Natural language processing to assess the epidemiology of delirium-suggestive behavioural disturbances in critically ill patients

Marcus Young, Natasha Holmes, Raymond Robbins, Nada Marhoon, Sobia Amjad, Ary Serpa Neto and Rinaldo Bellomo

ABSTRACT

Background: There is no gold standard approach for delirium diagnosis, making the assessment of its epidemiology difficult. Delirium can only be inferred though observation of behavioural disturbance and described with relevant nouns or adjectives.

Objective: We aimed to use natural language processing (NLP) and its identification of words descriptive of behavioural disturbance to study the epidemiology of delirium in critically ill patients.

Study design: Retrospective study using data collected from the electronic health records of a university-affiliated intensive care unit (ICU) in Melbourne, Australia.

Participants: 12 375 patients

Intervention: Analysis of electronic progress notes. Identification using NLP of at least one of a list of words describing behavioural disturbance within such notes.

Results: We analysed 199 648 progress notes in 12 375 patients. Of these, 5108 patients (41.3%) had NLP-diagnosed behavioural disturbance (NLP-Dx-BD). Compared with those who did not have NLP-Dx-DB, these patients were older, more severely ill, and likely to have medical or unplanned admissions, neurological diagnosis, chronic kidney or liver disease and to receive mechanical ventilation and renal replacement therapy ($P < 0.001$). The unadjusted hospital mortality for NLP-Dx-BD patients was 14.1% versus 9.6% for patients without NLP-Dx-BD. After adjustment for baseline characteristics and illness severity, NLP-Dx-BD was not associated with increased risk of death (odds ratio [OR], 0.94; 95% CI, 0.80–1.10); a finding robust to multiple sensitivity, subgroups and time of observation subcohort analyses. In mechanically ventilated patients, NLP-Dx-BD was associated with decreased hospital mortality (OR, 0.80; 95% CI, 0.65–0.99) after adjustment for baseline severity of illness and year of admission.

Conclusions: NLP enabled rapid assessment of large amounts of data identifying a population of ICU patients with typical high risk characteristics for delirium. Moreover, this technique enabled identification of previously poorly understood associations. Further investigations of this technique appear justified.

Crit Care Resusc 2021; 23 (2): 144-153

Delirium is a common syndrome in patients admitted to the intensive care unit (ICU)[1] and is associated with mortality, institutionalisation, and long term cognitive impairment.[2-9] Its definition by the fifth edition of the *Diagnostic and statistical manual of mental disorders* (DSM-5) provides guidance to clinicians.[10,11] However, such definition cannot be verified or falsified against an objective standard. Therefore, despite such guidance and the frequency of delirium in ICU patients, its clinical diagnosis and the study of its epidemiology have proved challenging. This is because the diagnosis of delirium is affected, among others, by the degree of surveillance, observer awareness, its fluctuating nature, a background of chronic neurocognitive decline in some patients, the presence or absence of associated physical manifestations (eg, psychomotor agitation), and differences in presentation in ICU patients compared with ward patients.[12-15] Two methodologies, the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) and the Intensive Care Delirium Screening Checklist (ICDSC) have been applied in an attempt to resolve these difficulties,[16,17] but their evaluation has delivered discordant findings.[18-20] In particular, and of great importance, CAM-ICU is applied once or twice a day and is therefore unlikely to reliably capture the development or presence of delirium throughout the day–night cycle.[19] ICDSC is normally completed every 8–24 hours.[17,21] However, the checklist includes questions reviewing indicators over the previous 24 hours for which the individual completing the assessment may not have first-hand knowledge.

Importantly, all methodologies used to diagnose or assess delirium simply describe variable forms of abnormal behavioural phenotypes.

Given the above considerations, continuing patient assessment by nursing, medical or allied health personnel, as reported in their progress

notes, should logically provide a more comprehensive assessment of the patient's behaviour over the full day–night cycle. Such global assessment of behaviour has been shown to identify more patients with delirium than the use of the CAM-ICU test.[19] It is expressed by words, which suggest or imply the presence of behavioural disturbances typically associated with delirium (eg, agitation/agitated, confusion/confused, disorientation/disorientated).[22] Such words can now be analysed by natural language processing (NLP) techniques, which overcome the limitations of the human capacity to read and rapidly analyse thousands of notes and millions of words.[23-25]

NLP uses computer software to analyse the structure of natural language. This software may be applied to identify sentences within electronically recorded progress notes. Once identified, sentences can be converted to lists of words or "tokens" and compared with a reference list of words or expressions of interest. Furthermore, NLP techniques such as "stemming" may be used to reduce the impact of alternate and incorrect spelling on word comparison. Stemming reduces words to their "stem" by removing the last few letters, thereby making comparisons less dependent on word endings.

Accordingly, we used NLP techniques to assess the epidemiology of words suggestive of behavioural disturbance in ICU progress notes. We aimed to test the hypothesis that such words would be used to describe patients with clinical characteristics typical of patients at high risk of conventionally diagnosed delirium. Moreover, we hypothesised that patients identified by such words would have specific clinical characteristics and outcomes consistent with those of patients reported as having delirium in the literature.

Methods

Study design

We performed a retrospective study using data collected from the electronic health records of a university-affiliated ICU in Melbourne, Australia. This study was approved by the Austin Hospital Human Research Ethics Committee (LNR/19/Austin/38) without the need for informed consent given the non-interventional, data-based, anonymised nature of the study.

