



Early diagnosis of candidemia with explainable machine learning on automatically extracted laboratory and microbiological data: results of the AUTO-CAND project

Daniele Roberto Giacobbe, Cristina Marelli, Sara Mora, Sabrina Guastavino, Chiara Russo, Giorgia Brucci, Alessandro Limongelli, Antonio Vena, Malgorzata Mikulska, Maryam Tayefi, Stefano Peluso, Alessio Signori, Antonio Di Biagio, Anna Marchese, Cristina Campi, Mauro Giacomini & Matteo Bassetti

To cite this article: Daniele Roberto Giacobbe, Cristina Marelli, Sara Mora, Sabrina Guastavino, Chiara Russo, Giorgia Brucci, Alessandro Limongelli, Antonio Vena, Malgorzata Mikulska, Maryam Tayefi, Stefano Peluso, Alessio Signori, Antonio Di Biagio, Anna Marchese, Cristina Campi, Mauro Giacomini & Matteo Bassetti (2023) Early diagnosis of candidemia with explainable machine learning on automatically extracted laboratory and microbiological data: results of the AUTO-CAND project, *Annals of Medicine*, 55:2, 2285454, DOI: [10.1080/07853890.2023.2285454](https://doi.org/10.1080/07853890.2023.2285454)

To link to this article: <https://doi.org/10.1080/07853890.2023.2285454>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 27 Nov 2023.



[Submit your article to this journal](#)



Article views: 1034



[View related articles](#)






[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

Early diagnosis of candidemia with explainable machine learning on automatically extracted laboratory and microbiological data: results of the AUTO-CAND project

Daniele Roberto Giacobbe^{a,b#} , Cristina Marelli^{b#} , Sara Mora^{c#}, Sabrina Guastavino^d, Chiara Russo^{a,b}, Giorgia Brucci^{a,b}, Alessandro Limongelli^{a,b}, Antonio Vena^{a,b}, Malgorzata Mikulska^{a,b}, Maryam Tayefi^e, Stefano Peluso^f, Alessio Signori^g, Antonio Di Biagio^{a,b}, Anna Marchese^{h,i}, Cristina Campi^{d,j}, Mauro Giacomini^c  and Matteo Bassetti^{a,b}

^aDepartment of Health Sciences (DISSAL), University of Genoa, Genoa, Italy; ^bClinica Malattie Infettive, IRCCS Ospedale Policlinico San Martino, Genoa, Italy; ^cDepartment of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy; ^dDepartment of Mathematics (DIMa), University of Genoa, Genoa, Italy; ^eNorwegian Centre for E-Health Research, Tromsø, Norway; ^fDepartment of Statistics and Quantitative Methods, University of Milan - Bicocca, Milan, Italy; ^gSection of Biostatistics, Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy; ^hDepartment of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genoa, Italy; ⁱMicrobiology Unit, IRCCS Ospedale Policlinico San Martino, Genoa, Italy; ^jLife Science Computational Laboratory (LISCOMP), IRCCS Ospedale Policlinico San Martino, Genoa, Italy

ABSTRACT

Background: Candidemia is associated with a heavy burden of morbidity and mortality in hospitalized patients. The availability of blood culture results could require up to 48–72 h after blood draw; thus, early treatment decisions are made in the absence of a definite diagnosis.

Methods: In this retrospective study, we assessed the performance of different supervised machine learning algorithms for the early differential diagnosis of candidemia and bacteremia in adult patients on a large dataset automatically extracted within the AUTO-CAND project.

Results: Overall, 12,483 episodes of candidemia (1275; 10%) or bacteremia (11,208; 90%) were included in the analysis. A random forest classifier achieved the best diagnostic performance for candidemia, with sensitivity 0.98 and specificity 0.65 on the training set (true skill statistic [TSS]=0.63) and sensitivity 0.74 and specificity 0.57 on the test set (TSS = 0.31). Then, the random classifier was trained in the subgroup of patients with available serum β -D-glucan (BDG) and procalcitonin (PCT) values by exploiting the feature ranking learned in the entire dataset. Although no statistically significant differences were observed from the performance measures obtained by employing BDG and PCT alone, the performance measures of the classifier that included the features selected in the entire dataset, plus BDG and PCT, were the highest in most cases.

