Supplemental information

Combining clinical notes with structured

electronic health records enhances

the prediction of mental health crises

Roger Garriga, Teodora Sandra Buda, João Guerreiro, Jesús Omaña Iglesias, Iñaki Estella Aguerri, and Aleksandar Matić

Supplementary Information

Features in LogReg5
Number of crisis in the last year
Weeks since last crisis
Number of days hospitalized during the last crisis
Maximum length of stay during the last crisis
Maximum crisis severity during the last crisis

Supplementary Table 1. List of features used to build the LogReg5 baseline. Related to STAR Methods - Machine learning prediction models.
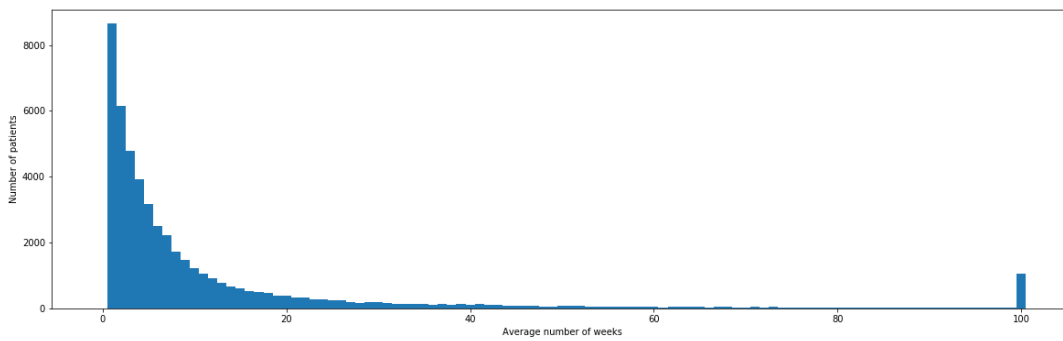
a Number of notes per patient
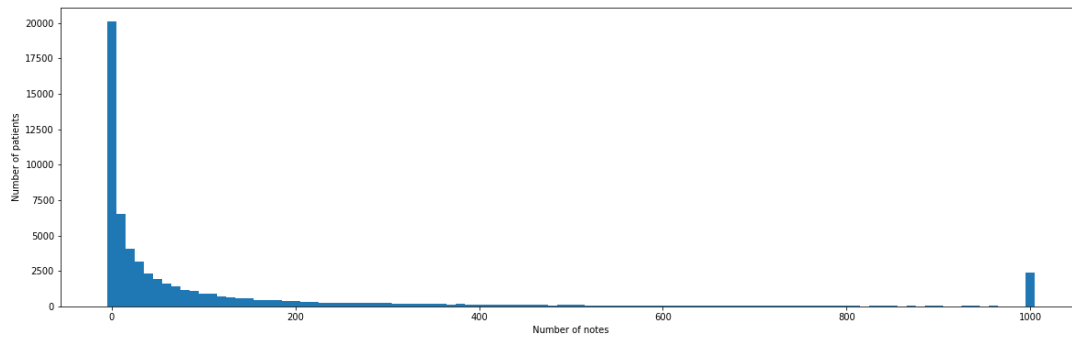
b Number of notes per patient-week
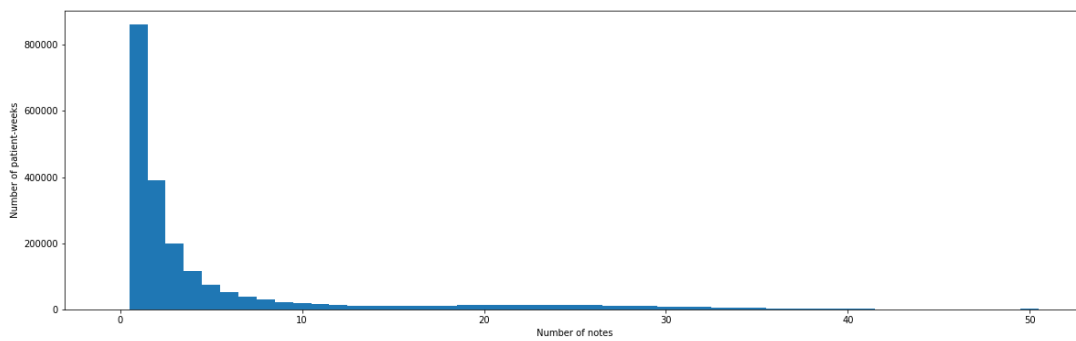
c Average number of weeks between notes per patient

