Discovering Social Determinants of Health from Case Reports using Natural Language Processing: Algorithmic Development and Validation

Shaina Raza [1,2*]; Elham Dolatabadi [1,3], Nancy Ondrusek[1,2], , Laura Rosella[2], Brian Schwartz[1,2]

[1] Public Health Ontario (PHO), Toronto, ON, Canada.
[2] Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada.
[3] Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, ON, Canada.

Corresponding Author:
Shaina Raza, PhD
Email: shaina.raza@oahpp.ca

Appendix 1

Table S1: *Named entities*

Clinical entities	Non-clinical entities (including SDoHs)
1. HEART_DISEASE 2. CLINICAL_DEPARTMENT 3. BLOOD_PRESSURE 4. DISEASE_SYNDROME 5. DOSAGE 6. TREATMENT 7. TEST 8. PSYCHOLOGICAL_CONDITION 9. SYMPTOM 10. RESPIRATION 11. INTERNAL_ BODY_PART 12. EXTERNAL_BODY_PART 13. ONCOLOGICAL 14. PROCEDURE 15. DIABETES 16. DRUG_NAME 17. VACCINE 18. HYPERLIPIDEMIA 19. HYPERTENSION 20. DEATH_ENTITY 21. RESPIRATION 22. KIDNEY_DISEASE 23. OBESITY 24. BMI 25. PULSE 26. INJURY_OR_POISONING	27. GENDER 28. HEIGHT 29. AGE 30. DATE 31. RACE_ETHNICITY 32. EMPLOYMENT 33. SMOKING 34. ADMISSION_DISCHARGE 35. ALCOHOL 36. SUBSTANCE 37. WEIGHT 38. RELATIVE_DATE 39. DURATION 40. TIME

Figure S1: Named entities extracted from the case report.

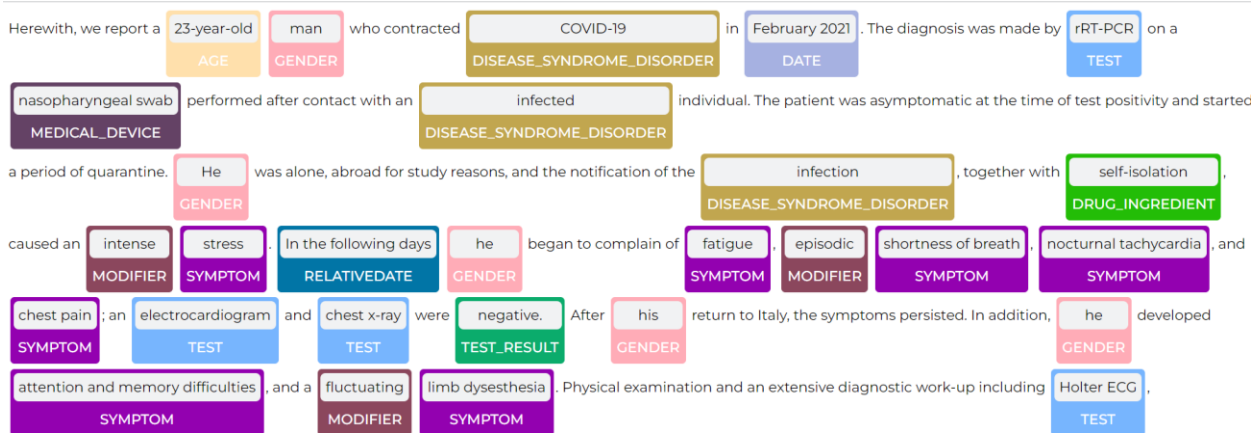
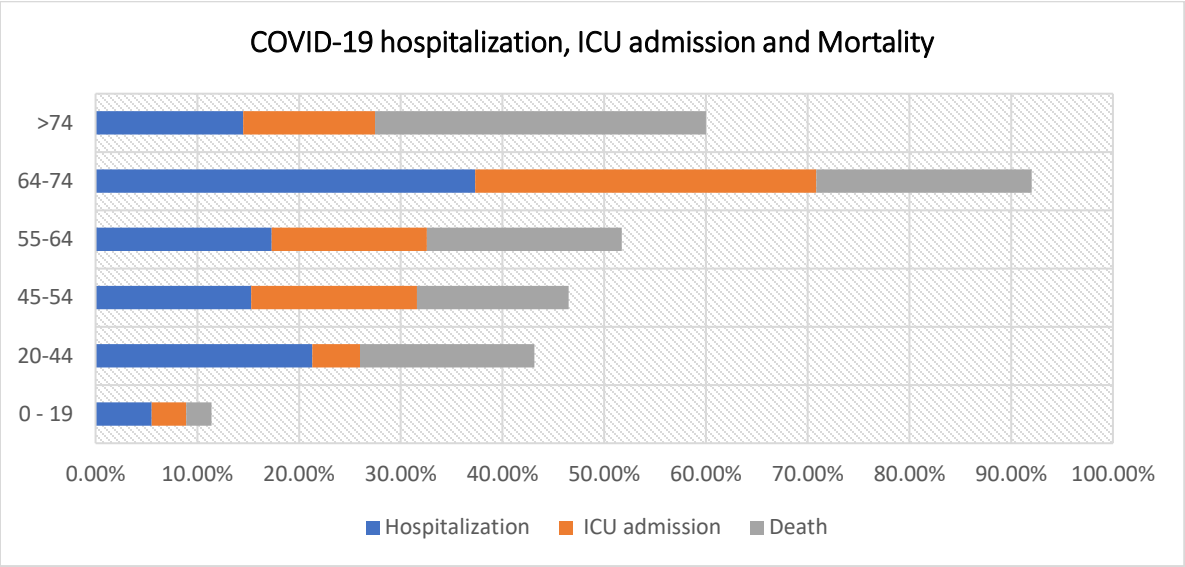


Table S2: *General Hyperparameters used along with best value and range in parenthesis.*

Hyperparameter	Value
LSTM state size	200 [200 - 300]
dropout rate	0.5 [0.2 - 0.7]
Epochs	40 [20- 80]
Batch size	16 [8 - 128]
Learning rate (lr)	1.e-05 [1.e-9 – 1.e-2]
lr decay coefficient (po)	0.005 [0.001, 0.01]
Warmup steps	10,000
Optimizer	ADAM [58], β1=0.9 and β2=0.999
Word dimension	300 [50 – 450]
Hidden size LSTM	300
Gradient clipping	5.0
The fine-tuning transformer-based architectures (BioBERT and others) are maximum sequence length of 128, number of layers as 12, number of attention heads also 12 and embedding size as 768. For different datasets, the fine-tuning takes different hours (2 hours, 3 hours, 4 hours and 10 hours for our dataset). In the NER task, we fixed the length of sentences to 512. Other hyperparameters used in the paper are given in Table 2.	

Figure S2: Hospitalization, ICU admission, and morality in COVID-19 patients with different age groups.



In Figure S2, we observe that, while the mortality rate in older adults (age>74) is 32.50%, the hospitalization and ICU admission rates (14.50% and 12.98%, respectively) are lower than in other age groups.