Setting and population

All adult patients (≥ 18 years old) admitted to the ICU of the Austin Hospital, Melbourne, Australia, between 2 February 2010 and 31 December 2018 were considered for inclusion. For patients who had multiple admissions during the study period, only the first admission was considered for analysis. No further exclusion criteria were considered.

Data collection and manipulation

All baseline and outcome data were collected as part of the Australian and New Zealand Intensive Care Society Adult Patient Database run by the Centre for Outcome and Resource Evaluation.[26]

Using a proprietary intensive care clinical information system, we obtained electronic data from all typed progress notes entered into the ICU-specific electronic health records by doctors, nurses, physiotherapists, and other allied health practitioners. NLP (Natural Language Toolkit; NLTK 3.5) sentence tokenising techniques were applied to convert progress notes into sentence vectors.[27] Each vector was searched for words, terms or expressions that were suggestive of behavioural disturbance (Online Appendix, table S1).

The selection of the terms describing behavioural disturbance potentially associated with delirium was informed by words selected by relevant personnel and described in a previous survey among health care providers including ICU staff.[22] Words suggestive of behavioural disturbance that were associated with negation (eg, "no", "nil" and "not") or resolution (eg, "resolved", "resolving" and "cleared") were excluded (Online Appendix, table S2). In addition, NLP stemming techniques were applied to adjust for spelling or typing mistakes.

Exposure

The primary exposure of the present study was the presence of behavioural disturbance. For the purpose of this study, behavioural disturbance is defined by the presence of one of the words suggestive of behavioural disturbance included in our list in any progress note during an ICU stay and abbreviated to NLP-Dx-BD. The day of NLP-Dx-BD was recorded as the first day when a word suggestive of behavioural disturbance was registered.

Outcomes

The primary outcome was all-cause in-hospital mortality. We additionally assessed ICU mortality, ICU length of stay and hospital length of stay.

Statistical analysis

All continuous data are reported as median with interquartile range (IQR) and categorical data as number and percentage. In the primary descriptive analysis, data from all patients fulfilling inclusion criteria were reported according to the presence (or absence) of words suggestive of behavioural disturbance. No missing data for any of the outcomes were

present in the dataset; therefore, all analyses were complete case analyses. Baseline and clinical characteristics of the patients were compared among the groups using Fisher exact tests and Wilcoxon rank sum tests.

To further assess the adjusted impact of the presence of NLP-Dx-BD on hospital mortality, the overall cohort of the study was narrowed to create nested time cohorts with progressively longer potential exposure to the risk of behavioural disturbance (0–1 day, 0–2 days, 0–3 days, 0–4 days). Importantly, each cohort had a period of potential exposure unaffected by informative censoring from either ICU discharge or death and the cut off of 4 days was chosen as just above the median duration of ICU stay. This approach was applied to balance informative censoring of patient data, thus maintaining a uniform exposure potential within each subset. The cohorts included patients with ICU length of stay of at least one day (0–1), 2 days (0–2), 3 days (0–3) or 4 days (0–4). Patients who died or were discharged before each time point were excluded from the cohort. The assessment of words suggestive of behavioural disturbances started at the time and date of ICU admission until day 1, 2, 3 and 4, according to the cohorts described above.

Multivariable logistic regression models were used to assess the impact of NLP-Dx-BD on hospital mortality. In all analyses, four models were fitted, one for each cohort. All models were adjusted by year of ICU admission as a categorical variable and by the Australian and New Zealand Risk of Death (ANZROD) after log transformation.[28] As previously shown, ANZROD is a powerful predictor and explains most of the mortality in ICUs in Australia and New Zealand. In addition, ANZROD is superior to the Acute Physiology and Chronic Health Evaluation (APACHE) III scores in predicting mortality in Australia and New Zealand, with an area under the receiver operating characteristic curve (AUROC) of 0.902.[29]

To further understand the impact of NLP-Dx-BD according to baseline characteristics, additional models including an interaction between NLP-Dx-BD and these characteristics were fitted. The following characteristics were assessed:

- use of mechanical ventilation;
- type of admission (elective surgery, urgent surgery or medical);
- source of admission (emergency room, operating room, ward or other);
- tertiles of age; and
- ICU admission diagnosis (cardiovascular, respiratory, sepsis, trauma, gastrointestinal or neurological).

All analyses were conducted in R v.3.6.3 (R Foundation for Statistical Computing, Vienna, Austria) and $P < 0.05$ was considered statistically significant.

Results

Patients

Using NLP techniques, we analysed 69 645 684 words in 199 648 progress notes. Such analysis identified 12 609 patients, and after exclusions, included 12 375 patients in the overall cohort and 11 626 in the time cohort 0–1 (patients with at least 24 hours of ICU length of stay) (Online Appendix, eFigure 1). In addition, according to planned methodology, the study cohort was further segmented into three additional time cohorts: cohort 0–2 (7517 patients, 64.6%), 0–3 (4893 patients, 42.1%), and 0–4 (3394 patients, 29.2%).

Overall, health care personnel used words suggestive of behavioural disturbance to characterise 5108 patients (41.3%) as having NLP-Dx-BD. The baseline characteristics of study patients at ICU admission are shown in Table 1. Overall, the median age was 64.0 years (IQR, 51.2–74.2 years), most patients (61.6%) were male, and 57.0% received mechanical ventilation. Patients with NLP-Dx-BD were older, more severely ill, less likely to be admitted from the operating room and more likely to be admitted from the ward. They were also less likely to have a planned ICU admission, more likely to be admitted after a rapid response team review, more likely to be admitted under a medical unit, and had different diagnostic categories, especially greater percentage of neurological diagnosis. Moreover, they had a greater prevalence of chronic diseases (eg, cirrhosis, chronic kidney disease, and hepatic failure). Finally, patients with NLP-Dx-BD were more likely to be mechanically ventilated and treated with renal replacement therapy on the day of admission. Vital signs and laboratory tests on the day of admission (Online Appendix, table S3) showed greater derangement among patients classified as having NLP-Dx-BD. The baseline characteristics according to the different time cohorts (Online Appendix, table S4a) were broadly consistent with the overall characteristics across all time-based cohorts.

