A comparative analysis of sepsis digital phenotyping methods

ANNA FEDYUKOVA, DANIEL CAPURRO*, and DOUGLAS PIRES†*, University of Melbourne

Health data captured in Electronic health records (EHRs) have enabled the development of computational approaches to improve patient management and treatment, including early diagnosis of severe conditions such as sepsis. The validity of these efforts, however, largely relies on which sepsis definition is used and the quality of the underlying data. Here we tested different sepsis definitions to better understand how phenotyping approaches may impact the classification accuracy of sepsis prediction algorithms.

To assess the extent to which sepsis definitions (dis)agree with each other, we have analised a large cohort of patients admitted to the ICU (over 22,000) from MIMIC-IV. Cases were classified as septic and non-septic using the Sepsis-3 definition as a standard and compared with different ICD-10-based sepsis phenotyping criteria.

Most of administrative sepsis definitions agreed with each other when identifying positive sepsis cases. At the same time, we identified considerable disagreement between Sepsis-3 and administrative definitions. This discrepancy affected machine learning algorithms' predictive performance. Two algorithms out of three built on Sepsis-3 outperformed models based on other phenotypes. Experiments demonstrate that phenotype definitions can significantly influence a predictive model performance. This highlights the importance of consistent and validated digital phenotyping criteria.

CCS Concepts: • **Computing methodologies → Machine learning**; **Causal reasoning and diagnostics**; • **Applied computing → Health informatics**.

Additional Key Words and Phrases: digital phenotyping, supervised learning, sepsis, MIMIC-IV

ACM Reference Format:
Anna Fedyukova, Daniel Capurro, and Douglas Pires. 2020. A comparative analysis of sepsis digital phenotyping methods. In *Proceedings of the Australasian Computer Science Week (ACSW) 2021.* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Sepsis is a life-threatening organ dysfunction caused by a deregulated host response to infection [16], highly frequent in intensive care units and a high mortality (ICU) [15, 18]. [14] showed that sepsis-related deaths comprise 19.7% (11 million) of all global deaths in 2017 which is as twice as originally thought.

Sepsis is a treatable condition, however delays in the initiation of antibiotic therapy can lead to increased mortality even among patients who received antibiotics within 6 hours [10]. However, attempts to detect potential cases of sepsis earlier and reduce the time to treatment initiation can also generate unintended consequences. Studies have shown that fewer than 60% of patients with suspected sepsis end up having a definite or probable infection that benefited from antibiotic therapy [9]. Sepsis treatment is complicated by the fact that it is a syndrome and, in many cases, can be caused by other non-infectious inflammatory disorders [13].

[*] Co-senior authors
[†] Co-senior authors

Recent advances in predictive modelling and the increasing availability of Electronic Health Records (EHR) have led to the development and deployment of algorithms to predict the onset of sepsis [2].

As sepsis is a broad term applied to an incompletely understood process [16], researchers have used a diversity of sepsis definitions to guide the development of predictive models. This heterogeneity can complicate the comparison between the performance of proposed models. We hypothesize that heterogeneous sepsis definitions can significantly modify the accuracy of different predictive models.

The goal of this study is to quantify the extent to which different sepsis phenotyping definitions (dis)agree with each other and to assess the effect of heterogeneous sepsis definitions on the accuracy of sepsis predictive models on a large cohort of patients admitted to the ICU.

2 METHODS

2.1 Context

In February 2016, The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) were published [16]. Unlike previous definitions, Sepsis-3 is based on the coexistence of an infection and organ dysfunction. According to that definition, sepsis is diagnosed when there is an increase of more than 2 points in the total sequential organ failure assessment (SOFA) score [3]. This definition associates sepsis with more severe conditions than those diagnosed by previous definitions [3]. Moreover, this new definition requires identification of a time point at which the patient may be septic that can be useful for retrospective assessment of the trajectory of the patient's illness [8].

Several approaches for labelling patients as septic using EHR data have been published [1], that we propose to investigate. The Danish ICD-10 sepsis definition (DK) includes nineteen diagnostic codes [11], while [4] proposed two Swedish sepsis definitions, Narrow (SE Narrow) and Wide (SE Wide) (the former more restrictive) assuming most patients treated in ICUs have severe infections in their discharge diagnoses. We also considered the Australian (AU) [6, 7, 17] and Canadian sepsis definition (CA), the latter based on ICD-10-CA/CCI (Canadian ICD-10 and the Canadian Classification of Intervention) [12].

