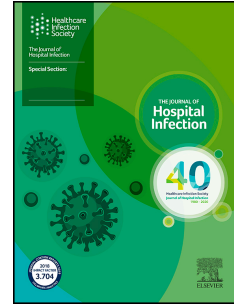# Journal Pre-proof

A machine learning approach to predict healthcare-associated infections at intensive care unit admission: findings from the SPIN-UTI project

Martina Barchitta, Andrea Maugeri, Giuliana Favara, Paolo Marco Riela, Giovanni Gallo, Ida Mura, Antonella Agodi, on behalf of the SPIN-UTI network

Please cite this article as: Barchitta M, Maugeri A, Favara G, Riela PM, Gallo G, Mura I, Agodi A, on behalf of the SPIN-UTI network, A machine learning approach to predict healthcare-associated infections at intensive care unit admission: findings from the SPIN-UTI project, *Journal of Hospital Infection*, https://doi.org/10.1016/j.jhin.2021.02.025.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1  **A machine learning approach to predict healthcare-associated infections at intensive**

2  **care unit admission: findings from the SPIN-UTI project**

3

4  Martina Barchitta[1,2], Andrea Maugeri[1,2], Giuliana Favara[1], Paolo Marco Riela[3], Giovanni

5  Gallo[3], Ida Mura[2,4], Antonella Agodi[1,2*], on behalf of the SPIN-UTI network[#]

6

7  [1] Department of Medical and Surgical Sciences and Advanced Technologies "GF Ingrassia",

8  University of Catania, Catania, Italy

9  [2] GISIO-SItI - Italian Study Group of Hospital Hygiene - Italian Society of Hygiene,

10 Preventive Medicine and Public Health, Italy

11 [3] Department of Mathematics and Informatics, University of Catania, Catania, Italy

12 [4] Department of Biomedical Sciences, University of Sassari, Sassari, Italy

13

14 [*] Corresponding author: Antonella Agodi, Department of Medical and Surgical Sciences and

15 Advanced Technologies "GF Ingrassia", University of Catania, Catania, Italy. Via S. Sofia,

16 87 – 95123 Catania, Italy. agodia@unict.it

17

18 **Running title:** Machine learning for infection prediction

19

20

21

22

23

24

25

26  **Summary**

27  **Background:** Identifying patients at higher risk of healthcare-associated infections (HAIs) in

28  intensive care unit (ICU) represents a major challenge for public health. Machine learning

29  could improve patient risk stratification and lead to targeted infection prevention and control

30  interventions.

31  **Aim:** To evaluate the performance of the Simplified Acute Physiology Score (SAPS) II for

32  HAIs risk prediction in ICUs, using both traditional statistical and machine learning

33  approaches.

34  **Methods:** We used data of 7827 patients from the "Italian Nosocomial Infections

35  Surveillance in Intensive Care Units" project.  The Support Vector Machines (SVM)

36  algorithm was applied to classify patients according to sex, patient origin, non-surgical

37  treatment for acute coronary disease, surgical intervention, SAPS II at admission, presence of

38  invasive devices, trauma, impaired immunity, antibiotic therapy in 48 hours before ICU

39  admission.

40  **Findings:** The performance of SAPS II for predicting the risk of HAIs provides a ROC

41  (Receiver Operating Characteristics) curve with an AUC (Area Under the Curve) of 0.612

42  (p<0.001) and an accuracy of 56%. Considering SAPS II along with other characteristics at

43  ICU admission, we found an accuracy of the SVM classifier of 88% and an AUC of 0.90

44  (p<0.001) for the test set. In line, the predictive ability was lower when considering the same

45  SVM model but removing the SAPS II variable (accuracy= 78% and AUC= 0.66).

46  **Conclusions:** Our study suggested the SVM model as a tool to early predict patients at higher

47  risk of HAI at ICU admission.

48

49  **Keywords:** healthcare-associated infections; machine learning; intensive care unit; risk

50  prediction

51

52 **Abbreviations**

53 Healthcare Associated Infection (HAI)

54 Intensive Care Unit (ICU)

55 European Center for Disease Prevention and Control (ECDC)

56 Simplified Acute Physiology Score (SAPS II)

57 Italian Nosocomial Infections Surveillance in Intensive Care Units (SPIN-UTI)

58 Italian Study Group of Hospital Hygiene (GISIO)

59 Italian Society of Hygiene, Preventive Medicine and Public Health (SItI)

60 Support Vector Machines (SVM)

61 K-Nearest Neighbor (K-NN)

62 Synthetic Minority Over-sampling Technique (SMOTE)

63 Radial basis function Kernel (RBF)

64 Interquartile range (IQR)

65 Receiver Operating Characteristics (ROC)

66 Area under the curve (AUC)

67 Acute Physiologic Assessment and Chronic Health Evaluation II (APACHE II)

68 Sequential Organ Failure Assessment (SOFA)

69

70

71

72

73

74

75

## Introduction

Healthcare-associated Infections (HAIs) are one of the major threats for public health worldwide, due to their significant impact on mortality, hospital stays, and assistance costs [1-3]. In particular, frequency of HAIs is higher among people staying in Intensive Care Units (ICUs), because they have more severe clinical conditions, they are often immune-compromised, and more likely to be intubated and catheterized than those staying in other hospital wards [4,5]. Furthermore, high antibiotic resistance rates have been reported together with increasing trends of resistant microorganisms, highlighting the need for continuous comprehensive strategies targeting not only the prudent use of antibiotics, but also infection control measures to control the epidemic spread of resistant isolates, especially in ICUs [3,6-9]. As reported by the European Center for Disease Prevention and Control (ECDC), in 2017 on a total of approximately 143,000 patients staying in ICU, 8% presented at least one HAI on a given day. In line, among ICU-surveilled HAIs, pneumonia, bloodstream infection and urinary tract infections accounted for 6%, 4% and 2%, respectively [10].