NLP-Dx-BD over time

The prevalence of NLP-Dx-BD increased from 31.0% (95% CI, 28.3–33.5%) in 2010 to 47.9% (95% CI, 45.3–50.6%) in 2018, an average increase of 1.3% (95% CI, 1.0–1.7%) per year (Online Appendix, eFigure 2). However, the relationship between the presence or absence of NLP-Dx-BD and mortality remained constant over the same period (Online Appendix, eFigure 2, B). NLP-Dx-BD was mostly diagnosed in the first 4 days of ICU stay, with a peak on the day after ICU admission (Online Appendix, eFigure 3).

Table 1. Natural language processing-diagnosed behavioural disturbance (NLP-Dx-BD) versus no NLP-Dx-BD

	NLP-Dx-BD	No NLP-Dx-BD	*P*
Total number of patients	5108	7267	
Median age, years (IQR)	64.9 (51.3–75.4)	63.4 (51.1–73.4)	< 0.001
Sex, male	3205 (62.7%)	4425 (60.9%)	0.037
Median body mass index, kg/m^2 (IQR)	28.0 (24.7–31.7)	28.0 (24.8–32.0)	0.699
Severity of illness, median (IQR)			
APACHE III	57.0 (42.0–74.0)	45.0 (33.0–60.0)	< 0.001
ANZROD	5.9% (1.3–21.0%)	1.7% (0.5–7.6%)	< 0.001
Source of admission			< 0.001
Operating room	1962 (38.4%)	3741 (51.5%)	
Emergency room	1478 (28.9%)	1872 (25.8%)	
Ward	928 (18.2%)	891 (12.3%)	
Other hospital	541 (10.6%)	598 (8.2%)	
Other ICU	198 (3.9%)	106 (1.5%)	
Others	1 (0.0%)	59 (0.8%)	
Planned admission	1145 (24.1%)	2728 (41.3%)	< 0.001
Rapid response team admission	922 (19.4%)	949 (14.4%)	< 0.001
Cardiac arrest	206 (4.3%)	218 (3.3%)	0.016
Acute respiratory failure	144 (3.1%)	110 (1.7%)	< 0.001
Type of admission			< 0.001
Medical	2775 (58.5%)	2805 (42.5%)	
Elective surgery	1143 (24.1%)	2732 (41.4%)	
Urgent surgery	824 (17.4%)	1070 (16.2%)	
Diagnostic category			< 0.001
Cardiovascular	1550 (30.3%)	2967 (40.8%)	
Gastrointestinal	916 (17.9%)	1186 (16.3%)	
Respiratory	757 (14.8%)	1090 (15.0%)	
Sepsis	391 (7.7%)	594 (8.2%)	
Neurological	470 (9.2%)	263 (3.6%)	
Metabolic	316 (6.2%)	325 (4.5%)	
Muscle and skin	261 (5.1%)	264 (3.6%)	
Trauma	250 (4.9%)	197 (2.7%)	
Renal and genitourinary	149 (2.9%)	319 (4.4%)	
Haematological	44 (0.9%)	45 (0.6%)	
Gynaecological	2 (0.0%)	11 (0.2%)	
Other	2 (0.0%)	6 (0.1%)	
Co-existing disorders			
Diabetes	1168 (22.9%)	1498 (20.6%)	0.003
Cirrhosis	424 (8.9%)	336 (5.1%)	< 0.001
Chronic respiratory disease	302 (5.9%)	363 (5.0%)	0.029
Use of immunosuppressive drugs	280 (5.5%)	327 (4.5%)	0.014
Chronic kidney disease	272 (5.3%)	284 (3.9%)	< 0.001
Immunological disease	230 (4.5%)	271 (3.7%)	0.033
Metastatic cancer	200 (3.9%)	332 (4.6%)	0.079
Chronic cardiovascular disease	195 (3.8%)	206 (2.8%)	0.003
Hepatic failure	114 (2.2%)	67 (0.9%)	< 0.001
Leukaemia	65 (1.3%)	92 (1.3%)	0.999
Lymphoma	55 (1.1%)	70 (1.0%)	0.584
Acute HIV infection	10 (0.2%)	3 (0.0%)	0.011
Organ support			
Mechanical ventilation	3314 (64.9%)	3704 (51.4%)	< 0.001
Renal replacement therapy	622 (12.2%)	332 (4.6%)	< 0.001
Clinical outcomes			
ICU mortality	411 (8.0%)	527 (7.3%)	0.105
Median ICU LOS, days (IQR)	3.4 (1.8–6.5)	1.2 (0.8–2.3)	< 0.001
Median hospital LOS, days (IQR)	15.4 (8.1–36.2)	8.1 (5.2–14.9)	< 0.001
In-hospital mortality	721 (14.1%)	697 (9.6%)	< 0.001

ANZROD = Australian and New Zealand Risk of Death; APACHE = Acute Physiology and Chronic Health Evaluation; HIV = human immunodeficiency virus; ICU = intensive care unit; IQR = interquartile range; LOS = length of stay. Percentages may not total 100 because of rounding.

