; RF, random forest; SVM, support vector machine; XGB/GBT, extreme gradient boosting/gradient boosted tree.

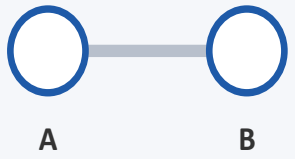
1. Sanderson et al. (2019) [22] — Administrative health system data; NN / GBT etc.
2. Sanderson et al. (2019) [24] — Feedforward NN vs LR (administrative data).
3. Sanderson et al. (2020) [23] — Post-ED parasuicide visit prediction (ML + admin data).
4. Choi et al. (2018) [25] — 10-year prediction using Cox regression + ML (South Korea).
5. Walsh et al. (2018) [27] — Adolescent suicide attempts (longitudinal clinical data + ML).
6. Zheng et al. (2020) [28] — Deep learning + EHR early warning for high-risk patients.
7. Simon et al. (2019) [29] — Required EHR data for accurate prediction (LR/LASSO/ML).
8. Ryu et al. (2019) [30] — Suicide prediction with machine learning (Korea).
9. Gradus et al. (2019) [31] — Sex-specific suicide risk (Denmark registries; ML).
10. Miché et al. (2019) [32] — Prospective prediction in community adolescents/young adults.

- Common comparator: Logistic Regression (LR) as the reference to connect algorithms.
- Algorithms compared: NN, LASSO, XGB/GBT, CR, SVM, RF.
- Outcome metric: AUC as the harmonized discrimination measure across studies.

Abbreviations: LR = logistic regression; NN = neural network; LASSO = least absolute shrinkage and selection operator; GBT = gradient boosting trees; XGB = extreme gradient boosting; RF = random forest; SVM = support vector machine; CR = Cox regression.

Why Network Meta-Analysis (NMA)?

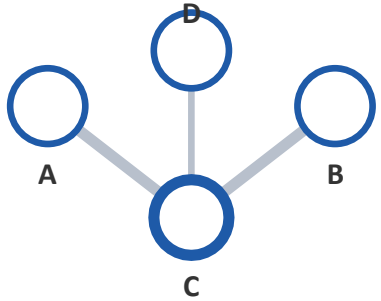
Pairwise MA



One comparison at a time

- ✓ Pairwise meta-analysis compares one pair at a time (A vs B).
- ✓ In practice, many models exist but head-to-head evidence is sparse.

Network MA (NMA)



Direct + indirect evidence

- ➢ NMA compares multiple models simultaneously using a connected evidence network.
- ➢ Network meta-analysis **combines direct and indirect evidence** to enable a **joint comparison of multiple algorithms** on a common performance scale.

The Stata Journal (2015)
15, Number 4, pp. 951–985

Network meta-analysis

Ian R. White
MRC Biostatistics Unit
Cambridge Institute of Public Health
Cambridge, UK
ian.white@mrc-bsu.cam.ac.uk

Abstract. Network meta-analysis is a popular way to combine results from several studies (usually randomized trials) comparing several treatments or interventions. It has usually been performed in a Bayesian setting, but recently it has become possible in a frequentist setting using multivariate meta-analysis and meta-regression, implemented in Stata with `mvmeta`. I describe a suite of Stata programs for network meta-analysis that perform the necessary data manipulation, fit consistency and inconsistency models using `mvmeta`, and produce various graphics.

Keywords: st0410, network components, network convert, network forest, network import, network map, network meta, network pattern, network query, network rank, network setup, network sidesplit, network table, network unset, network meta-analysis, multiple treatments meta-analysis, mixed-treatment comparisons

> Network meta-analysis (NMA), also termed multiple treatment meta-analysis or mixed treatment comparisons, was developed as an extension of pairwise meta-analysis to allow comparisons of more than two interventions in a single coherent analysis of all the relevant studies.

Key Assumptions and Diagnostics for NMA

Three core assumptions

- **Similarity**
  Studies comparing different model pairs are sufficiently comparable.
- **Transitivity**
  No systematic differences in effect modifiers across comparisons.
- **Consistency**
  Direct and indirect evidence agree within the network.