Conclusions: Random forest classifiers trained on large datasets of automatically extracted data have the potential to improve current diagnostic algorithms for candidemia. However, further development through implementation of automatically extracted clinical features may be necessary to achieve crucial improvements.

ARTICLE HISTORY

Received 24 August 2023
Revised 10 November 2023
Accepted 13 November 2023

KEYWORDS

Candidemia; machine learning; supervised; biomarker; glucan; procalcitonin; random forest


Background

Candidemia (i.e. bloodstream infection due to yeasts of the genus *Candida*) carries a heavy burden of morbidity and mortality in hospitalized patients, with mortality rates surpassing 50% in critically ill patients [1–4]. Furthermore, the development of candidemia is

associated with prolonged hospital stays and increased healthcare costs [1].

There are no specific symptoms or signs associated with candidemia that can clearly differentiate it from bacteremia (i.e. bloodstream infection due to bacteria), which is more frequently encountered in hospitalized subjects [5–7]. Availability of blood culture results

CONTACT Daniele Roberto Giacobbe  danieleroberto.giacobbe@unige.it , Department of Health Sciences (DISSAL), Via A. Pastore 1, Genoa 16132, Italy.
[#]Equal contributors.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07853890.2023.2285454>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

could require up to 48–72 h after blood draw; thus, early treatment decisions (i.e. whether to start antifungals besides empirical antibacterials) are made in the absence of a definite diagnosis. Appropriate early treatment choices should: (i) guarantee early initiation of antifungals in patients who truly have candidemia (or mixed candidemia and bacteremia), given the reported increase in mortality associated with delays in therapy [8, 9]; and (ii) avoid as much as possible useless antifungal treatment in patients with bacteremia only, in line with antifungal stewardship principles [10, 11].

Considering the similar presentation of bacteremia and candidemia, physicians often base their decision on whether to administer early antifungals on risk scores for candidemia and/or rapid serum fungal antigens, such as serum β -D-glucan (BDG), or molecular tests [12–28]. Although useful, these approaches are not perfect; thus, further advances in the accuracy of candidemia prediction could help improve the appropriate early treatment of patients with a consistent clinical picture. Recently, there has been a growing interest in exploiting the aid of machine learning (ML) algorithms for this purpose, although frequently with the lack of large datasets (in terms of both features and training examples) [29–33].

In the present study, we aimed to assess the performance of different supervised ML algorithms for the early differential diagnosis of candidemia and bacteremia in adult patients on a previously automatically extracted large dataset within the AUTO-CAND project.

Methods

Setting and objective

This retrospective study, conducted at IRCCS Ospedale Policlinico San Martino, a 1200-bed teaching hospital in Italy, comprised the second phase of the AUTO-CAND project. In the first phase, we validated the automated extraction of laboratory and microbiological data pertaining to candidemia and bacteremia episodes that occurred in our center from 1 January 2011 to 31 December 2019 [34, 35]. The details of the automated extraction process have been reported elsewhere [35]. Briefly, the automated system extracted 15,752 episodes of bloodstream infection that occurred during the study period [35]. More in detail, the following were extracted from the hospital laboratory information system (LIS): 14,112 episodes of bacteremia (90%), 1338 episodes of candidemia (8%), and 302 episodes of mixed candidemia/bacteremia (2%) [35]. The primary objective of the second phase of the AUTO-CAND

project, reported in the present study, was to assess the diagnostic performance of different supervised, explainable ML algorithms for the early differential diagnosis of candidemia and bacteremia in adult patients on the previously extracted dataset. The exclusion criterion was an age < 18 years. The AUTO-CAND project was approved by the pertinent local ethics committee (Liguria Region Ethics Committee, registry number 71/2020). The requirement for informed consent specific for the present study was waived owing to the retrospective nature of the analyses.