2.2 Data collection and inclusion criteria

For this study, we obtained data from the MIMIC-IV database [5]. This database is a de-identified database from patients admitted to critical care units or the emergency department between 2008 – 2019 at the Beth Israel Deaconess Medical Center (BIDMC, Boston Massachusetts, USA). It contains administrative data, including patient demographics, and clinical information including vital signs, laboratory tests, medication administration, diagnoses and procedures.

We included patients older than 18 who met the Sepsis-3 definition and whose episodes were coded using the ICD-10 classification. 69,619 admissions met the sepsis definition. We excluded 7,107 cases without any ICD codes, and 32,525 cases only coded ICD-9 terms. Since our task involved predicting the future development of sepsis, we also excluded the patients with a very early suspicion of infection, defined as patients with cultures ordered or antibiotics administered before or up to 24 hours of being admitted to the ICU. The final cohort consisted of 22,090 ICU stays.

2.3 Reference standard

Sepsis-3 [16] was chosen as the reference standard for sepsis definition as it better aligns with the contemporary understanding of the pathophysiology of sepsis [8]. In addition, Sepsis-3 criteria are independent from ICD-10 codes [8].

Table 1. Measurement of sepsis prediction performance

	Sepsis-3		AU		DK		SE Wide		SE Narrow		CA	
	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
LR	0.80	0.60	0.82	0.64	**0.83**	**0.67**	0.76	0.54	0.82	0.63	0.83	0.66
RF	**0.82**	**0.64**	0.78	0.56	0.80	0.61	0.66	0.33	0.78	0.57	0.80	0.61
GB	**0.82**	**0.64**	0.78	0.55	0.79	0.59	0.70	0.41	0.77	0.55	0.80	0.60

LR=Logistic Regression, RF=Random Forest, GB=Gradient Boosting.

From September 2020 MIMIC-IV was supplemented by an additional group of tables containing a sepsis cases classified according to the Sepsis-3 definition, thus generating an external label for sepsis cases.

2.4 Study design

The study consisted of two major steps. First, to understand the level of (dis)agreement between different sepsis definitions we analysed the proportion of shared positive cases among over 22,000 ICU admissions.

The second step was to build sepsis prediction models via supervised learning using the different sepsis definitions as the ground truth and evaluate the effect of the definitions on predictive performance.

First 24 hours clinical data was used as the evidence to train the predictive models, including vital sings, laboratory results, demographics, and contextual features. Six balanced data sets with equal sets of features but different sepsis definitions as reference standards were used for building of prediction models as training data. Each data set contains 2,194 data points that were chosen randomly. The number of positive cases was determined based on the maximum available positive cases for the Sepsis-3 definitions. Two data sets consisting of 412 positive and 412 negative cases were used for evaluation of the models' performance. The first one included common positive cases between five of the six definitions (412). SE Wide only shared 56 positive cases with other definitions and had to be tested on customised data set. This data set included 56 positive cases which are common for all definitions and 356 positive cases which are positive only for SE Wide.

Three machine learning techniques were evaluated: Logistic Regression (LR), Gradient Boosting (GB) and Random Forest (RF). The area under the ROC curve AUROC and Matthews Correlation Coefficient MCC were used to evaluate the performance of the models under 10-fold cross-validation and on blind tests.

3 RESULTS

3.1 Analysis of sepsis definitions

We started by analysing the percentage of sepsis-positive and sepsis-negative cases for each definition. The highest proportion of septic patients (5,165 incidences or 23%) among all of the analysed ICU stays was generated by SE Narrow definition. This is one and a half times higher than the mean share of septic patients for AU (3,059 cases) and DK (3,068 cases) ICD-10 definitions, 30% higher than CA definition (3,767), 3.2 times higher than for SE Wide and 4 times higher in comparison with Sepsis-3 definition (1,097 positive cases).

Analysis of overlapping data among all subsets showed that only 56 sepsis cases are common for all six data sets. However, closer inspection of the data shows that five out of six definitions, except for SE Wide, have a 'common core' of 412 sepsis incidences.

We built confusion matrices to compare each of the five administrative approaches against the Sepsis-3 definition and against each other (Figure 1). Both SE Wide and SE Narrow definitions have large unique proportion of cases.