Issues we explicitly evaluate

Treatment contrasts	network table / forest
Heterogeneity	random-effects meta model
Inconsistency	network sidesplit (node-splitting)
Within-study error	SE(AUC) + sensitivity checks

In Stata: network meta consistency + network sidesplit all + network map + forest

Methods Summary: NMA Implementation in Stata

Evidence

- 10 studies reported AUCs
- Models: LR (ref), NN, RF, GBT/XGB, SVM, LASSO, CR

01

Outcome

- Discrimination: AUC
- Effect size: ΔAUC vs LR

02

Uncertainty handling

- SE(AUC) reconstructed via Hanley–McNeil
- Random-effects multivariate model with between-study correlation
- Sensitivity: effective sample-size assumptions for cross-validated AUCs

03

Model & checks

- Consistency NMA (REML) in Stata (mvmeta/network)
- Diagnostics: heterogeneity (τ) and node-splitting (direct vs indirect)
- Outputs: ranking (SUCRA / PrBest)

04

Reconstructing SE(AUC) When Not Reported (Hanley–McNeil)

Most included papers reported AUC without a 95% CI. Network meta-analysis requires an SE for each study–algorithm AUC estimate.

We performed the network meta-analysis on the original AUC scale (primary analysis). Where 95% CIs were missing we reconstructed SE(AUC) using Hanley–McNeil (1982) based on AUC and case/control counts (n1/n0). Sample-size sensitivity analyses were run to check robustness.

Hanley–McNeil (1982): SE(AUC)

$$Q1 = A / (2 - A)$$

$$Q2 = 2A^2 / (1 + A)$$

$$Var(A) = \{ A(1-A) + (n1-1)(Q1-A^2) + (n0-1)(Q2-A^2) \} / (n1 \cdot n0)$$

$$SE(A) = \sqrt{Var(A)}$$

A = AUC; n1 = cases (positives); n0 = controls (negatives).

Source: Hanley JA & McNeil BJ (1982) Radiology. Used to derive SE(AUC) when CI is not reported.

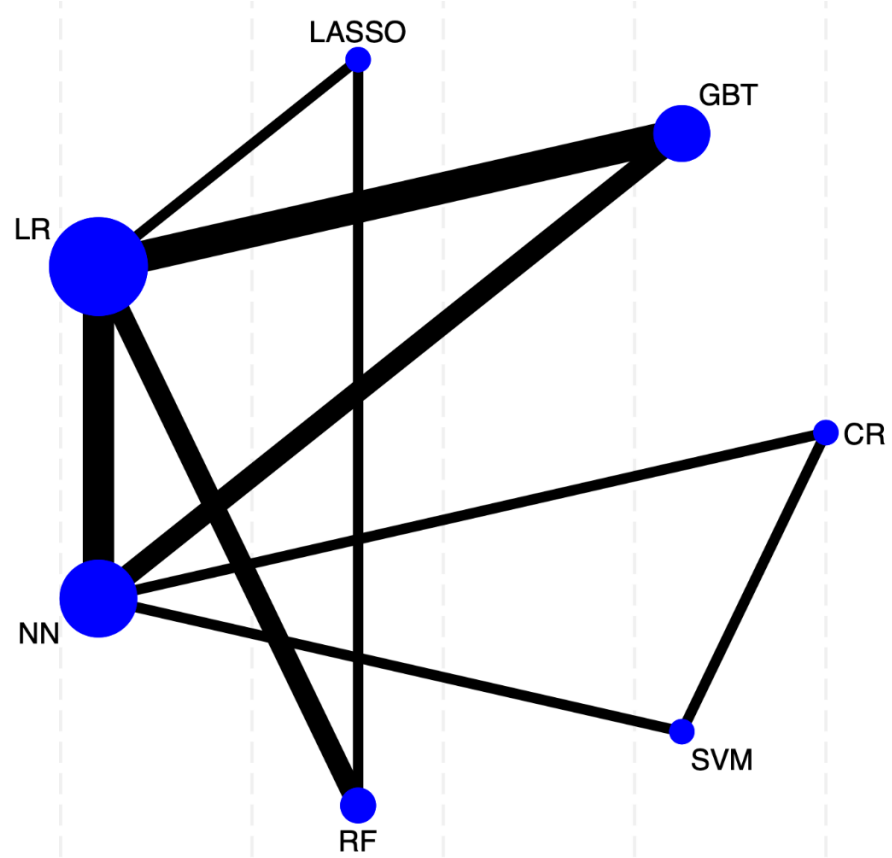
Reconstructing SE(AUC) When Not Reported (Hanley–McNeil)