Although HAIs depend on microorganisms' characteristics - such as infectivity, pathogenicity, modes of transmission – several patients' characteristics and the inappropriate use of invasive devices during the hospital stay represent some of the leading causes of HAIs in all the hospital wards, and especially in ICUs [4,11,12]. In the last decades, several early warning scores have been developed in clinical practice to measure health conditions or illness severity of ICU patients. In particular, the Simplified Acute Physiology Score (SAPS) II represents the most widely used instrument for the prediction of prognosis, HAIs risk, sepsis and mortality [13-17]. This validated score is calculated considering twelve routine physiological variables collected during the first hours of ICU admission, not including the type admission [18,19]. For these reasons, the identification of patients at higher risk of HAIs in

101 ICU still remains a major challenge for public health, with so many healthcare professionals

102 which have studied and continue to examine personal and clinical characteristics associated

103 with HAI risk [20-25]. In this scenario, recent advances in statistical and mathematical

104 approaches to automatically learn from a given dataset have made possible to identify

105 patients or subgroups of patients which are more likely to be affected by HAI during their

106 hospital stay [26-28]. Indeed, there is a strong need for reliable tools that can guide patient

107 management [29] by predicting the risk of HAIs and adverse associated outcomes, and thus

108 reducing their burden on healthcare systems [30, 31]. Furthermore, the availability of large

109 amount of patient and facility data and the appropriate application of machine learning

110 methods in healthcare epidemiology could help the understanding of risk factors for HAIs,

111 the development of patient risk stratification tools and the identification of pathways for the

112 spread of infections. This, in turn, could lead to targeted prevention interventions [32].

113 Particularly, machine learning has been proposed to predict specific adverse events and for

114 risk stratification in the ICU, becoming a useful way to improve quality of care [33].

115

116 Here, we aimed to identify and predict patients at higher risk of HAIs, according to their

117 characteristics at ICU admission. To do that, we used data from the "Italian Nosocomial

118 Infections Surveillance in Intensive Care Units" (SPIN-UTI) project, which was established

119 by the Italian Study Group of Hospital Hygiene (GISIO) of the Italian Society of Hygiene,

120 Preventive Medicine and Public Health (SItI) in 2006. The SPIN-UTI network, since then,

121 has collected data related to approximately 20,000 patients, more than 4,300 infections and

122 5,300 microorganisms [20-25, 34]. Our hypothesis is that, in the framework of predictive and

123 personalized medicine, machine learning algorithms could enrich conventional statistical

124 approaches, especially in terms of prediction of ICU prognosis, clinical deterioration and risk

125 assessment [35]. Accordingly, the current study first evaluates the performance of SAPS II for

126 HAI's risk prediction in ICUs using a traditional statistical method. Next, we applied a

127 Support Vector Machines (SVM) algorithm, considering SAPS II in combination with

128 additional features at ICU admission, in order to distinguish non-infected patients from those

129 who were diagnosed with at least one HAIs during their ICU stay and thus contribute to

130 efforts to enhance patient management by ensuring better prevention, prognosis and therapy.

131 **Methods**

132 *Study design and data collection*

133 In the current study, we used data collected during the seven editions of the SPIN-UTI project

134 according to the ECDC protocol [36]. From 2006 to 2019, the SPIN-UTI project prospectively

135 surveyed 20,060 patients staying in ICUs for more than 2 days, recording data at patient, ICU

136 and hospital levels during their stay in ICU. The study was approved by the ethics committee

137 of the involved institution (Ethics Committee "Catania 1", Catania, Italy; protocol numbers

138 111/2018/PO and 295/2019/EMPO). Study design, protocols and full details on data

139 collection were described elsewhere [20-25, 34].

140

141 For the current analysis, SAPS II was initially used as the main exposure variable. Its

142 computation was based on the following components, as previously described by Le Gall and

143 colleagues [37]: age; heart rate; systolic blood pressure; temperature; Glasgow coma scale;

144 continuous positive airway pressure; PaO2; FiO2; urine output; blood urea nitrogen; sodium;

145 potassium; bicarbonate; bilirubin; white Blood Cell; chronic diseases; type of admission. The

146 SAPS II components were measured 24 hours after admission to the ICU and the worst

147 values within those 24 hours were recorded. Each SAPS II component has a weighted value

148 in points and the total score must be computed adding the weighted values [37]. Additional

149 exposure variables were used to develop a machine learning algorithm for the prediction of

150 HAIs acquired in ICU. Specifically, the machine learning algorithm combined SAPS II with

151 all variables collected at ICU admission but not included in the SAPS II computation. Thus,

152 the original dataset contained only 39% of patients (n=7827) with a complete assessment of

153 variables considered in our study (Figure S I in Supplementary File I**.**).

154

155

156   *Training and Test Set composition and comparison*

157   Since machine learning approaches require large and balanced data set for training, we first

158   built a novel training data set made of recovered and synthetics data to tune the learning

159   algorithms. Specifically, methods for data imputation and balancing of the training set are

160   fully described in the in Supplementary File I. In brief, recovered data were obtained from

161   incomplete records of the original dataset by replacing the missing values using the K-

162   Nearest Neighbor (K-NN) imputation method, as described by Malarvizhi and Thanamani [38].

163   Instead, synthetic data were generated to balance the two classes of infected and non-infected

164   patients using the Synthetic Minority Over-sampling Technique (SMOTE). This technique is

165   a common oversampling method to resample the minority class data following those in the

166   majority class. While the classic oversampling technique duplicates minority data from the

167   minority data population, the SMOTE works by utilizing a K-NN algorithm to create

168   synthetic data. In simple terms, SMOTE first start by choosing random data from the

169   minority class population, then identifies the K-NN, and finally generates synthetic data from

170   the random data and the randomly selected K-NN [39].

171

172   The test set was instead composed by real data of patients with a complete assessment of the

173   following variables at ICU admission: sex (dichotomous), patient's origin (categorical: other

174   ward/healthcare facility, community), non-surgical treatment for acute coronary disease

175   (dichotomous), surgical intervention (dichotomous), SAPS II at admission (continuous),

176   presence of invasive devices at ICU admission (three dichotomous variables for urinary

177   catheter, intubation and central venous catheter, respectively), trauma (dichotomous),

178   impaired immunity (dichotomous), antibiotic therapy in 48 hours before ICU admission

179   (dichotomous).