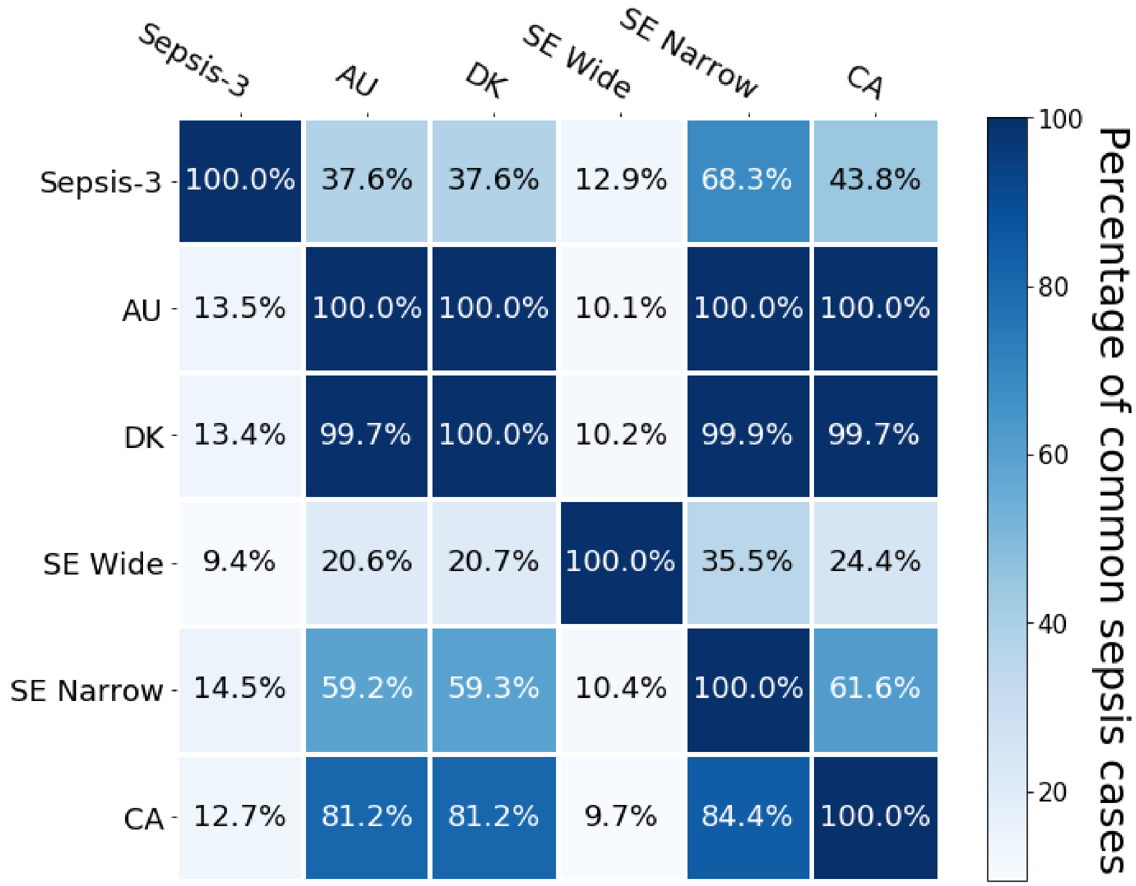

Fig. 1. Confusion matrix of administrative coding approaches against Sepsis-3 definition as the ground true.

For SE Wide definition this percentage is 58.6% (883 cases among 1,507) From the data, it is apparent that SE Wide demonstrates a lot of disagreement with the other ICD-10 sets. For SE Narrow the ratio of unique cases is 28,9% (1,494 unique cases out of 5,165). For Sepsis-3 the proportion of unique cases is also relatively high (23,8% or 261 cases among 1,097).

AU, DK, CA and SE Narrow definition, on the other hand, presented a significant overlap, with 2,647 common sepsis cases, with the AU subset fully covered by the DK, CA and SE Narrow subsets.

AU, DK and CA phenotypes demonstrate even stronger consistency in terms of positive cases. SE Narrow on the one hand has 51.25% of common cases with the three aforementioned definitions, with 29% of unique instances.