	study	algorithm	auc	n1	n0	Q1	Q2	var_auc	se_auc
1	Sanderson_2019_22	NN	.8419	3548	35480	.72696657	.76963527	.00001756	.0042024
2	Sanderson_2019_22	GBT	.8493	3548	35480	.73807252	.78009033	.00001704	.00412794
3	Sanderson_2019_22	LR	.8179	3548	35480	.69190424	.73597053	.00001953	.00441971
4	Sanderson_2019_24	NN	.8352	3548	35480	.71703298	.76019949	.0000182	.00426667
5	Sanderson_2019_24	LR	.8179	3548	35480	.69190424	.73597053	.00001953	.00441971
6	Sanderson_2020_23	LR	.8632	268	33426	.75932443	.79982208	.00020456	.0143026
7	Sanderson_2020_23	GBT	.8786	268	33426	.78348493	.82182259	.00018649	.01365603
8	Zheng_2020_28	NN	.769	203	117892	.62469536	.66858224	.00038069	.01951117
9	Zheng_2020_28	LR	.604	203	117892	.43266473	.45488276	.00044426	.02107742
10	Zheng_2020_28	GBT	.702	203	117892	.54083208	.57908816	.00042546	.02062657
11	Choi_2018_25	NN	.683	764	245222	.51860292	.55435417	.00011522	.01073404
12	Choi_2018_25	CR	.688	764	245222	.52439027	.56083415	.00011472	.01071095
13	Choi_2018_25	SVM	.687	764	245222	.52322922	.55953642	.00011483	.01071571
14	Walsh_2018_27	LR	.7	974	7059	.53846152	.57647057	.00009565	.00978034
15	Walsh_2018_27	RF	.9	974	7059	.81818178	.85263154	.00004493	.00670331
16	Ryu_2019_30	RF	.947	331	5442	.89933528	.92122141	.00007423	.00861571
17	Gradus_2019_31	RF	.8406	14103	265183	.72503021	.76780221	4.409e-06	.00209965
18	Miche_2019_32	LR	.828	137	2656	.70648465	.75009192	.00047839	.02188348
19	Miche_2019_32	RF	.824	137	2656	.70068027	.74449123	.00048654	.02205774
20	Miche_2019_32	LASSO	.826	137	2656	.70357748	.74729022	.00048273	.02197122
21	Simon_2019_29	LASSO	.85	63805	10212048	.73913047	.78108111	9.198e-07	.00095904

```
*************************************************
* 2) Hanley–McNeil SE(AUC) on logit scale
*************************************************
gen double Q1 = auc/(2-auc)
gen double Q2 = 2*auc^2/(1+auc)

gen double var_auc = ( auc*(1-auc) ///
    + (n1-1)*(Q1 - auc^2) ///
    + (n0-1)*(Q2 - auc^2) ) / (n1*n0)

gen double se_auc    = sqrt(var_auc)
```

Results: Evidence Network (Network Map)

```
network sidesplit all
network map
network forest

restore
```



Network plot from Stata: network map.

- Node size reflects how frequently each algorithm is evaluated across the included studies; edge thickness reflects the amount of direct head-to-head evidence for a given pair.
- The network is anchored by LR and NN. Direct evidence is densest for LR–NN–GBT–RF, whereas LASSO/SVM/CR are supported by fewer direct comparisons, implying sparser evidence and wider uncertainty for those nodes.