Unadjusted association between NLP-Dx-BD and patient characteristics and outcomes

The unadjusted primary and secondary outcomes (Table 2) show that NLP-Dx-BD patients had a significantly greater hospital mortality rate (but not a greater ICU mortality rate). Moreover, NLP-Dx-BD patients had a longer duration of ICU and hospital stay, a difference consistent across all four time cohorts (Online Appendix, table S4b). However, the difference in ICU and hospital mortality dissipated as time of observation extended to the 0–4 days cohort. The time to event survival plots for the different time cohorts demonstrate no difference in time to mortality across all time cohorts (Figure 1).

The unadjusted associations of NLP-Dx-BD and mortality overall and for different clinical subgroups across different time cohorts are shown in the Online Appendix, table S5. Overall, there was no significant association between NLP-Dx-BD and mortality across all time cohorts. However, for patients not receiving mechanical ventilation, the odds ratio (OR) for mortality varied between 1.82 and 1.90 ($P < 0.007$ to $P < 0.001$) across all time cohorts (Online Appendix, table S5). Conversely, for patients receiving mechanical ventilation, the OR for mortality varied between 0.75 and 0.95 ($P < 0.585$ to $P < 0.008$). In addition, several other subgroups and diagnostic categories had variable ORs, which achieved significance at different times and with

Table 2. Clinical outcomes

	NLP-Dx-BD	No NLP-Dx-BD	*P*
Total number of patients	5108	7267	
Primary outcome			
Hospital mortality	721 (14.1%)	697 (9.6%)	< 0.001
Secondary outcomes			
ICU mortality	411 (8.0%)	527 (7.3%)	0.105
Median ICU LOS, days (IQR)	3.4 (1.8–6.5)	1.2 (0.8–2.3)	< 0.001
Median hospital LOS, days (IQR)	15.4 (8.1–36.2)	8.1 (5.2–14.9)	< 0.001

ICU = intensive care unit; IQR = interquartile range; LOS = length of stay; NLP-Dx-BD = *n*atural language processing-diagnosed behavioural disturbance. Percentages may not total 100 because of rounding.

Figure 1. Kaplan–Meier survival plots of time to event for different time of observation cohorts