180    Thus, the training set was made by recovered (n= 7,758) and synthetics records (n=2,544),

181    while the test set included 7,827 real data. The distribution of infected and non-infected

182    patients between the training and test sets is summarized in Table S1 in Supplementary File I**.**

183    To evaluate the goodness of the training set records, we compared the distributions of each

184    single variable with those of the test set to assess that the training data are compliant with the

185    real data. As reported by Figure S II (Supplementary File I), SAPS II and age followed the

186    same distribution in the training and test sets. Likewise, Figures S III and S IV

187    (Supplementary File I) show that the distributions of categorical variables were similar

188    between training and test sets.

189

190    *Learning model generation*

191    To improve the predicting performance of the model, a machine learning algorithm

192    combining the SAPS II with additional variables collected at ICU admission (i.e. sex,

193    patient's origin, non-surgical treatment for acute coronary disease, surgical intervention,

194    presence of intubation, presence of urinary catheter, presence of central vascular catheter;

195    trauma, impaired immunity, antibiotic therapy in 48 hours before ICU admission) was

196    applied. Specifically, we chose the SVM as modeling tool. SVM is a supervised learning

197    algorithm which can be used for classification - especially for binary classification - and

198    regression problems. In the case of two or three exposure variables, the functions used to

199    classify between features are a line or a plane, respectively. In the case of more than three

200    exposure variables, like our model, the function classifying features is referred to a

201    hyperplane. Accordingly, the rationale behind SVM is to find an optimal hyperplane that

202    clearly classifies the different classes (in our case, infected and non-infected patients). The

203    separating hyperplane found by the algorithm provides the largest margin between the two

204    classes. However, our dataset was not linearly separable even in a feature space, not allowing

205    to satisfy all the constraints of SVM [40]. For this reason, we used a non-linear Kernel function

206    (i.e., the Gaussian Kernel, also called as Radial basis function Kernel, RBF). Gaussian RBF is

207    a popular Kernel function used in SVM models to map data that are not originally linearly

208    separable into a higher dimensional feature space where they are made linearly separable. It

209    is worth mentioning that linear kernels are less time consuming than non-linear ones, but they

210    provides less accuracy [40]. To compare the predictive ability of SVM with that obtained

211    through SAPS II alone, we calculated the accuracy (i.e., the proportion of total records that

212    are correctly predicted by the model) and the area under the curve (AUC; ranging from 0.5

213    for no prediction to 1.0 for perfect prediction [13, 15, 41, 42]). In addition, we calculated two

214    evaluation metrics for classification problems, namely precision and recall. Precision - also

215    called positive predictive value - is the fraction of the positive predictions which are truly

216    positive. Recall - often referred to as sensitivity - is the fraction of the initial positives which

217    have been predicted correctly. A perfect classifier should have precision and recall both equal

218    to 1. Data analyses were performed through Python and the SciPy stack. Full details on the

219    computational methods are given in the in Supplementary File I**.**

220

221    *Statistical Analysis*

222    Statistical analyses were performed using SPSS software (version 26.0, SPSS, Chicago, IL).

223    The Kolmogorov- Smirnov test was used to check the normal distribution of continuous

224    variables. Patients' characteristics were described using median and interquartile range (IQR)

225    or percentage. Comparisons between variables were analyzed by the Chi-squared test for

226    categorical variables, while the Mann-Whitney U test was used for continuous variables with

227    skewed distribution. To test the accuracy of the SAPS II in HAI's risk prediction along the

228    range of possible values, we used the Receiver Operating Characteristics (ROC) curve

229    analysis. A ROC curve is a useful graphical tool to evaluate the performance of a binary

230 classifier as its discrimination threshold is varied. In particular, the predictive performance

231 was assessed by calculating the accuracy and the area under the curve (AUC) [13, 15, 41, 42]. All

232 statistical tests were two-sided, and p-values $< 0.05$ were considered statistically significant.

233

234

235 **Results**

236 *Study population*

237 On a total of 20060 SPIN-UTI participants, the current analysis was performed on a

238 subsample of 7827 patients (median age= 69 years; 60.6% males) enrolled from 2006 to

239 2019. The remaining 12233 participants (61%) were excluded because of missing data on the

240 assessment at ICU admission. In this subsample, patients coming from other wards/hospitals

241 and reporting a surgical type of ICU admission were 73.9% and 52.4%, respectively. In

242 general, median SAPS II at admission was 40 (IQR= 28) and length of ICU stay was 5 days

243 (IQR= 10). Patients who reported trauma and impaired immunity were 3.4% and 8.6%,

244 respectively. With respect to medical treatments, 10.2% and 40.9% of patients underwent to

245 non-surgical treatment for acute coronary disease or surgical intervention, while 59% of

246 patients were on antibiotic therapy. In particular, the presence of urinary catheter, intubation

247 and central venous catheter was 77.5%, 59.8% and 41%, respectively. Finally, we observed

248 that percentage of ICU-acquired sepsis among patients enrolled was 6.1%, whereas ICU

249 mortality was 23.2%.

250

251 *Characteristics of infected patients*

252 Overall, **Table I** also shows the comparison between infected ($n = 1,225$; 15.7%) and non-

253 infected patients ($n = 6,602$; 84.3%) for characteristics at ICU admission. Infected patients

254 were more likely to come from the community and to report a medical type of ICU admission

255 than those non-infected. In particular, infected group consisted of patients who were more

256 likely to report impaired immunity, also including more patients with trauma. This translated

257 into higher SAPS II among infected patients if compared with non-infected.

