**Supplemental Materials**

**Materials and Methods**

*Study Design*

The conception, planning, execution, and interpretation of this study were co-developed with recovered ICU patients and family members. The acceptability and feasibility of family-administered delirium detection in the ICU, along with the study protocol and recruitment procedures, were assessed in a pilot study(33). To improve the recruitment rate following the pilot study a family member of a former ICU patient (BGS) was engaged as a researcher on the team. Data collection was planned before the family-administered delirium detection tools and the reference standard assessments were performed.

*Measures*

The RASS is a 10-point agitation-sedation scale centered at 0 (indicates calm and alert), with scores ranging from +4 to -5; more negative scores indicate greater levels of sedation and more positive scores indicate higher levels of agitation(37). The RASS is assessed by on-duty bedside nurses as part of standard care every four hours and recorded in eCritical. Delirium was only assessed in individuals with a RASS of ≥-3 (indicating an arousable patient).

*Reference Standard*

Thirteen ICU research nurses participated in the study, with an average of 4.3 (standard deviation [SD] 2.2) years of ICU experience. A neuropsychiatrist (ZI) trained the ICU research nurses on operationalizing Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)(1) criteria to identify delirium. As part of the training, the ICU research nurses independently read four case vignettes and used DSM-5 criteria to identify delirium. The study team, including the neuropsychiatrist, bedside nurses, and intensivists, developed a standardized assessment form operationalizing DSM-5 criteria, which was employed by the ICU research nurse for each reference standard assessment. Each ICU research nurse and the neuropsychiatrist independently conducted a minimum of 10 assessments during a pre-study training period and met to establish agreement. Research nurses completing the reference standard assessment were asked to record whether they remained blind to the delirium status of the patient.

Delirium subtypes were identified according to the results of the reference standard assessment of delirium (patients with delirium only) and scores on the RASS(38). Hyperactive delirium was defined as those with a RASS between +1 and +4, hypoactive delirium as a RASS between -1 and -3, and mixed delirium as those with RASS values fluctuating between positive and negative values.

*Data Analysis*

Binomial regression was conducted to model delirium status on each tool as a function of the reference standard assessment. The resulting coefficients represent the likelihood ratio positive, which is the probability of a positive test given delirium according to the reference standard assessment over the probability of a positive test given no delirium according to the reference standard assessment. Effect modification was explored using interaction terms, and models were also used to create adjusted estimates to identify confounding. Variables considered potential effect modifiers or confounders *a priori* were patient age, patient sex, caregiver age, caregiver sex, caregiver symptoms of depression, caregiver symptoms of anxiety, patient severity of illness (APACHE-II score upon ICU admission), mechanical ventilation at the time of assessment, and RASS score at the time of assessment.

The performance of the FAM-CAM and Sour Seven compared to the reference standard was assessed in five subgroups: (1) 65 years or older; (2) with higher severity of illness (APACHE-II score upon admission greater or equal to the median value of 20); (3) considered frail (Clinical Frailty Score upon admission of ≥5)(40); (4) invasive mechanical ventilation status (yes/no at the time off assessment); and (5) level of sedation (RASS ≥+1 vs. RASS <+1 at the time of assessment).

**Results**

*Criterion Validity of the CAM-ICU*

 Compared to the RN reference standard assessment, the CAM-ICU had a sensitivity of 76.2% (95%CI: 69.0-82.4), specificity of 82.8% (95%CI: 78.0-86.9), and AUC of 80.0% (95%CI: 76.0-83.0). The criterion validity of the delirium measurement tools using the CAM-ICU as the reference standard are presented in Supplemental Table 1 (Supplemental Digital Content 1, http://links.lww.com/CCM/F460).

*Criterion Validity of the FAM-CAM and Sour Seven*

Other cutpoints (from 1-17) for the Sour Seven were explored, however they did not result in improved measures of diagnostic accuracy over the previously published cutpoints (data not shown).

With the likelihood ratio positive as the effect measure, there was effect modification by patient sex (p=0.049) for the FAM-CAM: the likelihood ratio positive for females (n=58) was 8.7 (95%CI: 2.12-35.5) and for males (n=89) 1.6 (95%CI: 0.7-3.9). The ratio of the probability of a participant with delirium according to the reference standard testing positive for delirium on the FAM-CAM relative to the probability of a participant without delirium according to the reference standard testing positive on the FAM-CAM (the likelihood ratio positive) is higher among women than it is among men. No effect modification was observed for any other variable on the FAM-CAM and the Sour Seven. We observed confounding by family member age for both the FAM-CAM (crude: 2.9; adjusted: 3.2) and Sour Seven (crude: 4.8; adjusted: 5.8), where older age was associated with an increased likelihood ratio positive.

*Ease of Use of the FAM-CAM and Sour Seven*

No adverse events were reported by family members while completing the index tests; on the contrary, family members indicated satisfaction with filling out these tools.