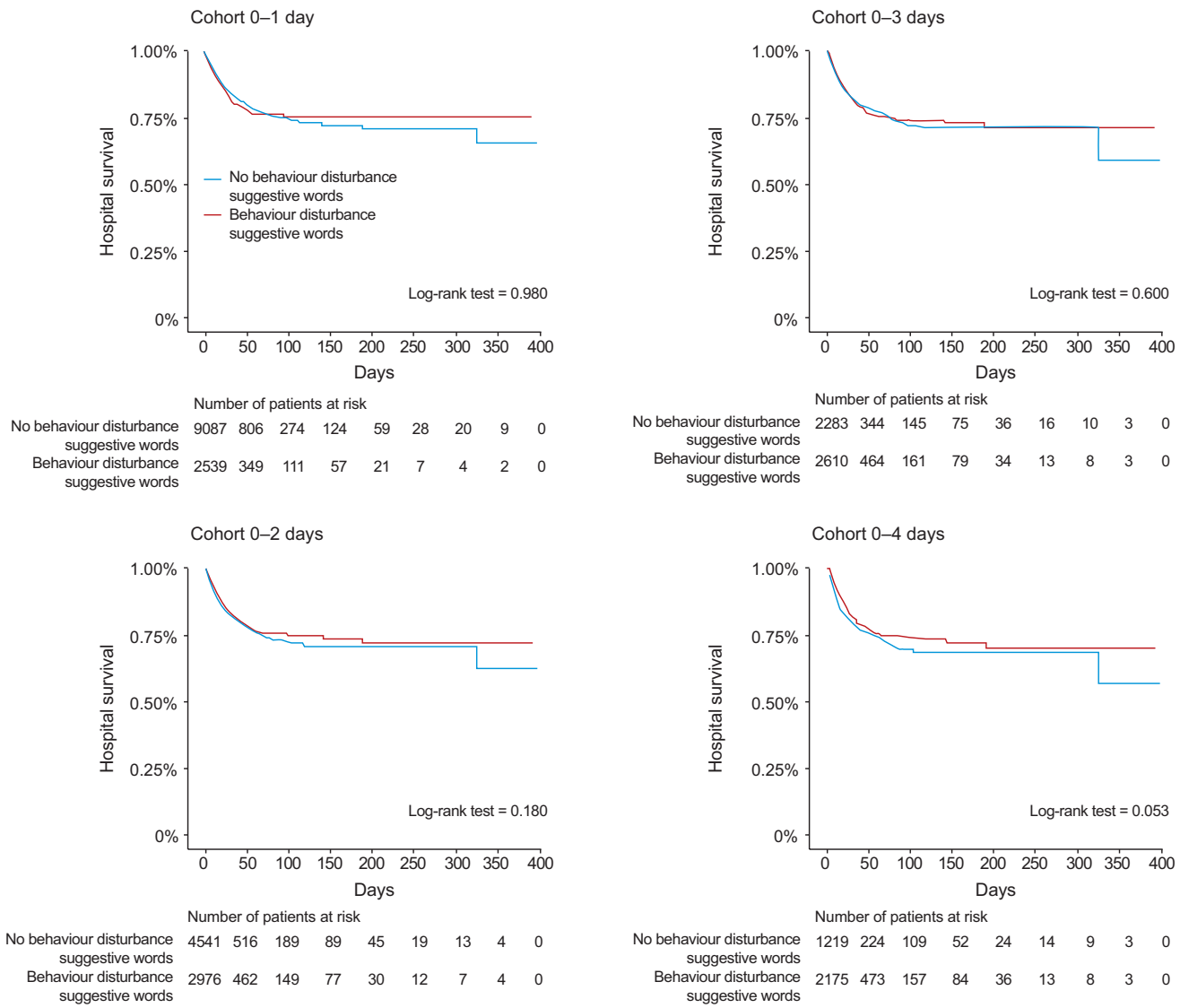


Table 3. Multivariable model of the association of natural language processing-diagnosed behavioural disturbance (NLP-Dx-BD) on hospital mortality according to the different cohorts

	Cohort 0–1 day (*n* = 11 626)		Cohort 0–2 days (*n* = 7517)		Cohort 0–3 days (*n* = 4893)		Cohort 0–4 days (*n* = 3394)	
	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*
Median overall population	0.94 (0.80–1.10)	0.453	0.88 (0.75–1.03)	0.125	0.97 (0.81–1.16)	0.734	0.84 (0.68–1.03)	0.097
Mechanical ventilation								
Yes	0.80 (0.65–0.99)	0.042	0.71 (0.58–0.86)	< 0.001	0.82 (0.66–1.00)	0.056	0.70 (0.55–0.88)	0.003
No	1.29 (0.99–1.68)	0.057	1.42 (1.05–1.93)	0.022	1.45 (0.99–2.12)	0.053	1.45 (0.88–2.45)	0.154
Type of admission								
Elective surgery	1.17 (0.56–2.26)	0.654	1.03 (0.52–1.99)	0.918	0.86 (0.41–1.77)	0.677	1.61 (0.64–4.47)	0.331
Urgent surgery	1.08 (0.72–1.59)	0.697	0.68 (0.46–1.00)	0.057	0.83 (0.53–1.30)	0.410	0.59 (0.35–1.01)	0.053
Medical	0.95 (0.78–1.16)	0.625	0.92 (0.75–1.12)	0.405	0.83 (0.53–1.30)	0.410	0.77 (0.60–0.99)	0.045
Source of admission								
Emergency room	0.74 (0.54–1.01)	0.061	0.82 (0.61–1.10)	0.192	0.84 (0.61–1.16)	0.296	0.88 (0.60–1.29)	0.497
Operating room	1.21 (0.86–1.69)	0.267	0.72 (0.51–1.00)	0.058	0.74 (0.50–1.07)	0.113	0.70 (0.45–1.10)	0.124
Ward	1.19 (0.89–1.60)	0.238	1.28 (0.94–1.74)	0.116	1.64 (1.14–2.38)	0.007	1.06 (0.69–1.62)	0.801
Other	0.83 (0.54–1.25)	0.383	0.79 (0.53–1.18)	0.257	0.83 (0.54–1.27)	0.399	0.71 (0.44–1.13)	0.145
Age								
≤ 56 years	0.83 (0.58–1.17)	0.301	0.83 (0.60–1.15)	0.273	0.79 (0.55–1.12)	0.190	0.70 (0.46–1.04)	0.079
57–70 years	0.94 (0.70–1.26)	0.691	0.71 (0.54–0.95)	0.020	0.73 (0.53–0.99)	0.048	0.73 (0.51–1.05)	0.087
> 70 years	1.01 (0.79–1.28)	0.933	1.10 (0.86–1.41)	0.444	1.41 (1.06–1.89)	0.020	1.06 (0.75–1.51)	0.726
Diagnosis								
Cardiovascular	0.93 (0.65–1.30)	0.662	0.87 (0.63–1.18)	0.376	0.68 (0.48–0.95)	0.024	0.61 (0.42–0.90)	0.012
Respiratory	0.99 (0.66–1.47)	0.977	0.93 (0.62–1.39)	0.732	0.74 (0.46–1.18)	0.205	0.55 (0.32–0.94)	0.030
Sepsis	1.12 (0.69–1.80)	0.635	1.12 (0.69–1.82)	0.633	1.76 (1.05–2.97)	0.031	1.43 (0.77–2.70)	0.257
Trauma	0.96 (0.41–2.15)	0.916	0.42 (0.17–0.99)	0.054	0.58 (0.22–1.44)	0.245	0.47 (0.17–1.30)	0.147
Gastrointestinal	1.21 (0.84–1.74)	0.299	1.07 (0.74–1.55)	0.720	1.04 (0.67–1.62)	0.856	1.07 (0.64–1.82)	0.792
Neurological	0.49 (0.29–0.82)	0.007	0.47 (0.27–0.79)	0.004	0.90 (0.47–1.74)	0.742	0.81 (0.37–1.85)	0.611

variable strength both in the direction of increased and decreased risk (Online Appendix, table S5).