258

259    With respect to the presence of invasive devices, infected patients were also more likely to be

260    intubated at ICU admission and less likely to be catheterized than those non-infected. As

261    expected, infected patients exhibited higher length of ICU stay (20.0 days vs. 4.0 days;

262    p<0.001) compared to non-infected patients. In line with these findings, mortality was also

263    higher in infected patients (35.1%) than in those non-infected (21.0%; p<0.001). No

264    differences were evident for age, sex, non-surgical treatment for acute coronary disease,

265    antibiotic therapy in 48 hours before ICU admission and presence of central venous catheter

266    at ICU admission.

267

268    *ROC Curve Analysis using traditional statistical approach*

269    Using traditional statistical analysis, we aimed to evaluate the performance of SAPS II  at

270    ICU admission in predicting HAIs for all patients staying in ICU for more than two days.

271    Figure 1 shows the ROC curve with an AUC of 0.612 (95% Confidence Interval = 0.60-0.63;

272    *p*<0.001). Although this test was statistically significant, the accuracy of SAPS II for

273    predicting the risk of HAIs was of 56%.

274

275    *ROC Curve Analysis using SVM model*

276    To improve the accuracy for predicting the risk of HAIs, we employed the SVM algorithm,

277    working on SAPS II along with other characteristics at ICU admission. Figure 2 shows the

278    ROC curve of SVM prediction model for the test set. We report that the accuracy of the SVM

279    classifier was 88% on the test set. Specifically, precision and recall were 0.95 and 0.91 for

280    non-infected patients and 0.60 and 0.73 for those who were diagnosed with at least one HAIs

281    during their ICU stay. In line, the predictivity was assessed using ROC curve, which provided

282    an AUC of 0.90 (95% Confidence Interval = 0.88-0.91; *p*<0.001). Our results indicated the

283    reliability of our SVM- model against overfitting. Finally, we aimed to compare our prediction

284    performance with those obtained on the same SVM model, without accounting for the SAPS

285    II variable in the test set. Figure 3 shows the ROC curve of SVM prediction model for the test

286    set, reporting an accuracy of 78%. Accordingly, precision and recall were 0.87 and 0.87 for

287    non-infected patients and 0.31 and 0.32 for those infected, respectively. As expected, the AUC

288    value provided by the ROC curve was 0.66 (95% Confidence Interval = 0.65-0.68; $p<0.001$),

289    indicating a lower predictive ability.

290

291

292 **Discussion**

293 Identifying patients at higher risk of HAIs still represents a major challenge for public health,

294 suggesting the need for novel tools that can guide patient management in ICUs [29-31]. In the

295 past decades, several early warning scores have been developed to evaluate disease severity

296 and to predict the risk of adverse outcomes during ICU stay [43-46]. Among many scores,

297 however, SAPS II still represents one of the most widely used in ICU setting and, therefore,

298 the most represented in the SPIN-UTI dataset. Thus, we first aimed to evaluate the accuracy

299 of SAPS II – calculated at ICU admission – for identifying patients who developed at least

300 one HAI during their ICU stay. In line with previous studies [14, 15, 17], patients who developed

301 at least one HAI exhibited higher SAPS II on ICU admission than those who did not.

302 However, the ROC curve analysis discouraged a predictive application of SAPS II, because

303 both AUC and accuracy were very low albeit statistically significant. Indeed, when AUC

304 obtained from ROC curve analysis is near to 0.5, it means the model has a poor predictive

305 performance.

306

307 Beyond SAPS II, other factors have been associated with the risk of HAIs and related

308 outcomes [4]. For instance, the prolonged use of invasive devices, impaired immunity, surgical

309 intervention and comorbidity were considered as the main risk factors for HAIs in ICU [4, 47].

310 Since infected and non-infected patients included in our study differed in other information

311 available at the time of ICU admission, we hypothesized that combining SAPS II with other

312 patients' characteristics could improve the predictive performance of our model. In this

313 scenario, machine learning approaches represent a possible strategy for healthcare facilities,

314 making possible to build a specific prediction model targeted to demographics and clinical

315 characteristics of patients [26, 27]. Indeed, there is current consensus that machine learning

316 algorithms could support and enrich conventional statistical approaches, especially in terms

317 of prediction of ICU prognosis, clinical deterioration and risk assessment [26, 27, 35]. Machine

318 learning systems have been developed in many fields of medicine including infectious

319 diseases control and clinical decision support [48]. Particularly, machine learning technique has

320 been applied in patients with sepsis [49], to predict candidemia [50] or complications related to

321 *Clostridium difficile* infection [51], to improve the prediction of antimicrobial resistance [52], and

322 for surveillance purpose [53].

323

324 To the best of our knowledge, the present study is the first employing machine learning

325 methods to identify patients at higher risk of HAIs, according to their individual

326 characteristics at ICU admission. To do that, we trained and tested a machine learning model,

327 which combined SAPS II with additional patients' characteristics collected at ICU admission.

328 Our intention was to use variables that are easily and routinely collected at ICU admission,

329 such as patients' demographic, origin and type of admission, medical history, and disease

330 severity. Notably, our findings demonstrated a high performance of our model, as indicated

331 by an AUC of 0.90. In line, the accuracy on the test set was 88%, with precision and recall

332 values of 95% and 91% for non-infected patients, and 60% and 73% for those who developed

333 at least one HAI. It is worth mentioning that an excellent model has AUC near to the 1, which

334 means it has a good predictive performance. Although SAPS II was the predictor that

335 weighted more on the model – as demonstrated by the sensitivity analysis – the inclusion of

336 additional characteristics significantly improved the prediction of patients who developed

337 HAIs in ICU (i.e. AUC raised from 0.6 for SAPS II alone to 0.9 for the machine learning

338 model).

339

340 Our findings confirm that applying machine learning algorithms for classification and

341 predictive problems might help solve many public health issues, including those plaguing

342  critical ill patients. These machine learning algorithms, if properly applied, could overcome

343  limitations of existing traditional tools such as early warning scores [43-46, 54-60]. However, it

344  will be our duty to compare the performance of our model with other early warning scores,

345  including the Acute Physiologic Assessment and Chronic Health Evaluation II (APACHE II)

346  and the Sequential Organ Failure Assessment (SOFA) [13, 16].

