Comparative Implementation of a Brief App-directed Delirium Identification Protocol by Hospitalists, Nurses, and Nursing Assistants

On-line Supplement

Technical Appendix

Methods for Supplement Tables

Supplement Tables:

Technical Appendix

In this technical Appendix, we describe details of the statistical analysis used to generate the data presented in the tables and associated text in the paper. We also present the statistical software and procedures used to conduct the analyses.

Most of the analyses presented (Tables 1, 2, 3 and associated text) involve descriptive statistics. We report mean and standard deviation for continuous variables, and percentages for categorical variables. When the distribution of the continuous variable is asymmetric, we also report the median, interquartile range, and minimum and maximum value. For Table 1 and patient characteristics, the unit of observation is the patient. For Tables 2 and 3, which present characteristics of the Reference Standard Delirium Assessments (RSDA) and timing of the clinician delirium identification protocols, the unit of observation is the delirium assessment.

The test characteristics of the UB-2 and 2-step delirium identification protocols are presented in Table 4 and associated text. Clinicians' protocols performed by physicians, nurses, and certified nursing assistants (CNAs) are compared to RSDAs. We present modeled point estimates and 95% confidence intervals for each test characteristic, with each clinical discipline considered separately. Test characteristics reported include overall accuracy (agreement of the clinician protocol with the RSDA), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratio positive (LRP), and likelihood ratio negative (LRN). No comparisons of statistical difference between the assessments are performed. The unit of observation is the delirium assessment and each participant has up to two assessments.

We perform modeling to account for cluster effects inherent in our study design. We consider cluster effects from three sources—up to two assessments within a patient, multiple

assessments performed by each clinician, and multiple assessments performed by each reference standard rater. Estimates of accuracy, sensitivity, specificity, PPV, NPV, LRP, LRN come from the mean of the fitted probabilities derived from the models.

For example, accuracy is defined to be 1 if the clinician's assessment of delirium agrees with the RSDA and 0 otherwise regardless of whether delirium is present or not. This probability of agreement, or accuracy, is estimated from the generalized linear mixed effects model (GLIMMIX) with logit link:

$$log\frac{P_{ijk}(accuracy=1|Z_{ijk})}{P_{ijk}(accuracy=0|Z_{ijk})} = \beta_0 + \alpha_i + \delta_j + \gamma_k + e_{ijk}$$

To address clustering, this model includes random effects from patients α_i (two assessments for each patient on two different days), reference standard raters δ_j, clinicians γ_k , and the error term e_{ijk}. These random effects have a normal distribution with mean zero, and variance components of $\sigma_\alpha^2, \sigma_\delta^2, \sigma_\gamma^2$, for patients, reference standard raters, and clinicians, respectively. The estimation of these variance components account for the cluster effects and are used in the estimation of the intercept β_0, which is the quantity of interest and used to derive estimates of the test characteristics, in this case, accuracy.

The model as described above did not converge due to the large imbalance in the distribution of the cluster sizes related to the reference standard raters and clinicians. Therefore, we sought to simplify the model by identifying and including in the model only non-ignorable cluster effects, as recommended by Kahan and Morris, *BMC Medical Research Methodology*, 2013 (full reference below). As recommended in this article, we estimated the correlation of the outcome (in this case accuracy) within each patient, within each reference standard rater, and within each clinician. We ran the models in SAS using Generalized Estimating Equations (PROC GEE)

with exchangeable correlation structure to estimate the correlation due to the above clusters. The cluster effect due to within patient correlation is largest (correlation near 0.3), and therefore non-ignorable. The correlation due to reference standard raters was near zero (<0.01); therefore, we judged this cluster effect to be ignorable. The cluster effect due to clinicians is small but not zero (between 0.01-0.03); therefore we decided to retain cluster effects for clinicians in the model.