Adjusted analyses

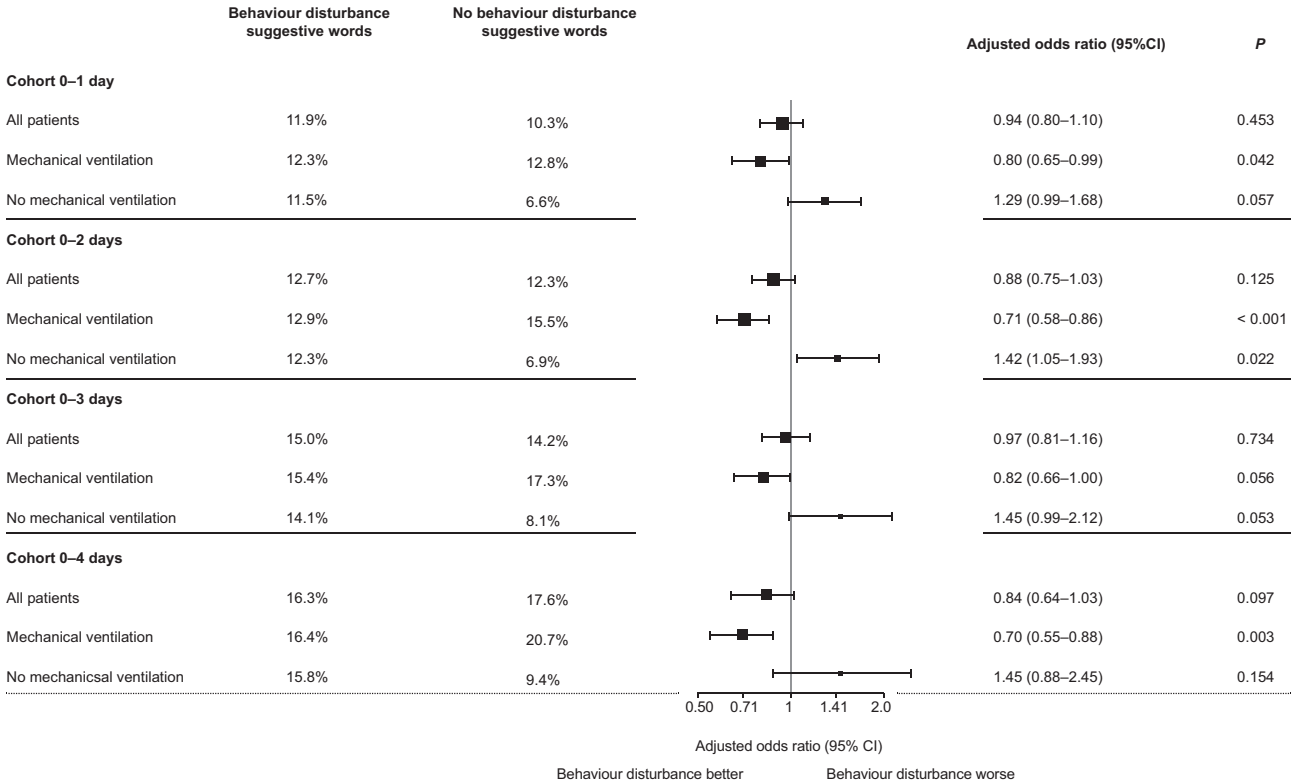
The findings of the multivariable model assessing the independent association between NLP-Dx-BD and mortality after adjustment for key baseline features and time cohorts are shown in Table 3. Overall, NLP-Dx-BD was not associated with an increased OR for hospital mortality. However, the presence or absence of mechanical ventilation significantly modified the OR for mortality across most time cohorts, such that the presence of mechanical ventilation was independently associated with a decreased OR for mortality (Figure 2). On interaction testing, this effect was highly significant across all time cohorts (Online Appendix, table S6). There was no other robust and recurrently significant interaction with any other variable across all time cohorts.

Discussion

Key findings

We used NLP techniques to analyse almost 200 000 medical, nursing and allied health progress notes from more than 12 000 critically ill patients and identified more than 5000 patients with NLP-Dx-BD. These patients were older, more severely ill and more likely to have medical or unplanned admissions, a neurological diagnosis, chronic kidney and liver disease, and to receive mechanical ventilation and renal replacement therapy — all clinical characteristics consistent with the epidemiology of a high risk cohort for delirium. As expected, unadjusted hospital mortality was greater in NLP-Dx-BD patients. However, after adjustment for baseline characteristics and illness severity, NLP-Dx-BD was not independently associated with an increased risk of death, a finding robust to multiple sensitivity subgroups

Figure 2. Forest plot illustrating the impact of mechanical ventilation on outcome for the different time of observation cohorts



and time of observation subcohort analyses. Moreover, in patients receiving mechanical ventilation, NLP-Dx-BD was consistently associated with decreased hospital mortality.

Relationship to previous findings

Making a diagnosis of delirium is challenging because delirium is a fluctuating neurological state and may not be present at the time of assessment. This is because the hypoactive phenotype may not be easily noted and/or because no objective quantitative gold standard test exists to define it.[20] Consequently, multiple diagnostic methodologies have been proposed. All are essentially based on the observation of a behavioural disturbance (agitation, confusion, disorientation etc) or the inability of the patient to satisfactorily answer a series of questions (CAM-ICU). These methodologies have limitations because they are also observer- and frequency of assessment-dependent. For example, the CAM-ICU methodology, while widely considered to have high specificity, has low sensitivity when undertaken by bedside nursing staff during the normal course of patient care.[18] Consistent with this, other investigators have reported that the rates of delirium diagnosis fell significantly after the introduction of CAM-ICU compared with previous unstructured bedside assessments.[19]

In contrast, delirium may also be identified and characterised through the words used by the bedside caregivers who describe behavioural disturbance. These caregivers are in constant contact with and continuously observe the patient and, thus, describe such constant observations in their notes. As shown in a recent survey,[22] when used in clinical progress notes, words such as "confused" and "aggressive" or "disorientated", for example, are readily understood by clinicians to indicate a behavioural disturbance and an acutely altered neurological state and likely delirium. This approach is gaining momentum because of its semantic and semiotic logic,[30] as shown in several small pilot studies, and because of its increased applicability through analytic software.[31]