Results: Network Table (Study × Algorithm Inputs)

```
90 . network table
```

study	Treatment and Statistic: LR auc	LR se_auc	LR n0	CR auc	CR se_auc	CR n0
Choi_2018_25				.6880000234	.0107109493	245222
Miche_2019_32	.8280000091	.021883475	2656			
Sanderson_2019_22	.817900002	.0044197135	35480			
Sanderson_2019_24	.817900002	.0044197135	35480			
Sanderson_2020_23	.8632000089	.0143025975	33426			
Walsh_2018_27	.6999999881	.0097803359	7059			
Zheng_2020_28	.6039999723	.0210774214	117892			

study	Treatment and Statistic: GBT auc	GBT se_auc	GBT n0	LASSO auc	LASSO se_auc	LASSO n0
Choi_2018_25						
Miche_2019_32				.8259999752	.0219712209	2656
Sanderson_2019_22	.8493000269	.0041279382	35480			
Sanderson_2019_24						
Sanderson_2020_23	.8786000013	.0136560309	33426			
Walsh_2018_27						
Zheng_2020_28	.7020000219	.0206265725	117892			

study	Treatment and Statistic: NN auc	NN se_auc	NN n0	RF auc	RF se_auc	RF n0
Choi_2018_25	.6830000281	.010734041	245222			
Miche_2019_32				.824000001	.0220577422	2656
Sanderson_2019_22	.841899991	.0042024008	35480			
Sanderson_2019_24	.8352000117	.0042666674	35480			
Sanderson_2020_23						
Walsh_2018_27				.8999999762	.0067033095	7059
Zheng_2020_28	.7689999938	.0195111714	117892			

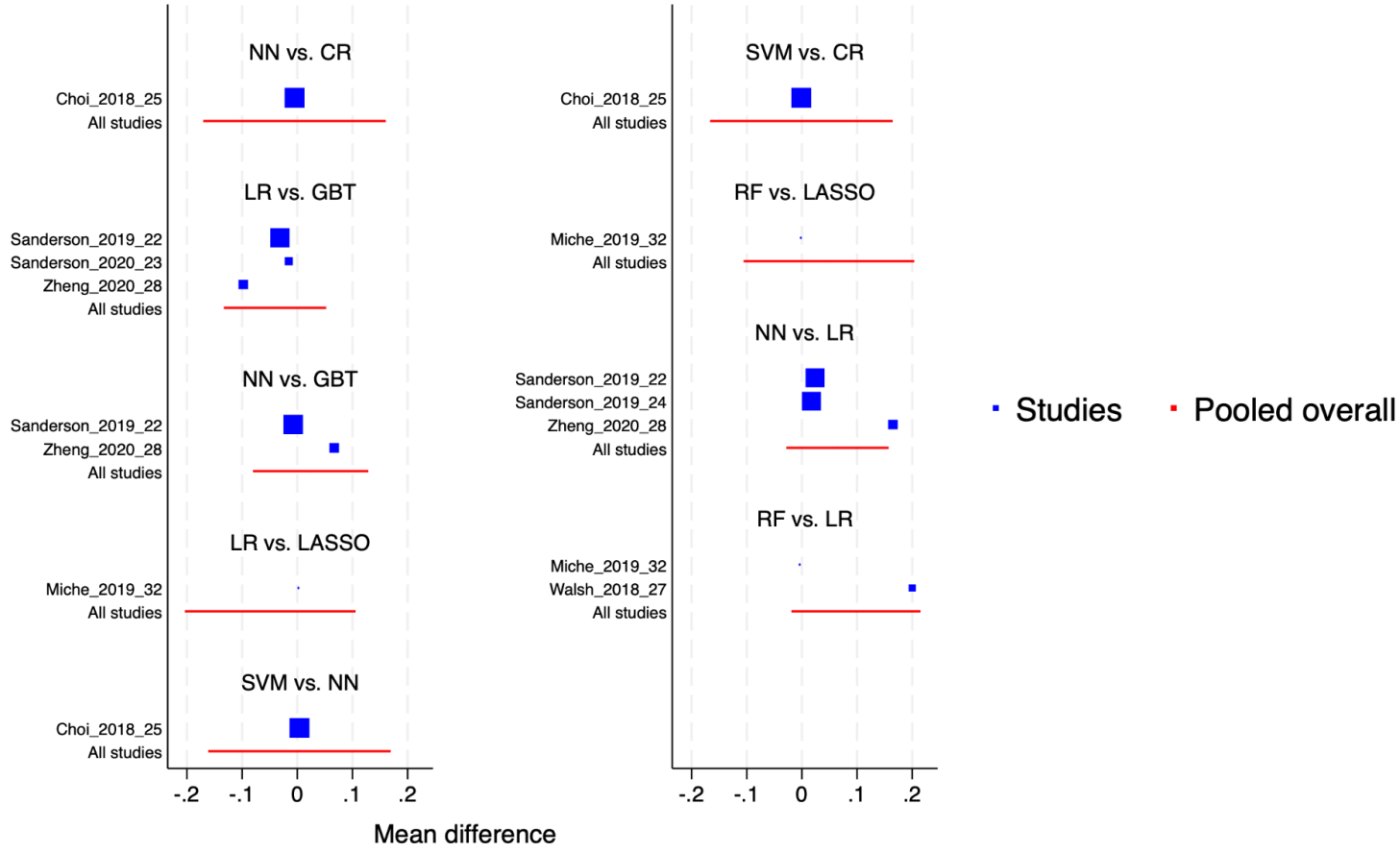
study	Treatment and Statistic: SVM auc	SVM se_auc	SVM n0
Choi_2018_25	.6869999766	.0107157143	245222
Miche_2019_32			
Sanderson_2019_22			
Sanderson_2019_24			
Sanderson_2020_23			
Walsh_2018_27			
Zheng_2020_28			