We then performed GLIMMIX modeling with two random effects for patients and clinicians, dropping reference standard rater cluster effects, as described above. This model for accuracy converged, and we were able to extract marginal probabilities from the adjusted intercept to estimate accuracy. However, similar models with two random effects did not converge for sensitivity, specificity, LRP or LRN estimates. Therefore, as recommended by Kahan and Morris, we turned to using fixed effects for clinicians (indicator variable using the clinician IDs). However, the number of clinicians (53 physicians, 236 nurses, 110 CNAs) was too large for the fixed effect model accounting for individual clinicians to converge given our sample size. Thus, we turned to using fixed effects for characteristics of the clinicians instead of the ID indicator. Specifically, we included four factors: clinician age, clinician gender, clinician years practicing at the hospital, and clinician hospital site. This model containing four fixed effects for clinicians converged for all test characteristics of interest, and for all stratified and subset analyses. Therefore, we adopted this approach throughout our paper.

We used this modeling approach for all test characteristics reported in manuscript **Table 4** and **Supplement Tables S7-10,** employing SAS PROC GLIMMIX with random effects for participant clustering, the four fixed effects for clinicians noted above, and no clustering by reference standard. For accuracy, all RSDAs were used in the modeling. For sensitivity, we used the subsample of RSDAs positive for delirium, for specificity the subsample of RSDAs negative for

delirium, for PPV the subsample of clinician assessments positive for delirium, and for NPV the subsample of clinician assessments negative for delirium. For LRP and LRN, we used fitted probabilities from the sensitivity and specificity model. The 95% confidence intervals were computed from taking the mean of the upper and lower bounds of the standard errors of the fitted probabilities. We used the SAS/STAT software version 9.4 64-bit MS Windows 10 for all modeling work. We used PROC GEE to evaluate clustering, and PROC GLIMMIX for modeling of test characteristics, as described above.

Reference: Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. BMC Medical Research Methodology. 2013 Dec; 13: 1-9.

Methods for Supplement Tables

Tables S1, S2, and S3 present means and standard deviations for continuous variables, and percentages for categorical variables. When the distribution of the continuous variable is asymmetric, we also report the median, interquartile range, and minimum and maximum value. For **Table S1**, the unit of observation is the patient participant. For **Table S2**, the unit of observation is the clinician participant. For **Table S3**, the unit of observation is the delirium assessment. **Table S4** presents the same results as **Table 3** in the main manuscript, with outliers greater than 3 standard deviations from the mean trimmed from the sample. **Table S5** presents the same results as **Table 4** without accounting for clustering. **Table S6** presents the raw numbers of assessments underlying the test characteristics reported in **Tables 4 and S5**.

Table S7 presents clinician delirium identification protocol duration data, stratified by dementia status. Similar to **Table 3** in the main manuscript, outliers are not excluded. **Table S8** presents the key test characteristics (overall accuracy, sensitivity, specificity) stratified by dementia status, using the same modeling approach as **Table 4** of the main paper, accounting for clustering by patient and clinician.

Tables S9 and S10 present sensitivity analyses focused on overall accuracy, accounting for clustering. **Table S9** presents several subset analyses, limiting to clinician assessments done within 2 hours of the RSDA, those done by the patient's primary hospital team, those done on study day 1 and day 2 separately, and finally, those done by clinicians who did 10 or more assessments. **Table S10** presents the overall accuracy results for the UB-2 by order of assessment on each day independent of clinician type. We performed this analysis to ensure that the accuracy of the delirium identification protocol does not go down with repeated exposure on the same day.