Previous studies have suggested that delirium may be associated with increased mortality,[2,9,32,33] some of which found an increased risk of mortality even after adjustment for covariates including severity of illness. These findings have created the view that delirium poses a mortality

risk. However, the diagnosis of delirium by conventional methods may have missed up to 70% of cases through a combination of underdiagnosis and underdocumentation, making such associations open to challenge.[7,19,34] Moreover, more recent detailed studies[35,36] found that delirium is, in fact, not independently associated with mortality. Our study of behavioural disturbance aligns with such observations. It also provides novel information on the association between behavioural disturbance and mortality in mechanically ventilated patients, where we found that NLP-Dx-BD was associated with decreased risk. We believe these observations may reflect the bias that such patients would have had to be awakened and considered for weaning in order to manifest behavioural disorders and were thus less likely to be severely ill. This is consistent with previous studies showing that rapidly reversible sedation-related delirium was associated with a reduction in one-year mortality and hospital length of stay compared with persistent delirium.[37]

Implications of study findings

Our findings imply that NLP software can be used to search for words that logically, clinically and epidemiologically define a population of critically ill patients at high risk of behavioural disruptions possibly representing a surrogate for delirium. Moreover, independent of whether these patients have delirium or not, however correctly or incorrectly defined by conventional methodologies, they have the very characteristics used in everyday practice by clinicians to describe its presence. Finally, they imply that NLP keyword-based techniques provide an unprecedented opportunity to analyse millions of words in thousands of clinical progress notes for the purpose of studying the epidemiology of NLP-Dx-BD.

Strengths and limitations

To our knowledge, our study is the first to use NLP techniques to study the epidemiology and outcomes of critically ill patients with behavioural disturbance. Moreover, our study results show the potential of NLP techniques to analyse thousands of clinical progress notes for the purpose of identifying such patients. This technique opens the door to unprecedented large-scale assessment of the epidemiology of this condition. Finally, we used keywords and terms which have face validity and are widely applied by caregivers at the bedside every day to describe patients with possible or probable delirium.

Nevertheless, we acknowledge several limitations of this study. First, we used NLP techniques to identify patients with words suggestive of behavioural disturbance in their clinical notes. We did not investigate if these patients had also been diagnosed with delirium through the application of alternate methodologies. However, the patient cohort we diagnosed with NLP-Dx-BD was consistent with a population at high risk of delirium. Second, bedside staff may recognise and document agitated behavioural disturbances more readily than non-agitated behavioural disturbances, which may cause our technique underdetect non-agitated behavioural disturbances. However, this study investigates NLP-Dx-BD and not the possible phenotypes of NLP-Dx-BD and their rate of occurrence. We intend such analysis to be the subject of a future study. Third, although our study reviewed a large number of clinical progress notes, it is a single-centre study and our findings may not be applicable to other ICUs. However, the study was conducted in a large tertiary ICU with a patient population typical of other ICUs in high income countries and we may reasonably expect that clinical notes in other ICUs would exhibit similar characteristics. Finally, our observations regarding mortality, although consistent with recent work, challenge conventional wisdom and need to be confirmed or refuted in further studies.

Conclusion

NLP-Dx-BD identified a population of ICU patients expected to also be at high risk of delirium. Moreover, this technique produced a rapid assessment of large amounts of data and enabled the identification of previously poorly understood associations. This approach may open the door to large-scale epidemiological studies of the timing, mode of development, manifestations, severity and duration of behavioural disturbance.

Competing interests

No relevant disclosures.

Author details

Marcus Young[1,2]
Natasha Holmes[1]
Raymond Robbins[1]
Nada Marhoon[1]
Sobia Amjad[1,3]
Ary Serpa Neto[1]
Rinaldo Bellomo[1,2,4,5,6]

1 Data Analytics Research and Evaluation (DARE) Centre, Austin Health and University of Melbourne, Melbourne, VIC, Australia.
3 School of Computing and Information Systems, University of

Melbourne, Melbourne, VIC, Australia.
4 Australian and New Zealand Intensive Care Research Centre, School of Publish Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia.
5 Department of Intensive Care, Austin Hospital, Melbourne, VIC, Australia.
6 Centre for Integrated Critical Care, School of Medicine, University of Melbourne, Melbourne, VIC, Australia.
2 Department of Critical Care, School of Medicine, University of Melbourne, Melbourne, VIC, Australia.