Definitions

Candidemia was defined as the presence of at least one positive blood culture for *Candida* spp.; bacteremia was defined as the presence of at least one positive blood culture for one or more bacteria. At the time of collection, blood samples were labeled in the ward with a unique number provided by the hospital software and sent to the internal hospital laboratory for immediate incubation and registration in the LIS. The origin of each episode was defined as the day, hour, and minute when the first blood culture (later positive for *Candida* spp. and/or bacteria) was registered in the LIS. A blood culture positive for the same *Candida* spp. or bacteria causing a previous episode was considered the origin of a novel episode only if collected at least 30 days after the last positive blood culture collected within the previous episode. Episodes of candidemia and bacteremia starting within 48 h in the same patient were defined as mixed candidemia and bacteremia episodes [36]. Bacteremia episodes caused by coagulase-negative staphylococci or other common skin colonizers were considered only in the presence of two positive blood cultures (for the same organism) collected less than 48 h apart from two different sites or sets [37]. All tasks were automatically performed using an extraction system [35].

Data collected for the study

Besides age, sex, and ward of stay at the time of candidemia or bacteremia, the developed automated system was able to extract the laboratory test results as they were, if performed, at each of the following time points: (i) at the episode origin and (ii) in each of the seven days before the origin [35]. The automated system extracted the results of the following laboratory tests: white cell count, red cell count, platelet count, neutrophil cell count, lymphocyte cell count, basophil

cell count, eosinophil cell count, monocyte cell count, hemoglobin, hematocrit, creatinine, urea, uric acid, lactate, lactate dehydrogenase (LDH), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin, direct bilirubin, activated partial thromboplastin time (aPTT), prothrombin time, international normalized ratio (INR), glucose, glycated hemoglobin, total proteins, albumin, triglycerides, C-reactive protein (CRP), procalcitonin (PCT), and BDG. The automated system also extracted the following microbiological information pertaining to the 30 days before the origin of the episode (based on the results of cultures from respiratory, urinary, and gastrointestinal sites [25]): (i) respiratory colonization by *Candida* spp., (ii) urinary colonization by *Candida* spp., (iii) gastrointestinal colonization by *Candida* spp., (iv) *Candida* colonization (yes vs. no), and (v) multifocal *Candida* colonization. The system was able to automatically detect if a specific laboratory test was performed at a given time point and if a specific colonization site was explored (i.e. whether cultures were collected within 30 days before the origin of the episode) to correctly define missing values for statistical analysis [35].

Statistical analysis

The devised pipeline, described in detail in the following paragraphs, is also summarized graphically in Figure 1. All the following analyses were performed in a Python 3 (Anaconda, Jupyter Notebook 6.5.2) environment. The source code of AUTO-CAND is available at the GitHub repository [38].

Dataset preparation

As expected, based on routine clinical practice, a large percentage of missing data (especially sequential data for a given laboratory test in subsequent days) was present either by episode or by features (not all tests were performed daily but only selected ones according to clinical needs). Therefore, for each laboratory test, we decided to consider only one result, i.e. the closest to the origin within the two previous days (i.e. the closest to the onset of candidemia or bacteremia). While this precluded the assessment of the impact of sequential variations over days of laboratory values on the prediction of candidemia, it allowed us to conspicuously reduce the number of missing values for a given variable. Then, we excluded those episodes and features for which the percentage of missing values was still higher than 50% in order to preserve the reliability of multiple imputation [39]. Overall, 12,483

episodes and 31 of the 42 available features were retained (Figure 2 and Supplementary figure S1). Of note, among the excluded features were serum BDG and serum PCT, due to a percentage of missing values higher than 89% and 72%, respectively. Based on previous literature, these two tests are expected to show the best diagnostic performance for candidemia (as single markers) among the explored laboratory tests [27]. For this reason, after training on the entire dataset (without BDG and PCT), the ML model showing the best performance was also evaluated in the subgroup of patients with BDG and PCT values not missing to verify whether the classifier was able to improve the isolated performance of BDG and PCT (see below). Eventually, 29/31 variables were included in the analyses (ward of stay was excluded to focus only on laboratory results and demographic data in line with the aim of the study, and the number of explored sites was excluded as unnecessary owing to imputation of the variable *Candida* colonization, see below).

Dataset splitting

Before applying ML algorithms, we split the complete sample into two datasets: training set and test set. We randomly assigned 70% of the episodes to the training set and the remaining 30% to the test set, maintaining the same proportion of candidemia episodes (including mixed episodes; hereafter, we refer to this group as candidemia episodes) and bacteremia episodes (10% and 90%, respectively) in both the training and test sets. Consequently, we assigned 8738 episodes to the training set and 3745 episodes to the test set (see Figure 2). Among the two datasets, episodes without missing BDG and PCT were identified for further analysis (see below).