Table S1. Patient participant characteristics by site

	Overall (n = 527)	Site A (n = 269)	Site B (n = 258)
Age – mean years (SD)	79.7 (6.6)	79.3 (6.6)	80.2 (6.6)
Male – n (%)	226 (43)	118 (44)	108 (42)
Race* – n (%)			
White	458 (88)	209 (78)	249 (97)
Black or African American	43 (8)	42 (16)	1 (0.4)
Others	22 (4)	16 (6)	6 (3)
Hispanic or Latino* – n (%)	11 (2)	7 (3)	4 (2)
Education* – n (%)			
Less than high school	54 (10)	20 (7)	34 (13)
High school graduate	191 (37)	74 (28)	117 (46)
Some college	101 (19)	64 (24)	37 (14)
College graduate	73 (14)	41 (15)	32 (13)
Master's degree	70 (14)	45 (17)	25 (10)
Doctoral degree	33 (6)	23 (9)	10 (4)
Married* – n (%)	224 (43)	104 (39)	120 (47)
Lives Alone* – n (%)	194 (37)	104 (39)	90 (35)
Charlson Score – n (%)			
0	36 (7)	24 (9)	12 (5)
1	172 (33)	91 (34)	81 (31)
2+	319 (60)	154 (57)	165 (64)
ADL* – sum score (SD)	0.9 (1.4)	0.8 (1.5)	0.9 (1.3)
Impaired in ADL* – n (%)	202 (38)	88 (33)	114 (44)
IADL* – sum score (SD)	1.8 (1.8)	1.4 (1.7)	2.2 (1.8)
Impaired in IADL* – n (%)	333 (63)	144 (54)	189 (73)
Dementia – n (%)	183 (35)	80 (30)	103 (40)

Table Abbreviations and Footnotes:
SD=Standard Deviation, Charlson= Charlson comorbidity score, ADL=basic Activities of Daily Living, scored 0-6, 6 worst, IADL=Instrumental Activities of Daily Living, 0-7, 7 worst
*Number of missing data: race (4), ethnicity (2), education (5), marital status (2), living situation (2), ADL (2), IADL (2).

Table S2. Clinician participant characteristics

	Physician (n = 53)	Nurse (n = 236)	CNA (n = 110)
Age – mean years (SD)	36.2 (6.7)	30.7 (9.0)	32.9 (13.4)
Male – n (%)	31 (58)	25 (11)	10 (9)
Race – n (%)			
White	33 (62)	207 (88)	51 (46)
Black or African American	2 (4)	12 (5)	37 (34)
Others	18 (34)	17 (7)	22 (20)
Hispanic or Latino* – n (%)	1 (2)	14 (6)	13 (12)
Native English Speaker – n (%)	48 (91)	217 (92)	75 (68)
Education* – n (%)			
Less than high school	0 (0)	0 (0)	2 (2)
High school graduate	0 (0)	0 (0)	76 (70)
Some college	0 (0)	15 (6)	12 (11)
College graduate	0 (0)	202 (87)	16 (14)
Master's degree	0 (0)	17 (7)	3 (3)
Doctoral degree	53 (100)	0 (0)	0 (0)
Certified in Geriatrics/Gerontology – n (%)	1 (2)	2 (1)	4 (4)
Years of Practice ≤ 5 – n (%)	33 (62)	160 (68)	69 (63)
Years in the Current Hospital ≤ 5 – n (%)	39 (74)	179 (76)	82 (75)

Table Abbreviations and Footnotes:
CNA=certified nursing assistants, SD=standard deviation
* Number of missing data: ethnicity (3), education (3).

Table S3. Reference standard delirium assessment CAM features

		Overall (n = 924)	Without Delirium (n = 770)	With Delirium (n = 154)
Acute Change	Not present	647 (70)	635 (82)	12 (8)
	Present	277 (30)	135 (18)	142 (92)
Fluctuating Course	Not present	839 (91)	763 (99)	76 (49)
	Present	85 (9)	7 (1)	78 (51)
Inattention	Not present	313 (34)	313 (41)	0 (0)
	Present, mild	401 (43)	348 (45)	53 (34)
	Present, marked	210 (23)	109 (14)	101 (66)
Disorganized Thinking	Not present	747 (80)	726 (94)	21 (14)
	Present, mild	135 (15)	36 (5)	99 (64)
	Present, marked	42 (5)	8 (1)	34 (22)
Altered Level of Consciousness	Not present	883 (96)	768 (100)	115 (74)
	Present, mild	40 (4)	2 (0)	38 (25)
	Present, marked	1 (0)	0 (0)	1 (1)
Disorientation	Not present	546 (59)	532 (69)	14 (9)
	Present, mild	261 (28)	196 (25)	65 (42)
	Present, marked	117 (13)	42 (6)	75 (49)
Memory Impairment	Not present	521 (56)	500 (65)	21 (14)
	Present, mild	339 (37)	250 (32)	89 (57)
	Present, marked	64 (7)	20 (3)	44 (29)
Perceptual Disturbances	Not present	771 (83)	698 (91)	73 (47)
	Present, mild	137 (15)	69 (9)	68 (44)
	Present, marked	16 (2)	3 (0)	13 (9)
Psychomotor Agitation	Not present	889 (96)	761 (99)	128 (83)
	Present, mild	31 (4)	7 (1)	24 (16)
	Present, marked	4 (0.4)	2 (0)	2 (1)
Psychomotor Retardation	Not present	851 (92)	740 (96)	111 (72)
	Present, mild	66 (7)	28 (4)	38 (25)
	Present, marked	7 (1)	2 (0)	5 (3)
Sleep-Wake Cycle Disturbance	Not present	536 (58)	446 (58)	90 (58)
	Present, mild	384 (42)	324 (42)	60 (39)
	Present, marked	4 (0)	0 (0)	4 (3)