347

348  The main strengths of our work also include some methodological aspects worthy of note.

349  Indeed, our model was trained and tested on large datasets obtained through patient-based

350  prospective surveillance across Italian regions. Moreover, the surveillance is based on

351  structured and standardized tools provided by the ECDC protocols. This allows not only to

352  establish an Italian benchmark for planning preventive strategies in the future, but also to

353  compare and to validate our findings with those that will be obtained in other European

354  countries. On the other hand, however, there are some points to keep in mind when

355  interpreting our results. The first one is that machine learning algorithms should not be seen

356  as substitutes of existing scores, but rather they could support clinicians in the decision-

357  making process. For instance, on the basis of our findings, it could be hypothesized to

358  develop an automated tool able to identify patients who need more attention because of their

359  high risk of HAI. However, our approach must not be seen as a fixed model, but it could be

360  integrated and/or modified according to specific needs. Similarly, in the near future, it is our

361  intention to apply a similar approach to stratify for specific type of HAIs and to predict other

362  adverse outcomes (e.g. mortality) for critical ill patients, also considering length of stay in

363  ICU. The second point to be considered is that approximately 60% of SPIN-UTI records were

364  incomplete. Although this did not exclude potential bias that cannot be controlled in the

365  current analysis, we used incomplete records to generate recovered data. If on the one hand it

366  would be preferable to use real data, on the other hand our novel approach gives an

367   alternative when data are scarce or incomplete. The third point is that machine learning is

368   frequently referred to as a "black box" for clinicians, which however expect to become

369   familiar with it and to be able to pinpoint why a decision is suggested [61, 62]. For this reason,

370   data scientists are trying to develop more interpretable algorithms in medical fields, even if

371   we are only just beginning to build trust in these new technologies [61, 62].

372

373   **Conclusions**

374   Our findings provide a promising evaluation of a better predictive performance of the SVM

375   algorithm than conventional statistical approaches, suggesting the SVM as a possible tool to

376   identify and predict patients at higher risk of HAIs at ICU admission, providing clinicians

377   sufficient time to potentially prevent HAI and mitigate its severity, targeting specific

378   infection prevention and control interventions to high-risk groups in order to improve quality

379   of care.

380

**Figure legends**

**Figure 1.** ROC curve of the SAPS II predicting healthcare associated infections. The figure shows the ability of SAPS II to identify patients who developed at least one HAI during their ICU stay. The curve plots true positive rate (i.e., sensitivity) versus false positive rate (i.e., 1 - specificity) at different classification thresholds. The blue curve represents the ability of SAPS II to discriminate patients who developed at least one HAI from those who did not (Area Under the Curve, AUC = 0.612; 95% Confidence Interval = 0.60-0.63; $p$<0.001). The black dotted line is the reference for no predictive ability (AUC=0.500).

**Figure 2.** ROC curve of support vector machine algorithm predicting healthcare associated infections. The model is based on a Support Vector Machines (SVM) algorithm, which combines SAPS II with additional features at ICU admission. The curve plots true positive rate (i.e., sensitivity) versus false positive rate (i.e., 1 - specificity). The blue curve represents the ability of the SVM algorithm to predict patients who developed at least one HAI from those who did not (Area Under the Curve, AUC= 0.90; 95% Confidence Interval = 0.88-0.91; $p$<0.001). The black dotted line is the reference for no predictive ability (AUC=0.500).

**Figure 3.** ROC curve of support vector machine algorithm predicting healthcare associated infections, by excluding SAPS II. The model is based on a support vector machine algorithm, which combines patients' characteristics collected at ICU admission. The curve plots true positive rate (i.e., sensitivity) versus false positive rate (i.e., 1 - specificity). The blue curve represents the ability of the SVM algorithm to predict patients who developed at least one HAI from those who did not (Area Under the Curve, AUC= 0.66; 95% Confidence Interval = 0.65-0.68; $p$<0.001). The black dotted line is the reference for no predictive ability (AUC=0.500).

**Competing interests:**

429  Ospedaliera Universitaria Policlinico Vittorio Emanuele, Catania); Anna Maria Longhitano,

430  Giorgio Scrofani, Maria Concetta Monea (Azienda Ospedaliera Cannizzaro, Catania); Marina

431  Milazzo (Presidio Ospedaliero di Vittoria, Ragusa; Presidio Ospedaliero Civile-OMPA,

432  Ragusa; ASP 7, Ragusa); Antonino Giarratano, Giuseppe Calamusa, Maria Valeria

433  Torregrossa (Azienda Ospedaliera Universitaria Policlinico Paolo Giaccone, Palermo);

434  Antonino Di Benedetto (ASP di Palermo); Giuseppa Maria Gisella Rizzo (ASP di Trapani);

435  Giuseppe Manta (ASP di Caltanissetta; Presidio Ospedaliero Sant'Elia, Caltanissetta);

436  Romano Tetamo, Rosa Mancuso (ARNAS "Civico-Ascoli-Di Cristina", Palermo); Laura

437  Maria Mella (Azienda Ospedaliera "Luigi Sacco", Milano); Ignazio Dei (ASL Lanusei,

438  Presidio Ospedaliero "Nostra Signora della Mercede", Nuoro); Irene Pandiani (Presidio

439  Ospedaliero SS. Annunziata, Taranto); Antonino Cannistrà, Paola Piotti (Clinica S. Rocco di

440  Franciacorta S.p.A. Ome, Brescia); Massimo Girardis, Elena Righi, Alberto Barbieri

441  (Policlinico di Modena); Patricia Crollari (Presidio Ospedaliero di Vizzolo Predabissi,

442  Milano); Albino Borracino, Salvatore Coniglio, Rosaria Palermo, Sergio Pintaudi, Daniela Di

443  Stefano (ARNAS Garibaldi, Catania); Antonina Romeo (Azienda Ospedaliera Villa Sofia-

444  Cervello, Presidio Ospedaliero Villa Sofia, Palermo); Giovanna Sticca (Fondazione di Ricerca

445  e Cura "Giovanni Paolo II" UCSC, Campobasso); Massimo Minerva (Azienda Ospedaliera di