Imputation and standardization

There are several methods to handle missing data in clinical research [40]. We decided to impute missing values by means of Multiple Imputation with Chained Equations (MICE), through a nearest neighbors method, which is one of the most used in clinical settings [39, 40]. Continuous features were standardized by subtracting their means and dividing them by their standard deviations.

Machine learning algorithms and performance evaluation

Since the outcome variable was dichotomous (candidemia yes vs. no), the differential diagnosis between candidemia and bacteremia was considered a

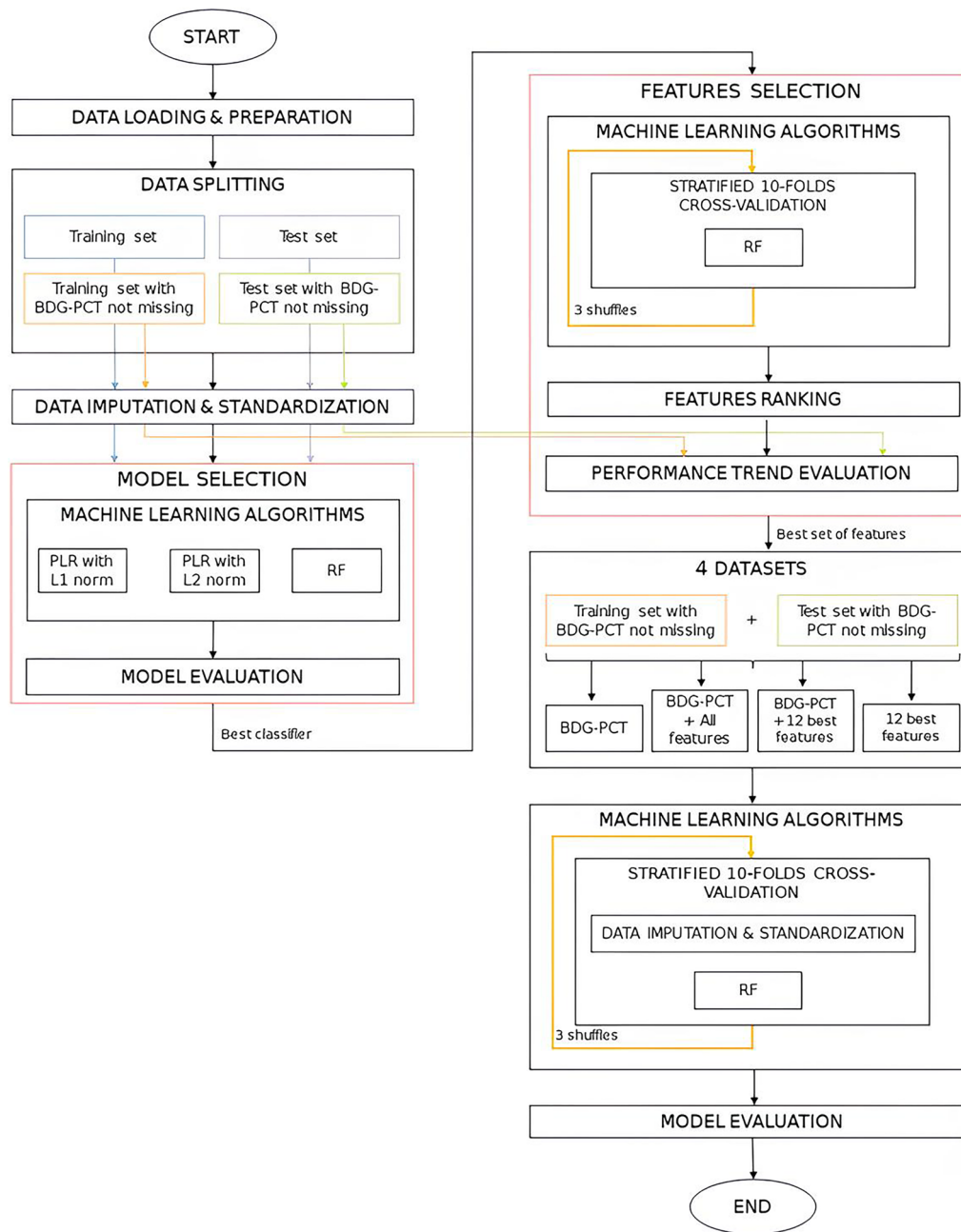


Figure 1. Complete diagram of the devised pipeline. The diagram can be divided into six main steps: Data Loading & Preparation, Data Splitting, Data Imputation & Standardization, Model Selection, Features Selection, Evaluation of model and best set of features on the complete subset of data with serum β -D-glucan (BDG) and procalcitonin (PCT) not missing. PLR: Penalized Logistic Regression; RF: Random Forest.

classification problem. The associations of training set features with candidemia were first descriptively tested in univariable analysis using the chi-square test (for categorical features) and the Kruskal-Wallis test (for continuous features), before imputation of missing values. Regarding multivariable analyses with ML algorithms, performed after imputation of missing values, we trained the following three classifiers to guarantee model explainability: (i) L2-penalized logistic regression (PLR with L2 norm); (ii) L1-penalized logistic regression

(PLR with L1 norm); and (iii) random forest. Details about the parameters selection for each model are provided in Table S7. To evaluate the performance of the classifiers in the training and test sets, a confusion matrix was built, and the diagnostic performance was assessed in terms of the following: (i) true skill statistic (TSS, also known as Youden Index); (ii) positive predictive value (PPV, also known as precision), (iii) negative predictive value (NPV), (iv) accuracy, (v) weighted F1-score, (vi) specificity, (vii) sensitivity (also known as