Table Abbreviations and Footnotes:
CAM=Confusion Assessment Method

Table S4. Duration of clinician delirium identification protocols with trimmed outliers[*]

	UB-2			2-Step		2-Step, No Skip		2-Step, Skip (UB-CAM)	
	CNA (n = 856)	Nurse (n = 869)	Physician (n = 852)	Nurse (n = 859)	Physician (n = 850)	Nurse (n = 416)	Physician (n = 416)	Nurse (n = 443)	Physician (n = 434)
Mean (SD), sec	59 (37)	54 (31)	53 (33)	100 (91)	103 (96)	118 (112)	127 (114)	83 (61)	81 (68)
Median (Q1,Q3), sec	49 (35, 73)	46 (32, 69)	43 (31, 66)	59 (34, 140)	61 (33, 153)	54 (32, 203)	66 (32, 202)	62 (36, 112)	58 (33, 103)
Min, sec	6	6	6	6	6	6	6	9	6
Max, sec	261	189	208	475	528	475	525	393	528

Table Abbreviations and Footnotes:
UB-2=Ultra-brief 2-Item Screen, 2-Step=2-Step Delirium Identification Protocol, UB-CAM=Ultra-brief CAM,
[*]Outliers ≥ 3 SD from the mean were removed: UB-2 CNA (6), Nurse (4), Physician (4) and 2-Step Nurse (7), Physician (4).
SD=Standard Deviation, sec=Seconds, Q1=quartile 1, Q3=quartile 3, Min=Minimum, Max=Maximum

Table S5. Test characteristics of clinician delirium identification protocols not accounting for clustering*

	UB-2			2-Step		2-Step, No Skip		2-Step, Skip (UB-CAM)	
	CNA (n = 898)	Nurse (n = 910)	Physician (n = 895)	Nurse (n = 902)	Physician (n = 893)	Nurse (n = 441)	Physician (n = 440)	Nurse (n = 461)	Physician (n = 453)
Accuracy, %	68.5 (65.3-71.5)	73.8 (70.9-76.7)	70.1 (66.9-73.0)	88.9 (86.7-90.9)	86.9 (84.5-89.0)	88.9 (85.6-91.7)	86.6 (83.0-89.6)	88.9 (85.7-91.7)	87.2 (83.8-90.1)
Sensitivity, %	87.8 (81.3-92.6)	85.5 (78.9-90.7)	81.9 (74.7-87.7)	65.1 (56.9-72.7)	63.5 (55.2-71.3)	65.7 (53.4-76.7)	70.7 (59.0-80.6)	64.6 (53.0-75.0)	56.2 (44.1-67.8)
Specificity, %	64.7 (61.2-68.1)	71.5 (68.1-74.7)	67.7 (64.2-71.0)	93.6 (91.6-95.3)	91.5 (89.3-93.4)	93.3 (90.2-95.6)	89.9 (86.3-92.8)	94.0 (91.1-96.1)	93.2 (90.1-95.5)
PPV, %	32.7 (28.1-37.6)	37.6 (32.5-42.9)	33.6 (28.8-38.7)	66.9 (58.6-74.5)	59.9 (51.8-67.6)	64.8 (52.5-75.8)	58.9 (48.0-69.2)	68.9 (57.1-79.2)	61.2 (48.5-72.9)
NPV, %	96.4 (94.4-97.9)	96.1 (94.2-97.5)	94.9 (92.7-96.6)	93.1 (91.1-94.8)	92.7 (90.5-94.4)	93.5 (90.5-95.8)	93.7 (90.6-96)	92.8 (89.7-95.1)	91.7 (88.5-94.3)
LRP	2.5 (2.2-2.8)	3.0 (2.6-3.4)	2.5 (2.2-2.9)	10.2 (7.2-13.3)	7.5 (5.5-9.5)	9.8 (5.7-13.8)	7.0 (4.6-9.3)	10.7 (6.1-15.3)	8.2 (4.7-11.7)
LRN	0.2 (0.1-0.3)	0.2 (0.1-0.3)	0.3 (0.2-0.4)	0.4 (0.3-0.5)	0.4 (0.3-0.5)	0.4 (0.2-0.5)	0.3 (0.2-0.4)	0.4 (0.3-0.5)	0.5 (0.3-0.6)