446  Melegnano - Presidio Ospedaliero di Cernusco sul Naviglio, Milano); Leila Fabiani,

447  Alessandra Gentile (Presidio Ospedaliero S. Salvatore, L'Aquila); Paolo Stefanini (AUSL di

448  Reggio Emilia; Ospedale Civile Di Guastalla, Reggio Emilia); Marcello Mario D'Errico,

449  Abele Donati (Azienda Ospedaliero-Universitaria Ospedali Riuniti, Ancona); Santa De

450  Remigis, Federica Venturoni (Presidio Ospedaliero di Teramo); Manuela Antoci (Presidio

451  Ospedaliero Civile-OMPA, Ragusa; ASP 7, Ragusa); Riccardo Pagliarulo (Azienda Sanitaria

452  Locale Ospedale "Di Venere", Bari); Aida Bianco, Maria Pavia (Azienda Ospedaliera

453  Universitaria "Mater Domini", Catanzaro); Marcello Pasculli, Cesare Vittori (Azienda

454 Ospedaliera Universitaria Senese, Siena); Giovanni Battista Orsi (Azienda Ospedaliera

455 Sant'Andrea, Roma); Cristina Arrigoni (IRCCS Policlinico San Matteo di Pavia); Maria

456 Patrizia Olori (Stabilimento Ospedaliero "C. e G. Mazzoni", Ascoli Piceno); Massimo

457 Antonelli, Patrizia Laurenti (Fondazione Policlinico Gemelli, Roma); Franco Ingala, Carmela

458 Conte, Salvatore Russo, Laura Condorelli (ASP 8, Siracusa); Patrizia Farruggia (Presidio

459 Ospedaliero Unico Aziendale dell'Azienda USL di Bologna); Cristina Maria Luisa (Ente

460 Ospedaliero Ospedali Galliera di Genova); Italia Galassi (Presidio Ospedaliero "SS. Filippo e

461 Nicola" di Avezzano).

462

463

464

465 **Supplementary File I.**

466 Supplementary methods and files related to the development of the Support Vector Machine

467 Algorithm.

**References**

1.  Alp E, Damani N Healthcare-associated infections in intensive care units: epidemiology and infection control in low-to-middle income countries. *J Infect Dev Ctries* 2015; **9**: 1040-5.

2.  Haque M, Sartelli M, McKimm J, Abu Bakar M Health care-associated infections - an overview. *Infect Drug Resist* 2018; **11**: 2321-33.

3.  Serra-Burriel M, Keys M, Campillo-Artero C *et al.* Impact of multi-drug resistant bacteria on economic and clinical outcomes of healthcare-associated infections in adults: Systematic review and meta-analysis. *PLoS One* 2020; **15**: e0227139.

4.  Marcel JP, Alfa M, Baquero F *et al.* Healthcare-associated infections: think globally, act locally. *Clin Microbiol Infect* 2008; **14**: 895-907.

5.  Suetens C, Latour K, Kärki T *et al.* Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: results from two European point prevalence surveys, 2016 to 2017. *Euro Surveill* 2018; **23**.

6.  Agodi A, Barchitta M, Quattrocchi A *et al.* Antibiotic trends of Klebsiella pneumoniae and Acinetobacter baumannii resistance indicators in an intensive care unit of Southern Italy, 2008-2013. *Antimicrob Resist Infect Control* 2015; **4**: 43.

7.  Barchitta M, Quattrocchi A, Maugeri A *et al.* The "Obiettivo Antibiotico" Campaign on Prudent Use of Antibiotics in Sicily, Italy: The Pilot Phase. *Int J Environ Res Public Health* 2020; **17**.

8.  Barchitta M, Maugeri A, La Rosa MC *et al.* Carbapenem Consumption and Rate of carbapenem-resistant gram-negative bacteria: results from the Sicilian Surveillance System. *Ann Ig* 2020.

9.  Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Murolo G, Agodi A Three-Year Trends of Healthcare-Associated Infections and Antibiotic Use in Acute Care Hospitals: Findings from 2016-2018 Point Prevalence Surveys in Sicily, Italy. *Antibiotics (Basel)* 2020; **10**.

10. European Center for Disease Prevention and Control. Healthcare-associated infections acquired in intensive care units. Annual Epidemiological Report for 2017. https://www.ecdc.europa.eu/sites/default/files/documents/AER_for_2017-HAI.pdf: 2017.

11. European Center for Disease Prevention and Control. Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals 2011-2012. https://www.ecdc.europa.eu/en/publications-data/point-prevalence-survey-healthcare-associated-infections-and-antimicrobial-use-0

12. Barchitta M, Maugeri A, Favara G *et al.* Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care unit: findings from the SPIN-UTI network. *J Hosp Infect* 2020.

13. Gilani MT, Razavi M, Azad AM A comparison of Simplified Acute Physiology Score II, Acute Physiology and Chronic Health Evaluation II and Acute Physiology and Chronic Health Evaluation III scoring system in predicting mortality and length of stay at surgical intensive care unit. *Niger Med J* 2014; **55**: 144-7.

14. Sadaka F, EthmaneAbouElMaali C, Cytron MA, Fowler K, Javaux VM, O'Brien J Predicting Mortality of Patients With Sepsis: A Comparison of APACHE II and APACHE III Scoring Systems. *J Clin Med Res* 2017; **9**: 907-10.

15. Mungan Ib, Bektaş Se, Çavuş MA, Sarı S, Turan S. The predictive power of SAPS-3 and SOFA scores and their relations with patients outcomes in the Surgical Intensive Care Unit. 2019.

16. Haddadi A, Ledmani M, Gainier M, Hubert H, Tagne J, De Micheaux P. Comparing the APACHE II, SOFA, LOD, and SAPS II scores in patients who have developed a nosocomial infection. Bangladesh Crit Care J March 2014; 2 (1): 4-9

17. Agodi A, Barchitta M, Auxilia F *et al.* Epidemiology of intensive care unit-acquired sepsis in Italy: results of the SPIN-UTI network. *Ann Ig* 2018; **30**: 15-21.