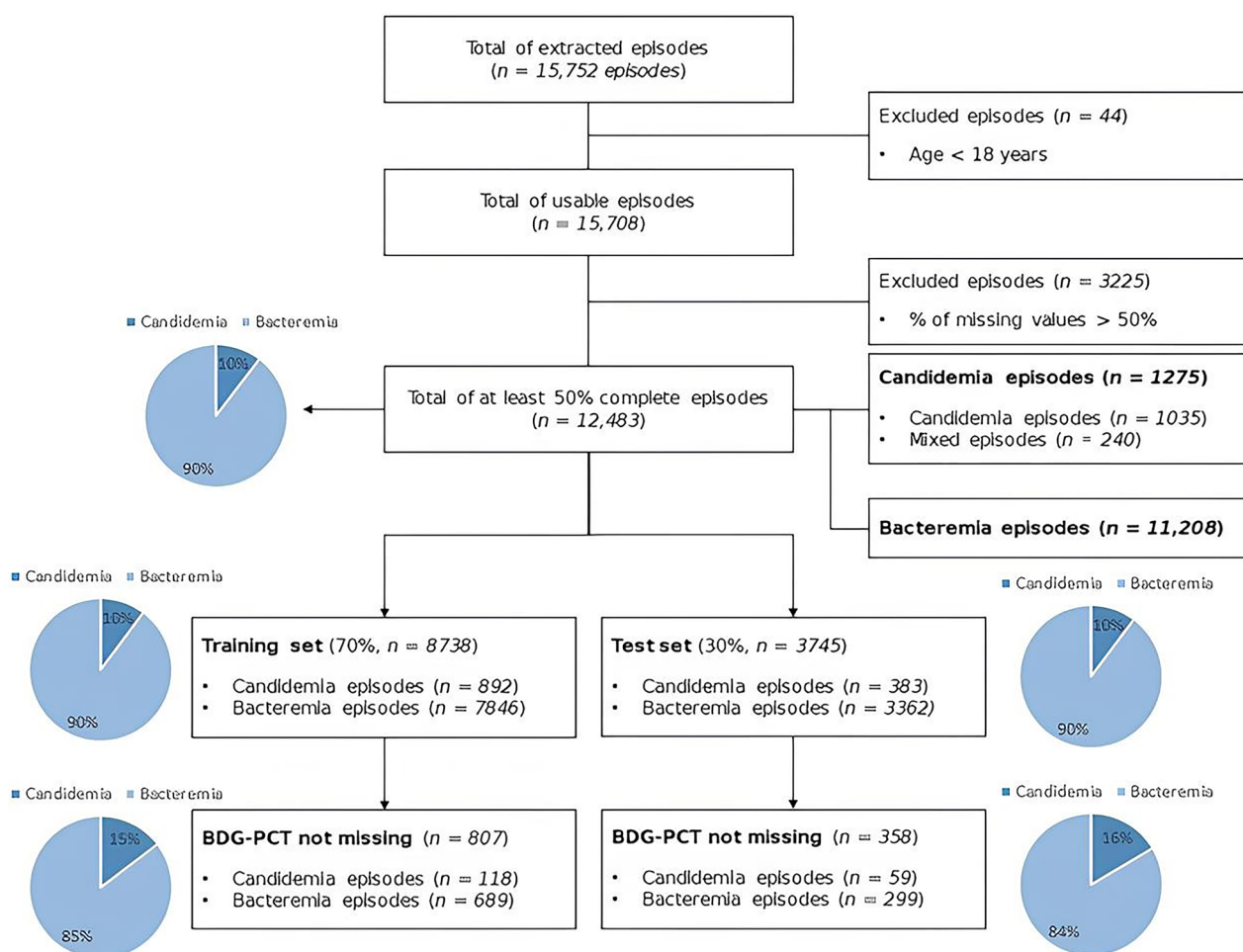


Figure 2. Complete process of dataset preparation and splitting. The proportion of candidemia and bacteremia episodes is constant, except for the two samples with serum β -D-glucan (BDG) and procalcitonin (PCT) not missing because they are directly derived from the previous two groups.

recall), (viii) negative likelihood ratio (LR⁻), and (ix) positive likelihood ratio (LR⁺). These measures depend on the threshold τ chosen to assign an episode to 0 or 1 according to the probability predicted by the classifier. By moving the threshold τ within the 0–1 interval, the sensitivity can be advantaged over specificity, or *vice versa*. In our study, we decided to favor sensitivity over specificity to minimize the number of false-negative results for candidemia. Indeed, not recognizing a true episode of candidemia was deemed the least desirable situation from a clinical perspective, since it could lead to perilous delays in antifungal treatment, in turn impacting survival. The sensitivity of BDG previously registered at our center was defined as the minimum performance requirement for the evaluated ML models [22, 23]. Consequently, we deemed any chosen τ value to fulfill the following condition: $sensitivity \geq specificity \geq 0.60$. To advantage sensitivity over specificity, we explored τ values of ≤ 0.5 . Eventually, the optimal τ value among those fulfilling the above

conditions was selected as the one that maximized TSS. For each classifier and for each explored threshold τ , a 10-fold cross-validation was performed using a user-defined score function built to maximize the TSS calculated on the validation set in order to choose the best set of hyper-parameters for each classifier. Based on the threshold τ , the predicted probabilities were then assigned a value of 