- ***Inputs:*** *Study × model AUC and within-study SE(AUC).*
- ***Network contribution:*** *10 studies reported AUCs; 7 had sufficient information to obtain comparable SE(AUC) and entered the primary NMA network.*
- ***Implication:*** *Network is connected (anchored by LR/NN), but sample sizes vary and some models are supported by few studies.*

Results: Primary Pooled Contrasts (Forest Plot)

```
network sidesplit all
network map
network forest

restore
```

- **Effect measure:** Pooled mean difference in AUC (ΔAUC) from a **random-effects consistency NMA**.
- **Main finding: All pooled contrasts were small and 95% CIs included 0**, indicating **no statistically detectable improvement** over LR.
- **Interpretation:** RF and NN showed the largest positive point estimates, but uncertainty remained substantial—especially for algorithms supported by sparse direct evidence.



Forest plot from Stata: network forest.

Reading guide: blue squares = study-level estimates; red line = pooled overall effect for each contrast (mean difference in AUC).

Results: Consistency Model Estimates (vs LR)

- None of the alternative algorithms demonstrated a statistically significant improvement over LR, as all 95% confidence intervals for ΔAUC crossed 0.
- The largest point estimate was observed for RF versus LR (mean ΔAUC ≈ +0.098; p = 0.101).
- Uncertainty remained substantial, evidenced by wide confidence intervals, which likely reflects sparse head-to-head evidence and between-study heterogeneity under the random-effects framework.

```
11 . network meta consistency
   Command is: mvmeta _y _S  , bscovariance(exch 0.5) longparm suppress(uv mm)  vars(_y_A _y_B _y_C _
   > y_E _y_F _y_G)
   Note: using method reml
   Note: using variables _y_A _y_B _y_C _y_E _y_F _y_G
   Note: 7 observations on 6 variables
   Note: variance-covariance matrix is proportional to .5*I(6)+.5*J(6,6,1)

   Initial:      Log likelihood = -8.0938716
   Rescale:      Log likelihood =  .40926558
   Rescale eq:   Log likelihood =   1.113312
   Iteration 0:  Log likelihood =   1.113312
   Iteration 1:  Log likelihood =  1.5237388  (not concave)
   Iteration 2:  Log likelihood =  1.8190252  (not concave)
   Iteration 3:  Log likelihood =  1.8324281
   Iteration 4:  Log likelihood =  1.8366224
   Iteration 5:  Log likelihood =  1.8368536
   Iteration 6:  Log likelihood =  1.8368536

   Multivariate meta-analysis
   Variance-covariance matrix = proportional .5*I(6)+.5*J(6,6,1)
   Method = reml                                   Number of dimensions    =     6
   Restricted log likelihood = 1.8368536           Number of observations  =     7
```

		Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_y_A	_cons	.0695687	.0967682	0.72	0.472	-.1200934	.2592309
_y_B	_cons	.0403781	.0473274	0.85	0.394	-.0523819	.133138
_y_C	_cons	.0490012	.0789714	0.62	0.535	-.1057799	.2037823
_y_E	_cons	.0645897	.0473234	1.36	0.172	-.0281624	.1573418
_y_F	_cons	.0980024	.0596962	1.64	0.101	-.0189999	.2150048
_y_G	_cons	.0685687	.0967682	0.71	0.479	-.1210935	.2582308

Table excerpt: network meta consistency (REML).