***Correspondence*:** Rinaldo.Bellomo@austin.org.au

References

1 Girard TD, Thompson JL, Pandharipande PP, et al. Clinical phenotypes of delirium during critical illness and severity of subsequent long-term cognitive impairment: a prospective cohort study. *Lancet Respir Med* 2018; 6: 213-22.
2 Ely EW, Shintani A, Truman B, et al. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *J Am Med Assoc* 2004; 291: 1753-62.
3 Marcantonio ER. Delirium in hospitalized older adults. *N Engl J Med* 2017; 377: 1456-66.
4 Wintermann GB, Weidner K, Strauss B, et al. Single assessment of delirium severity during postacute intensive care of chronically critically ill patients and its associated factors: post hoc analysis of a prospective cohort study in Germany. *BMJ Open* 2020; 10: e035733.
5 Pisani MA, Kong SYJ, Kasl S V, et al. Days of delirium are associated with 1-year mortality in an older intensive care unit population. *Am J Respir Crit Care Med* 2009; 180: 1092-7.
6 Witlox J, Eurelings LSM, De Jonghe JFM, et al. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA* 2010; 304: 443-51.
7 Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *Lancet* 2014; 383: 911-22.
8 Andrews PS, Wang S, Perkins AJ, et al. Relationship between intensive care unit delirium severity and 2-year mortality and health care utilization. *Am J Crit Care* 2020; 29: 311-7.
9 Salluh JIF, Wang H, Schneider EB, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. *BMJ* 2015; 350: 1-10.
10 Boustani M, Rudolph J, Shaughnessy M, et al. The DSM-5 criteria, level of arousal and delirium diagnosis: Inclusiveness is safer. *BMC Med* 2014; 12: 1-4.
11 American Psychiatric Association. Diagnostic and statistical manual of mental disorders : DSM-5; 5th ed. Arlington, VA: APA, 2013.
12 Fong TG, Davis D, Growdon ME, et al. The interface between delirium and dementia in elderly adults. *Lancet Neurol* 2015; 14: 823-32.
13 Flaherty JH, Yue J, Rudolph JL. Dissecting delirium: phenotypes, consequences, screening, diagnosis, prevention, treatment, and program implementation. *Clin Geriatr Med* 2017; 33: 393-413.
14 Kotfis K, Marra A, Wesley Ely E. ICU delirium — a diagnostic and therapeutic challenge in the intensive care unit. *Anaesthesiol Intensive Ther* 2018; 50: 128-40.
15 Canet E, Amjad S, Robbins R, et al. Differential clinical characteristics, management and outcome of delirium among ward compared with intensive care unit patients. *Intern Med J* 2019; 49: 1496-504.
16 Ely EW, Margolin R, Francis J, et al. Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *Crit Care Med* 2001; 29: 1370-9.
17 Bergeron N, Dubois MJ, Dumont M, et al. Intensive care delirium screening checklist: evaluation of a new screening tool. *Intensive Care Med* 2001; 27: 859-64.
18 Van Eijk MM, Van Den Boogaard M, Van Marum RJ, et al. Routine use of the Confusion Assessment Method for the Intensive Care Unit: a multicenter study. *Am J Respir Crit Care Med* 2011; 184: 340-4.
19 Reade MC, Eastwood GM, Peck L, et al. Routine use of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) by bedside nurses may underdiagnose delirium. *Crit Care Resusc* 2011; 13: 217-25.
20 Reade MC, Aitken LM: The problem of definitions in measuring and managing ICU cognitive function. *Crit Care Resusc* 2012; 14: 236-43.
21 Ouimet S, Riker R, Bergeon N, et al. Subsyndromal delirium in the ICU: evidence for a disease spectrum. *Intensive Care Med* 2007; 33: 1007-13.
22 Holmes NE, Amjad S, Young M, et al. Using language descriptors to recognise delirium: a survey of clinicians and medical coders to identify delirium-suggestive words. *Crit Care Resusc* 2019; 21: 299-302.
23 Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Informatics Assoc* 2011; 18: 594-600.
24 Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural language processing of clinical notes on chronic diseases: systematic review. *J Med Internet Res* 2019; 21: 1-18.
25 Doing-Harris KM, Weir CR, Igo S, et al. POETenceph — automatic identification of clinical notes indicating encephalopathy using a realist ontology. *AMIA Annu Symp Proc* 2015; 2015: 512-21.
26 Stow PJ, Hart GK, Higlett T, et al. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. *J Crit Care* 2006; 21: 133-41.
27 Bird S, Loper E, Klein E. Natural language processing with Python. O'Reilly Media, 2009.
28 Paul E, Bailey M, Kasza J, et al. The ANZROD model: better benchmarking of ICU outcomes and detection of outliers. *Crit Care Resusc* 2016; 18: 25-36.

29 Paul E, Bailey M, Pilcher D. Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: Development and validation of the Australian and New Zealand Risk of Death model. *J Crit Care* 2013; 28: 935-41.

30 Elizabeth Workman T, Weir C, Rindflesch TC. Differentiating sense through semantic interaction data. *AMIA Annu Symp Proc* 2016; 2016: 1238-47.

31 Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace, 2009.

32 Klouwenberg PMCK, Zaal IJ, Spitoni C, et al. The attributable mortality of delirium in critically ill patients: Prospective cohort study. *BMJ* 2014; 349: 1-10.

33 Sanchez D, Brennan K, Al Sayfe M, et al. Frailty, delirium and hospital mortality of older adults admitted to intensive care: the Delirium (Deli) in ICU study. *Crit Care* 2020; 24: 1-8.

34 de la Cruz M, Fan J, Yennu S, et al. The frequency of missed delirium in patients referred to palliative care in a comprehensive cancer center. *Support Care Cancer* 2015; 23: 2427-33.

35 Soares Pinheiro FGDM, Santana Santos E, Barreto ÍDDC, et al. Mortality predictors and associated factors in patients in the intensive care unit: a cross-sectional study. *Crit Care Res Pract* 2020; 2020: 5-10.

36 Duprey MS, Van Den Boogaard M, Van Der Hoeven JG, et al. Association between incident delirium and 28- and 90-day mortality in critically ill adults: a secondary analysis. *Crit Care* 2020; 24: 1-10.

37 Patel SB, Poston JT, Pohlman A, et al. Rapidly reversible, sedation-related delirium versus persistent delirium in the intensive care unit. *Am J Respir Crit Care Med* 2014; 189: 658-65.