In terms of consistency with Sepsis-3 definition the highest agreement is shown by SE Narrow with 68.3% of common cases (Figure 1), followed by CA, which includes 43.8% of Sepsis-3 cases. AU and DK subsets include 37.6% of Sepsis-3 cases each. But only 12.9% of SE Wide positive cases are also covered by the Sepsis-3 definition.

A comparison of positive Sepsis-3 cases included into other phenotypes demonstrates that SE Narrow includes 14,5% of incidents. AU, DK and CA contain 13.5%, 13.4% and 12.7% of positive Sepsis-3 cases respectively. SE Wide phenotype has only 9.4% of Sepsis-3 cases.

Analysis of confusion matrices considering Sepsis-3 as a ground truth shows that performance of SE Narrow definition has the highest AUROC of 0.74, but relatively low MCC of 0.24. AU, DK and CA data sets have a considerable proportion of common cases and, therefore, presented similar performance. A AUROC of 0.62 was obtained for both AU and DK phenotypes and 0.64 for CA. MCC of AU, DK and CA was 0.16. The lowest AUROC was obtained by the SE Wide definition.

3.2 Analysis of predictive model performance based on different sepsis definitions

The results of supervised learning models (Table 1) show a large variation in predictive performance, that seems significantly depended on sepsis definition, with models based on Sepsis-3 on average outperforming other definitions, based on AUROC (an average of 0.813 AUROC).

Only the LR classifier built on the data labelled according to DK, CA and AU definitions presented slightly higher MCC then the models using Sepsis-3 definition. Models based on SE Wide and SE Narrow data achieved lower MCC and AUROC in comparison with the remaining models.

Performance of LR with Sepsis-3 data set is relatively lower because this model assigns more weights to ‘mean blood pressure’ and ‘bilirubin total min’ features than the other models. Values of these features do not have strong correlation with sepsis for some of patients’ records, resulting in misclassification of additional 26 false negative and 14 false positive cases that were correctly identified by the other models.

Depending on the definition, AUROC varies between 0.76 and 0.83 for LR, for RF the AUROC is in the range between 0.66 and 0.82, for GB 0.70-0.82. MCC for LR varies between 0.54 and 0.67, for RF 0.33 - 0.64 and for GB it is between 0.41 and 0.64.

4 LIMITATIONS

This study is based on data set collected from a single hospital which may not be representative of other medical institutions with different documentation practices. Analysis including a broader range of institutions and geographic locations may lead to better generalization of conclusions.

5 CONCLUSION

Here we presented a large scale evaluation of sepsis digital phenotyping approaches. The experiments with blind data set showed that models based on Sepsis-3 definition slightly overperform models built on other phenotypes, except for LR.

However, there is a significant disagreement between Sepsis-3 and administrative definitions. The share of sepsis cases related to administrative definitions which are included into Sepsis-3 does not exceed 44% for the majority of definitions. The percentage of Sepsis-3 cases included into other data sets is not more than 14.5%.

Further analysis of ICD-10 sets of codes has shown that AU, DK and CA definitions are fairly consistent with each other. DK and CA have more than 80% of common positive cases, while AU positive cases appeared to be completely included into CA, DK and SE Wide definitions.

SE Wide contains the broadest definition of sepsis, with the largest number of positive cases and shares many common cases with other administrative phenotypes and the largest number of unique cases. Models based on this definition show the lowest performance regardless of the chosen machine learning algorithm.

SE Narrow criteria has the highest agreement with Sepsis-3 and identifies the largest proportion of positive cases and could be the preferred method to find sepsis patients using administrative data.

The SE Wide sepsis phenotype shows the highest inconsistency with other definitions and has the lowest scores among all administrative classification approaches. Therefore, this classification approach should be used with a high degree of caution.

The results of this investigation show that most of administrative sepsis definitions, except for one, agree with each other in terms of positive sepsis cases definition. However, they significantly disagree with Sepsis-3, which is independent from ICD-10 codes. AUROC and MCC evaluation scores can increase up to 16 and 31 points respectively contingent on the definition and algorithm. A careful selection of robust and validated digital phenotypes is a key step in the development of predictive algorithms as it can heavily influence their performance.

REFERENCES

[1] Mette K Beck, Anders Boeck Jensen, Annelaura Bach Nielsen, Anders Perner, Pope L Moseley, and Søren Brunak. 2016. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Scientific reports* 6, 1 (2016), 1–9.