Consistency Check: Node-Splitting (Direct vs Indirect Evidence)

- Node-splitting compares direct versus indirect evidence for each contrast.
- Most contrasts show no evidence of inconsistency (p>0.05).However, LASSO–LR and LASSO–RF show inconsistency (p=0.018), possibly reflecting heterogeneity and/or sparse/imbalanced evidence; ranking claims involving LASSO should be interpreted cautiously.
- Overall consistency was generally acceptable, with inconsistency limited to LASSO-related contrasts.

```
12 . network sidesplit all

Side    Direct                Indirect              Difference
        Coef.      Std. Err.  Coef.      Std. Err.  Coef.      Std. Err.  P>|z|
D E *    .0687667   .0537656  -.0040955   .2179454   .0728622   .2244792  0.745
D F             .          .          .          .          .          .      .
A E *      -.005   .0844289   .1302056   6.782048  -.1352056   6.782574  0.984
A G             .          .          .          .          .          .      .
B D *   -.0482667   .0519349   .0897285   .2109066  -.1379952   .2172069  0.525
B E         .0298    .066444   .0018567   .1328625   .0279433   .1485505  0.851
C D *        .002   .0608973      -.406   .1611149       .408   .1722396  0.018
C F *      -.002    .0608973       .406   .1611149      -.408   .1722397  0.018
E G *    .0039999   .0844289  -.1312056   6.781368   .1352056   6.781893  0.984

* Warning: all the evidence about these contrasts comes from the trials which directly compare the
> m.
See help file for more information.
```

Table excerpt: network sidesplit all.

Model uncertainty: heterogeneity (τ) & between-study correlation (ρ)

- **Between-study heterogeneity (τ ≈ 0.084):** Random-effects SD of treatment effects across studies.
- **Between-study correlation (ρ = 0.5; exchangeable structure):** Specified common correlation among study-specific random effects in the multivariate random-effects model (*bscovariance(exch 0.5)*).
- **Interpretation:** With τ>0, uncertainty increases; interpret pooled effects and rankings cautiously, especially with sparse/imbalanced networks.

Estimated between-studies SDs and correlation matrix

	SD	_y_A	_y_B	_y_C	_y_E	_y_F	_y_G
_y_A	.08442197	1	.	.	.	.	.
_y_B	.08442197	.5	1	.	.	.	.
_y_C	.08442197	.5	.5	1	.	.	.
_y_E	.08442197	.5	.5	.5	1	.	.
_y_F	.08442197	.5	.5	.5	.5	1	.
_y_G	.08442197	.5	.5	.5	.5	.5	1

mvmeta command stored as F9

Stata output: estimated between-study SDs and correlation matrix for multivariate random effects.

Ranking Metrics: PrBest and SUCRA

```
network rank max, all gen(prob) reps(10000) seed(20260128)
sucra prob*

ds prob*
local plist `r(varlist)'

capture drop p_top1 p_top2 p_top3 p_top4
gen double p_top1 = 0
gen double p_top2 = 0
gen double p_top3 = 0
gen double p_top4 = 0

local i = 0
foreach v of local plist {
    local ++i
    replace p_top1 = p_top1 + `v' if `i'<=1
    replace p_top2 = p_top2 + `v' if `i'<=2
    replace p_top3 = p_top3 + `v' if `i'<=3
    replace p_top4 = p_top4 + `v' if `i'<=4
}
```

Estimated probabilities (%) of each treatment having each rank
- assuming the **maximum** parameter is the best
- using **10000** draws
- allowing for parameter uncertainty

	Treatment						
Rank	LR	CR	GBT	LASSO	NN	RF	SVM
Best	0.0	21.4	4.7	13.6	6.9	32.8	20.5
2nd	0.5	17.7	10.0	15.9	16.2	21.8	17.9
3rd	2.7	13.0	15.6	12.5	27.1	15.6	13.5
4th	8.0	12.2	19.1	11.5	24.3	12.9	12.1
5th	18.3	10.2	21.9	13.1	16.6	9.6	10.3
6th	32.0	11.1	19.1	13.7	7.2	5.1	11.9
Worst	38.6	14.4	9.7	19.7	1.7	2.2	13.8