18. Allyn J, Ferdynus C, Bohrer M, Dalban C, Valance D, Allou N Simplified Acute Physiology Score II as Predictor of Mortality in Intensive Care Units: A Decision Curve Analysis. *PLoS One* 2016; **11**: e0164828.

19. Nielsen A, Thorsen-Meyer H, Belling K *et al.* Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digital Health 2019; 1: e78–89.

20. Agodi A, Barchitta M, Quattrocchi A *et al.* Preventable proportion of intubation-associated pneumonia: Role of adherence to a care bundle. *PLoS One* 2017; **12**: e0181170.

21. Agodi A, Auxilia F, Barchitta M *et al.* Antibiotic consumption and resistance: results of the SPIN-UTI project of the GISIO-SItI. *Epidemiol Prev* 2015; **39**: 94-8.

22. Agodi A, Auxilia F, Barchitta M *et al.* Trends, risk factors and outcomes of healthcare-associated infections within the Italian network SPIN-UTI. *J Hosp Infect* 2013; **84**: 52-8.

23. Agodi A, Auxilia F, Barchitta M *et al.* Building a benchmark through active surveillance of intensive care unit-acquired infections: the Italian network SPIN-UTI. *J Hosp Infect* 2010; **74**: 258-65.

535    24.    Agodi A, Auxilia F, Barchitta M *et al.* [Control of intubator associated pneumonia in
536          intensive care unit: results of the GISIO-SItI SPIN-UTI Project]. *Epidemiol Prev* 2014; **38**:
537          51-6.

538    25.    Agodi A, Barchitta M, Mura I, Pasquarella C, Torregrossa MV, SItI G The commitment of
539          the GISIO-SItI to contrast Healthcare-Associated Infections and the experience of
540          prevalence studies in Sicily. *Ann Ig* 2018; **30**: 38-47.

541    26.    Komorowski M Artificial intelligence in intensive care: are we there yet? *Intensive Care*
542          *Med* 2019; **45**: 1298-300.

543    27.    Rajkomar A, Dean J, Kohane I Machine Learning in Medicine. Reply. *N Engl J Med* 2019;
544          **380**: 2589-90.

545    28.    Favara G, Riela P, Maugeri A, Barchitta M, Gallo G, Agodi A. Risk of Pneumonia and
546          associated outcomes in Intensive Care Unit: an integrated approach of Visual and Cluster
547          Analysis. IEEE World Congress on Services (SERVICES) 2019.

548    29.    Yee CR, Narain NR, Akmaev VR, Vemulapalli V A Data-Driven Approach to Predicting
549          Septic Shock in the Intensive Care Unit. *Biomed Inform Insights* 2019; **11**:
550          1178222619885147.

551    30.    Chen L, Dubrawski A, Wang D *et al.* Using Supervised Machine Learning to Classify Real
552          Alerts and Artifact in Online Multisignal Vital Sign Monitoring Data. *Crit Care Med* 2016;
553          **44**: e456-63.

554    31.    Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP Multicenter
555          Comparison of Machine Learning Methods and Conventional Regression for Predicting
556          Clinical Deterioration on the Wards. *Crit Care Med* 2016; **44**: 368-74.

557    32.    Wiens J, Shenoy ES Machine Learning for Healthcare: On the Verge of a Major Shift in
558          Healthcare Epidemiology. *Clin Infect Dis* 2018; **66**: 149-53.

559    33.    Michard F, Teboul JL Predictive analytics: beyond the buzz. *Ann Intensive Care* 2019; **9**:
560          46.

561    34.    Masia MD, Barchitta M, Liperi G *et al.* Validation of intensive care unit-acquired infection
562          surveillance in the Italian SPIN-UTI network. *J Hosp Infect* 2010; **76**: 139-42.

563    35.    Linnen DT, Escobar GJ, Hu X, Scruth E, Liu V, Stephens C Statistical Modeling and
564          Aggregate-Weighted Scoring Systems in Prediction of Mortality and ICU Transfer: A
565          Systematic Review. *J Hosp Med* 2019; **14**: 161-9.

566    36.    European Center for Disease Prevention and Control. European surveillance of healthcare-
567          associated infections in intensive care units- HAI-Net ICU protocol- Protocol version 1.02.

568  https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/healthcare

569  -associated-infections-HAI-ICU-protocol.pdf: 2015.

570  37.  Le Gall JR, Lemeshow S, Saulnier F A new Simplified Acute Physiology Score (SAPS II)

571  based on a European/North American multicenter study. *JAMA* 1993; **270**: 2957-63.

572  38.  Malarvizhi R, Thanamani A. K-nearest neighbor in missing data imputation.  Int J Eng Res

573  Dev 5(1):05–07 2012.

574  39.  Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling

575  Technique.  JAIR 16 (2002), 321--357.

576  40.  Cortes C, Vapnik V. Support-Vector Networks.  Mach. Learn. 20, 3 (September 1995), 273–

577  297.

578  41.  Martos-Benítez FD, Larrondo-Muguercia H, León-Pérez D, Rivero-López JC, Orama-

579  Requejo V, Martínez-Alfonso JL Performance of three prognostic models in critically ill

580  patients with cancer: a prospective study. *Int J Clin Oncol* 2020.

581  42.  D'Arrigo G, Provenzano F, Torino C, Zoccali C, Tripepi G. I test diagnostici e l'analisi della

582  curva ROC.  G I tal  N efrol  2011; 28 (6): 642-647.

583  43.  Gerry S, Bonnici T, Birks J *et al.* Early warning scores for detecting deterioration in adult

584  hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020; **369**:

585  m1501.

586  44.  Brennan TA, Leape LL, Laird NM *et al.* Incidence of adverse events and negligence in

587  hospitalized patients: results of the Harvard Medical Practice Study I. 1991. *Qual Saf Health*

588  *Care* 2004; **13**: 145-51; discussion 51-2.