[2] Jacob S Calvert, Daniel A Price, Uli K Chettipally, Christopher W Barton, Mitchell D Feldman, Jana L Hoffman, Melissa Jay, and Ritankar Das. 2016. A computational approach to early sepsis detection. *Computers in biology and medicine* 74 (2016), 69–73.

[3] Seitaro Fujishima. 2016. Organ dysfunction as a new standard for defining sepsis. *Inflammation and Regeneration* 36, 1 (2016), 24.

[4] Rolf Gedeborg, Mia Furebring, and Karl Michaëlsson. 2007. Diagnosis-dependent misclassification of infections using administrative data variably affected incidence and mortality estimates in ICU patients. *Journal of clinical epidemiology* 60, 2 (2007), 155–e1.

[5] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.

[6] Toni Henderson, Jennie Shepheard, and Vijaya Sundararajan. 2006. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Medical care* (2006), 1011–1019.

[7] Irwani Ibrahim, Ian G Jacobs, Steven AR Webb, Judith Finn, et al. 2012. Accuracy of International classification of diseases, 10th revision codes for identifying severe sepsis in patients admitted from the emergency department. *Critical Care and Resuscitation* 14, 2 (2012), 112.

[8] Alistair EW Johnson, Jerome Aboab, Jesse D Raffa, Tom J Pollard, Rodrigo O Deliberato, Leo Anthony Celi, and David J Stone. 2018. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine* 46, 4 (2018), 494.

[9] Peter MC Klein Klouwenberg, Olaf L Cremer, Lonneke A van Vught, David SY Ong, Jos F Frencken, Marcus J Schultz, Marc J Bonten, and Tom van der Poll. 2015. Likelihood of infection in patients with presumed sepsis at the time of intensive care unit admission: a cohort study. *Critical Care* 19 (2015), 319.

[10] Vincent X Liu, Vikram Fielding-Singh, John D Greene, Jennifer M Baker, Theodore J Iwashyna, Jay Bhattacharya, and Gabriel J Escobar. 2017. The timing of early antibiotics and hospital mortality in sepsis. *American journal of respiratory and critical care medicine* 196, 7 (2017), 856–863.

[11] Kreesten Meldgaard Madsen, Henrik Carl Schønheyder, Brian Kristensen, Gunnar Lauge Nielsen, and Henrik Toft Sørensen. 1998. Can hospital discharge diagnosis be used for surveillance of bacteremia? A data quality study of a Danish hospital discharge registry. *Infection Control & Hospital Epidemiology* 19, 3 (1998), 175–180.

[12] Hude Quan, Cathy Eastwood, Ceara Tess Cunningham, Mingfu Liu, Ward Flemons, Carolyn De Coster, William A Ghali, IMECCHI investigators, et al. 2013. Validity of AHRQ patient safety indicators derived from ICD-10 hospital discharge abstract data (chart review study). *BMJ open* 3, 10 (2013).

[13] Chanu Rhee, Shruti Gohil, and Michael Klompas. 2014. Regulatory mandates for sepsis care—reasons for caution. *New England Journal of Medicine* 370, 18 (2014), 1673–1676.

[14] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 395, 10219 (2020), 200–211.

[15] Yasser Sakr, Ulrich Jaschinski, Xavier Wittebole, Tamas Szakmany, Jeffrey Lipman, Silvio A Ñamendys-Silva, Ignacio Martin-Loeches, Marc Leone, Mary-Nicoleta Lupu, Jean-Louis Vincent, et al. 2018. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. In *Open forum infectious diseases*, Vol. 5. Oxford University Press US, ofy313.

[16] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama* 315, 8 (2016), 801–810.

[17] Vijaya Sundararajan, Christopher M MacIsaac, Jeffrey J Presneill, John F Cade, and Kumar Visvanathan. 2005. Epidemiology of sepsis in Victoria, Australia. *Critical care medicine* 33, 1 (2005), 71–80.

[18] Jean-Louis Vincent, Yasser Sakr, Charles L Sprung, V Marco Ranieri, Konrad Reinhart, Herwig Gerlach, Rui Moreno, Jean Carlet, Jean-Roger Le Gall, Didier Payen, et al. 2006. Sepsis in European intensive care units: results of the SOAP study. *Critical care medicine* 34, 2 (2006), 344–353.