Treatment Relative Ranking of Model 1

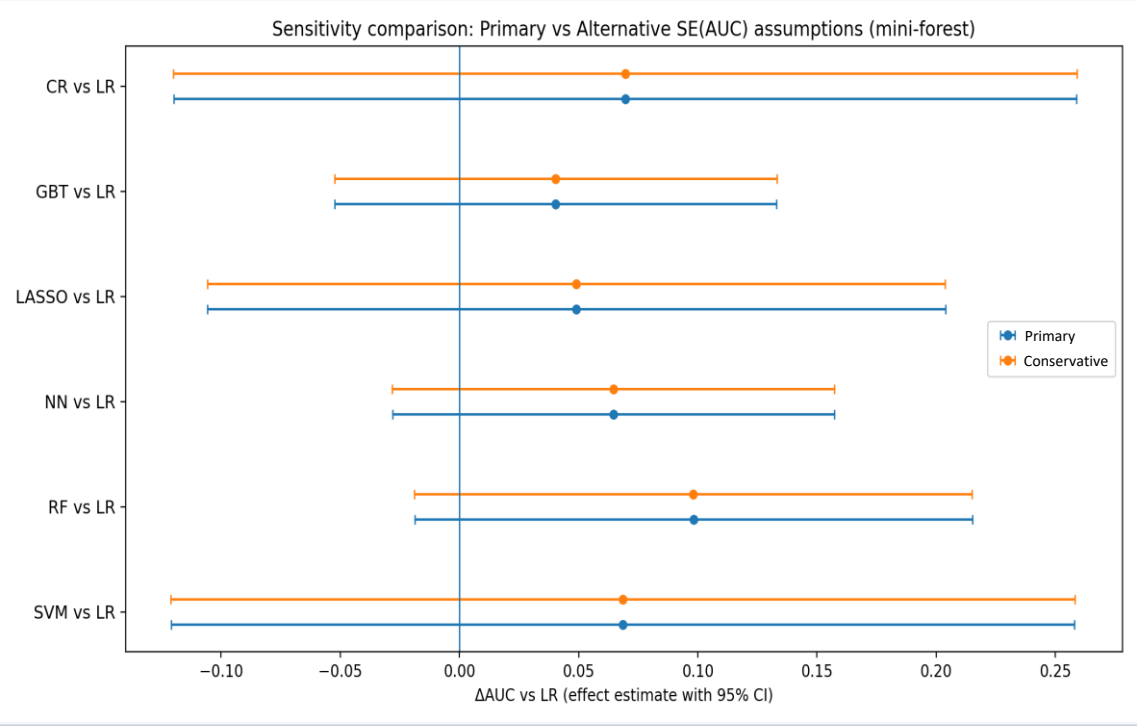
Treatm~t	SUCRA	PrBest	MeanRank
1	17.6	0.0	5.9
2	56.2	21.4	3.6
3	43.4	4.7	4.4
4	47.6	13.6	4.1
5	57.4	6.9	3.6
6	71.9	32.8	2.7
7	55.9	20.5	3.6

- **RF is most likely to be best (PrBest ≈ 32.8%; SUCRA ≈ 71.9), but uncertainty remains.**
- **Consistent with the NMA effect estimates, no model shows a statistically significant improvement over LR (95% CIs for ΔAUC overlap 0).**
- **Rankings are probabilistic—interpret alongside ΔAUC, CIs, and network/consistency checks.**

Note: Rankings are probabilistic; do not infer superiority without ΔAUC and 95% CIs.

Sensitivity Analysis: Effective Sample Size Under Cross-Validation

Sensitivity set-up: For AUCs obtained via internal validation, SE(AUC) depends on the effective test sample size. Because reporting of validation procedures varied, we treated the reported total N as an upper bound and re-derived SE(AUC) under conservative effective-N scenarios (80/50/20% of N; per-fold test N when k was reported), then re-ran the same NMA.