589  45.  Vincent C, Neale G, Woloshynowych M Adverse events in British hospitals: preliminary

590  retrospective record review. *BMJ* 2001; **322**: 517-9.

591  46.  Institute of Medicine (US) Committee on Quality of Health Care in America. To Err is

592  Human: Building a Safer Health System.  2000.

593  47.  Tan X, Rolls K, Wiseman T, Betihavas V. Risk factors for Healthcare Associated Infections

594  (HAI) or sepsis in trauma patients : an integrative literature review.  16Th International

595  Conference For Emergency Nursing (Icen18).

596  48.  Peiffer-Smadja N, Rawson TM, Ahmad R *et al.* Corrigendum to 'machine learning for

597  clinical decision support in infectious diseases: a narrative review of current applications'

598  clinical microbiology and infection (2020) 584-595. *Clin Microbiol Infect* 2020; **26**: 1118.

599  49.  Vellido A, Ribas V, Morales C, Ruiz Sanmartín A, Ruiz Rodríguez JC Machine learning in

600  critical care: state-of-the-art and a sepsis case study. *Biomed Eng Online* 2018; **17**: 135.

601   50.   Ripoli A, Sozio E, Sbrana F *et al.* Personalized machine learning approach to predict
602         candidemia in medical wards. *Infection* 2020.

603   51.   Li BY, Oh J, Young VB, Rao K, Wiens J Using Machine Learning and the Electronic
604         Health Record to Predict Complicated. *Open Forum Infect Dis* 2019; **6**: ofz186.

605   52.   Macesic N, Polubriaginof F, Tatonetti NP Machine learning: novel bioinformatics
606         approaches for combating antimicrobial resistance. *Curr Opin Infect Dis* 2017; **30**: 511-7.

607   53.   Roth JA, Battegay M, Juchler F, Vogt JE, Widmer AF Introduction to Machine Learning in
608         Digital Healthcare Epidemiology. *Infect Control Hosp Epidemiol* 2018; **39**: 1457-62.

609   54.   Hillman KM, Bristow PJ, Chey T *et al.* Duration of life-threatening antecedents prior to
610         intensive care admission. *Intensive Care Med* 2002; **28**: 1629-34.

611   55.   Scardoni A, Balzarini F, Signorelli C, Cabitza F, Odone A Artificial intelligence-based tools
612         to control healthcare associated infections: A systematic review of the literature. *J Infect*
613         *Public Health* 2020; **13**: 1061-77.

614   56.   Lovejoy CA, Buch V, Maruthappu M Artificial intelligence in the intensive care unit. *Crit*
615         *Care* 2019; **23**: 7.

616   57.   Desautels T, Calvert J, Hoffman J *et al.* Prediction of Sepsis in the Intensive Care Unit With
617         Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform*
618         2016; **4**: e28.

619   58.   Parreco JP, Hidalgo AE, Badilla AD, Ilyas O, Rattan R Predicting central line-associated
620         bloodstream infections and mortality using supervised machine learning. *J Crit Care* 2018;
621         **45**: 156-62.

622   59.   Deo RC Machine Learning in Medicine. *Circulation* 2015; **132**: 1920-30.

623   60.   Deo RC Machine Learning in Medicine: Will This Time Be Different? *Circulation* 2020;
624         **142**: 1521-3.

625   61.   The Lancet Respiratory Medicine Opening the black box of machine learning. *Lancet Respir*
626         *Med* 2018; **6**: 801.

627   62.   Luz CF, Vollmer M, Decruyenaere J, Nijsten MW, Glasner C, Sinha B Machine learning in
628         infection management using routine electronic health records: tools, techniques, and
629         reporting of future technologies. *Clin Microbiol Infect* 2020; **26**: 1291-9.

630

**Table I.** Characteristics of patients according to their infectious status

| Characteristics | Patients (*n*=7827) | Infected patients (*n*=1225) | Non- infected patients (n=6602) | *p*-value |
|---|---|---|---|---|
| **Age, years** | 69.0 (21.0) | 69.0 (21.0) | 69.0 (21.0) | 0.064 |
| **Sex (% men)** | 60.6% | 62.8% | 60.1% | 0.084 |
| **Patient's origin** | | | | |
| Other ward/healthcare facility | 73.9% | 67.7% | 75.1% | **<0.001** |
| Community | 26.1% | 32.3% | 24.9% | |
| **SAPS II score at admission** | 40.0 (28.0) | 47.0 (27.0) | 38.0 (27.0) | **<0.001** |
| **Type of ICU admission** | | | | |
| Medical | 47.6% | 53.6% | 46.5% | **<0.001** |
| Surgical | 52.4% | 46.4% | 53.5% | |
| **Trauma** | 3.4% | 5.0% | 3.2% | **0.001** |
| **Impaired immunity** | 8.6% | 10.4% | 8.2% | **0.015** |
| **Non-surgical treatment for acute coronary disease** | 10.2% | 8.9% | 10.4% | 0.109 |
| **Surgical intervention** | 40.9% | 36.7% | 41.7% | **<0.001** |
| **Antibiotic therapy in 48 hours before ICU admission** | 59% | 59.8% | 58.9% | 0.579 |
| **Presence of urinary catheter at ICU admission** | 77.5% | 74.4% | 78.0% | **0.006** |
| **Presence of intubation at ICU admission** | 59.8% | 63.8% | 59.1% | **0.002** |
| **Presence of central venous catheter at ICU admission** | 41% | 39.7% | 41.3% | 0.295 |
| **ICU-acquired sepsis (%yes)** | 6.1% | 37.6% | - | - |
| **Outcome (%death)** | 23.2% | 35.1% | 21.0% | **<0.001** |
| **Length of ICU stay, days** | 5.0 (10.0) | 20.0 (20.0) | 4.0 (6.0) | **<0.001** |

*Results are reported as median (interquartile range) for continuous variables, or percentage for categorical variables. Statistical analyses were performed using the Mann-Whitney or the Chi-squared test.

ROC curve