Supplementary Figure 1. Statistics regarding clinical notes, related to Table 1. **a** Histogram of the total number of notes about each patient, for those patients that had at least 1 note. Each bin has width 10 except for the rightmost one which includes all patients with 1000 or more notes. **b** Histogram of the number of notes taken per patient and per week. Each bin has width 1 except for the rightmost one which indlues all patient-weeks with 50 or more notes. **c** Histogram of the average number of weeks between notes taken per patient, for those patients that had notes from at least 2 distinct weeks. Each bin has width 1 except for the rightmost one which includes all patients with an average number of weeks of at least 100 weeks.
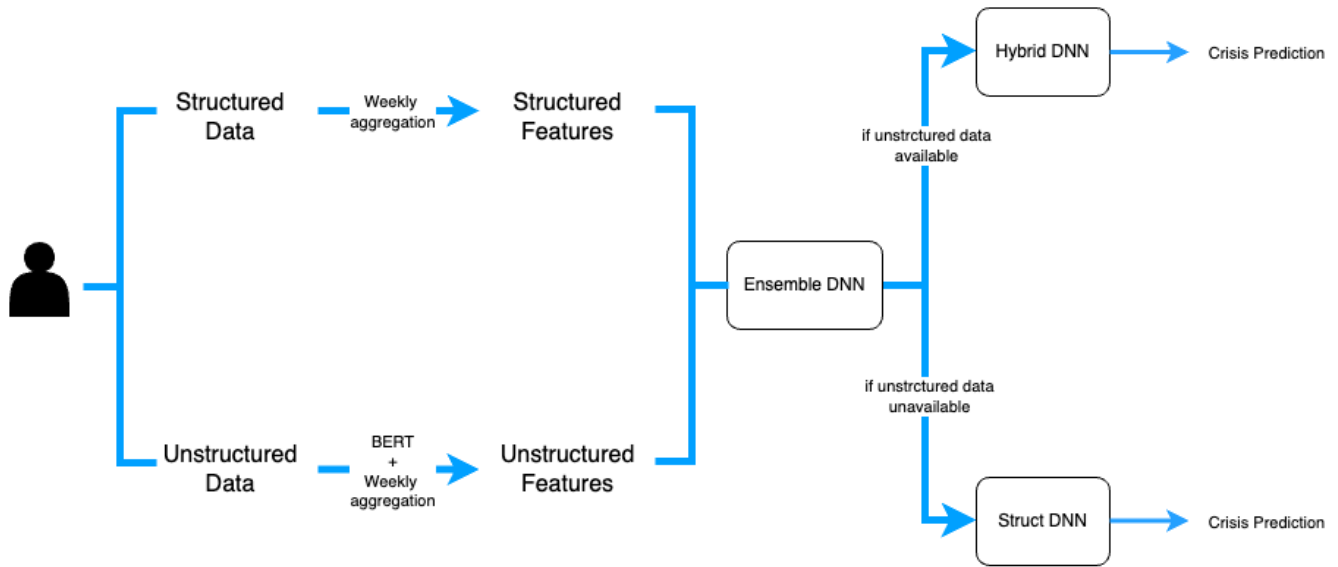
Supplementary Figure 2. Diagram illustrating the inference phase of the Ensemble DNN, detailing how it generates crisis predictions. Related to Figure 1.

Model	Unstruct DNN	Hybrid DNN	Struct XGB	Ensemble DNN
Unstruct DNN	-	0.314	0.260	0.300
Hybrid DNN	-0.314	-	0.107	0.242
Struct XGB	-0.260	-0.107	-	0.160
Ensemble DNN	-0.300	-0.242	-0.160	-

Supplementary Table 2. Net reclassification analysis. Comparison of models based on percentile based Net Reclassification Improvement, using 1000 categories. The model specified in the row header serves as the reference. Related to Table 2

Perc. of weeks with unstruc. data	Number of patients (%)	Crisis episodes		Prevalence (%)	
		Train	Test	Train	Test
(0-10]	41,392 (69.3)	48,660	8,161	0.856	0.592
(10-20)	8,507 (14.2)	16,118	3,250	2.818	2.072
(20-30]	4,011 (6.7)	9,180	2,034	3.921	3.082
(30-40]	2,344 (3.9)	6,783	1,323	4.975	3.646
(40-50]	1,363 (2.3)	4,523	861	5.769	4.395
(50-100]	2,133 (3.6)	8,615	1,468	8.733	5.773

Supplementary Table 3. Number of patients, crisis episodes and target prevalence per subgroup of patients based on the percentage of weeks with unstructured data. Related to Figure 2.