Study	Validation / evaluation approach	k (if CV)
Sanderson et al. (2019) [22]	k-fold cross-validation	10
Sanderson et al. (2019) [24]	k-fold cross-validation	10
Sanderson et al. (2020) [23]	k-fold cross-validation	10
Choi et al. (2018) [25]	Train/validation split + CV used for model tuning	10 (tuning)
Walsh et al. (2018) [27]	Bootstrap internal validation (optimism-correction)	N/A
Zheng et al. (2020) [28]	Separate validation cohort (hold-out / temporal validation); CV mentioned but k not reported	N/A / NR
Simon et al. (2019) [29]	k-fold cross-validation	10
Ryu et al. (2019) [30]	Train/test split; feature selection via k-fold CV within training set	10 (within training)
Gradus et al. (2019) [31]	k-fold cross-validation	10
Miché et al. (2019) [32]	Nested repeated k-fold cross-validation	10 (nested/repeated)

Robustness: Conclusions were unchanged under conservative assumptions (similar point estimates; slightly wider CIs, still including 0).

Discussion & conclusion

I. **Evidence base:** 10 studies reported AUCs across 6 ML models + LR, forming a network anchored by LR/NN. Evidence was strongest for LR–NN–GBT–RF, with sparser links for LASSO/SVM/CR.

II. **Main findings:** Pooled **ΔAUC vs LR** were small and **95% CIs included 0**, indicating **no statistically clear superiority** of any ML model over LR. Ranking suggested RF as most likely best (**PrBest ≈ 32.9%; SUCRA ≈ 72**), but ranking is **probabilistic** and not a decisive winner.

III. **Interpretation:** NMA synthesises direct and indirect evidence to support cross-model comparison; conclusions should be interpreted cautiously given **uncertainty and sparse evidence**.

AUC discrimination is broadly similar across models; no ML model shows robust superiority over LR.

Limitations & Next steps

Limitations

I. **Sparse/imbalanced evidence:** Several contrasts were informed by few studies → wide uncertainty.

II. **Heterogeneity & comparability:** Differences in populations, predictors, horizons, and validation may affect transitivity.

III. **Incomplete reporting:** SE/95% CI and CV details (k, effective n1/n0) were often missing, requiring reconstruction and sensitivity checks.
IV. **AUC-only:** Calibration and clinical utility were rarely reported.

Next steps

I. **Transparent reporting standards:** Standardised reporting of AUC uncertainty (SE/95% CI), n1/n0, and validation specifications (CV folds and how AUCs were aggregated) would improve cross-study comparability.

II. **Stronger comparisons:** More head-to-head, multi-model evaluations on shared datasets/pipelines.

III. **Translation:** External validation and implementation studies (plus calibration/decision-utility, not AUC alone).

References

A. Background & methods

Lejeune A, Berrouiguet S, et al. *Artificial intelligence for suicide prevention: a systematic review.* European Psychiatry. 2022.

World Health Organization. *Suicide worldwide in 2019: Global health estimates.* Geneva: WHO; 2021.

Hanley JA, McNeil BJ. *The meaning and use of the area under a receiver operating characteristic (ROC) curve.*Radiology. 1982;143(1):29–36.

White IR. *mvmeta: multivariate meta-analysis and meta-regression* (Stata module; mvmeta ado file). 2022.

White IR. *network: network meta-analysis* (Stata module; network.ado). 2018 / 2024 (see Stata log for installed version).

StataCorp LLC. *Stata Statistical Software: Release 19.5.* College Station, TX: StataCorp LLC.

B. Included studies used in the NMA (short list)

1. Sanderson et al., 2019 (administrative data; NN / GBT etc.).
2. Sanderson et al., 2019 (feedforward NN vs LR).
3. Sanderson et al., 2020 (post-ED parasuicide prediction).
4. Choi et al., 2018 (Korea; long-term prediction).
5. Walsh et al., 2018 (adolescent longitudinal cohort).
6. Zheng et al., 2020 (deep learning; EHR).
7. Simon et al., 2019 (EHR prediction; LR/LASSO/ML).
8. Ryu et al., 2019 (Korea; ML prediction).
9. Gradus et al., 2019 (Denmark registries; sex-specific prediction).
10. Miché et al., 2019 (prospective adolescent cohort).

Thank You.