Table Abbreviations and Footnotes:
UB-2=Ultra-brief 2-Item Screen, 2-Step=2-Step Delirium Identification Protocol, UB-CAM=Ultra-brief CAM, PPV=Positive Predictive Value, NPV=Negative Predictive Value, LRP=Likelihood Ratio Positive, LRN=Likelihood Ratio Negative
* The values in the parentheses are 95% confidence intervals.

Table S6. Clinician delirium identification protocol test characteristics, raw numbers

	UB-2			2-Step		2-Step, No Skip		2-Step, Skip (UB-CAM)	
	CNA (n = 898)	Nurse (n = 910)	Physician (n = 895)	Nurse (n = 902)	Physician (n = 893)	Nurse (n = 441)	Physician (n = 440)	Nurse (n = 461)	Physician (n = 453)
Accuracy	615/898	672/910	627/895	802/902	776/893	392/441	381/440	410/461	395/453
Sensitivity	129/147	130/152	122/149	97/149	94/148	46/70	53/75	51/79	41/73
Specificity	486/751	542/758	505/746	705/753	682/745	346/371	328/365	359/382	354/380
PPV	129/394	130/346	122/363	97/145	94/157	46/71	53/90	51/74	41/67
NPV	486/504	542/564	505/532	705/757	682/736	346/370	328/350	359/387	354/386

Table Abbreviations and Footnotes:
UB-2=Ultra-brief 2-Item Screen, 2-Step=2-Step Delirium Identification Protocol, UB-CAM=Ultra-brief CAM, PPV=Positive Predictive Value, NPV=Negative Predictive Value

Table S7. Duration of clinician delirium identification protocols stratified by dementia

	With Dementia					Without Dementia				
	UB-2			2-Step		UB-2			2-Step	
	CNA (n=314)	Nurse (n= 320)	Physician (n= 314)	Nurse (n= 318)	Physician (n= 313)	CNA (n= 584)	Nurse (n= 590)	Physician (n= 581)	Nurse (n= 584)	Physician (n= 580)
Mean (SD), sec	71 (50)	65 (39)	69 (56)	150 (114)	153 (122)	57 (51)	50 (29)	48 (34)	79 (80)	82 (86)
Median (Q1,Q3), sec	58 (40, 87)	56 (39, 81)	57 (40, 86)	111 (59, 227)	120 (61, 217)	45 (34, 65)	41 (31, 60)	37 (29, 56)	46 (31, 93)	44 (30, 100)
Min, sec	6	6	6	6	6	6	6	6	9	6
Max, sec	447	326	775	674	936	862	231	343	538	528

Table Abbreviations and Footnotes:
UB-2=Ultra-brief 2-Item Screen, 2-Step=2-Step Delirium Identification Protocol,UB-2=Ultra-brief Two Item Screen, 2-Step=Two Step Delirium Identification Protocol, 95% C.I.=95% Confidence Interval, CNA=Certified Nursing Assistant