Note subgroups	Mean AUPRC (std)			t-statistic (p-value)
	Struc XGB	Unstruc DNN	Hybrid DNN	Struc XGB vs Hybrid DNN
≤ 10%	**0.048 (0.012)***	0.032 (0.008)	0.032 (0.007)	13.196 (<0.001)
(10%, 20%]	0.108 (0.021)	0.074 (0.014)	**0.121 (0.023)***	-5.203 (<0.001)
(20%, 30%]	0.150 (0.036)	0.097 (0.023)	**0.160 (0.035)***	-3.468 (0.001)
(30%, 40%]	0.176 (0.043)	0.115 (0.031)	**0.185 (0.050)***	-2.821 (0.007)
(40%, 50%]	0.198 (0.047)	0.126 (0.038)	**0.223 (0.059)***	-4.509 (<0.001)
> 50%	0.236 (0.049)	0.150 (0.034)	**0.247 (0.045)***	-3.583 (0.001)

Supplementary Table 4. AUPRC comparison between the models trained in structured (Struc XGB), unstructured (Unstruc DNN) or both (Hybrid DNN) per each note subgroup. (* indicates a statistically significant difference <0.05). Related to Figure 2.

Note subgroups	Mean AUROC (std)			DeLong statistic (p-value)
	Struc XGB	Unstruc DNN	Hybrid DNN	Struc XGB vs Hybrid DNN
≤ 10%	**0.784 (0.012)***	0.743 (0.024)	0.738 (0.024)	24.276 (<0.001)
(10%, 20%]	0.777 (0.019)	0.715 (0.023)	**0.778 (0.020)**	-0.953 (0.341)
(20%, 30%]	0.772 (0.023)	0.705 (0.025)	**0.775 (0.021)***	-1.986 (0.047)
(30%, 40%]	0.770 (0.033)	0.704 (0.035)	**0.773 (0.029)***	-2.033 (0.042)
(40%, 50%]	0.762 (0.033)	0.700 (0.033)	**0.765 (0.034)**	-1.260 (0.208)
> 50%	0.772 (0.027)	0.692 (0.034)	**0.781 (0.028)***	-4.703 (<0.001)

Supplementary Table 5. AUROC comparison between the models trained in structured (Struc XGB), unstructured (Unstruc DNN) or both (Hybrid DNN) per each note subgroup. (* indicates a statistically significant difference <0.05). Related to Figure 2.

	Target prevalence %	Crises occurred %	Crises flagged %	Crises detected %
Gender				
Male	1.3%	51.3%	51.7%	52.7%
Female	1.3%	48.6%	48.2%	47.1%
Ethnic group				
White	1.4%	66.8%	68.7%	69.3%
Asian	1.3%	14.0%	12.9%	12.8%
Black	1.6%	8.4%	9.3%	8.8%
Mixed	1.5%	6.4%	6.4%	6.5%
Not known	0.7%	4.4%	2.4%	2.7%

Supplementary Table 6. Evaluation of the Ensemble DNN model's performance by gender and ethnicity using the following metrics: the percentage of correctly identified crisis episodes, the percentage of crisis flagged by the algorithm, and the comparison of crisis incidents per subgroup. This assessment considers the top 1000 patients per week, ranked by predicted risk score. Related to Table 2.

	Target prevalence %	AUROC	AUPRC
Gender			
Male	1.3%	0.865	0.130
Female	1.3%	0.865	0.140
Ethnic group			
White	1.4%	0.866	0.138
Asian	1.3%	0.853	0.133
Black	1.6%	0.859	0.153
Mixed	1.5%	0.854	0.143
Not known	0.7%	0.882	0.129

Supplementary Table 7. Evaluation of the Ensemble DNN model's performance by gender and ethnicity using AUROC and AUPRC. Related to Table 2.