Table S8. Clinician identification protocol test characteristics stratified by dementia*

	With Dementia					Without Dementia				
	UB-2			2-Step		UB-2			2-Step	
% 95% C.I.	CNA (n=314)	Nurse (n= 320)	Physician (n= 314)	Nurse (n= 318)	Physician (n= 313)	CNA (n= 584)	Nurse (n= 590)	Physician (n= 581)	Nurse (n= 584)	Physician (n= 580)
Accuracy	58.0 (44.9-70.0)	62.8 (50.4-73.4)	58.5 (45.7-69.9)	80.1 (68.7-87.9)	77.9 (66.1-86.3)	74.4 (64.9-82.0)	79.1 (71.0-85.3)	75.9 (67.0-82.8)	93.7 (87.7-96.8)	91.6 (85.1-95.4)
Sensitivity	93.0 (72.8-98.4)	92.4 (74.0-98.0)	86.4 (64.7-94.9)	66.9 (44.4-83.0)	71.2 (50.9-85.6)	77.5 (45.7-92.9)	77.0 (51.0-92.6)	72.0 (40.4-90.3)	63.8 (36.8-86.7)	45.1 (19.8-71.8)
Specificity	40.3 (26.5-56.7)	48.2 (33.8-62.6)	43.9 (29.3-59.5)	84.6 (71.1-92.4)	80.3 (65.9-89.5)	73.6 (63.6-81.7)	78.9 (70.3-85.5)	75.8 (66.4-83.1)	96.6 (90.7-98.6)	95.7 (90.2-98.1)

Table Abbreviations and Footnotes:
*Test characteristics reported this table account for clustering using the methods described in the paper
UB-2=Ultra-brief 2-item Screen, 2-Step=Two Step Delirium Identification Protocol, 95% C.I.=95% Confidence Interval, CNA=Certified Nursing Assistant

Table S9. Sensitivity Analyses for Overall Accuracy

%, Accuracy 95% C.I.	UB-2			2-Step	
	CNA	Nurse	Physician	Nurse	Physician
Overall Sample	68.6 (60.8-75.4)	73.4 (66.5-79.2)	69.7 (62.3-76.2)	88.9 (83.2-92.7)	86.8 (81.0-90.9)
Assessments with 2 hours of RSDA	68.1 (58.9-76.0)	73.8 (65.5-80.5)	70.3 (61.5-77.7)	90.0 (83.2-94.1)	86.9 (79.8-91.6)
Assessment by Primary Hospital Team Member	69.5 (57.7-79.2)	76.1 (65.7-83.9)	70.8 (59.0-79.8)	88.1 (79.1-93.1)	86.5 (76.4-92.2)
Day 1 Assessment	68.9 (59.2-77.2)	73.0 (64.3-80.2)	68.6 (59.3-76.4)	87.8 (80.7-92.4)	86.7 (78.9-91.7)
Day 2 Assessment	68.5 (56.9-78.1)	74.9 (64.5-82.7)	72.0 (61.5-80.5)	90.8 (82.5-95.2)	87.1 (77.7-92.6)
Assessment done by clinicians with >10 assessments	70.5 (61.5-78.1)	74.6 (64.8-82.1)	69.4 (61.4-76.3)	89.9 (81.5-94.5)	86.1 (79.7-90.7)

Table Abbreviations and Footnotes:
*Test characteristics reported this table account for clustering using the methods described in the paper
UB-2=Ultra-brief 2-Item Screen, 2-Step=Two Step Delirium Identification Protocol, 95% C.I.=95% Confidence Interval, CNA=Certified Nursing Assistant

Table S10. Accuracy by Interview Order, UB-2 only

% Accuracy 95% C.I.	UB-2
First Interview	68.8 (61.1-75.5)
Second Interview	70.4 (62.7-77.1)
Third Interview	70.9 (63.4-77.3)

Table Abbreviations and Footnotes:
*Test characteristics reported this table account for clustering using the methods described in the paper
UB-2=Ultra-brief 2-Item Screen, 95% C.I.=95% Confidence Interval