0 or 1 to build the confusion matrix and compute performance measurements on both the training and test sets. Once the optimal threshold τ was chosen, changes from the pre-test probability (defined as the baseline prevalence of candidemia) to the post-test probability of candidemia for all classifiers were graphically summarized using Fagan's nomograms [41]. Further analyses were performed to verify whether the resulting best classifier (based on features other than BDG and PCT) could help improve the diagnostic performance of BDG and PCT. Specifically, the subgroup of episodes with no missing BDG and PCT values was identified and used

to run a new classification process including the features that had the greatest influence on the outcome on the entire dataset. To identify the most influential features, the permutation feature importance (PFI) was used. It employs a permutation approach to calculate a feature contribution coefficient in terms of change in the model measure of performance [40], evaluated in terms of TSS (which was the most relevant metric for our purpose, see above). Eventually, the PFI was repeated 30 times for stability of the resulting feature ranking, as shown in Figure 1.

Results

The characteristics and laboratory values of episodes of candidemia and bacteremia in the entire study population, training set, and test set are reported in Supplementary Table S1, whereas the results of descriptive univariable comparisons in the training set are shown in Table 1.

As shown in Table 1, an association with candidemia in univariable comparisons was observed for the following features: age, basophils, eosinophils, hematocrit, hemoglobin, red cell count, platelet count, aPTT, INR, prothrombin time, uric acid, ALP, total bilirubin, creatinine, GGT, LDH, urea, albumin, CRP, total proteins, and *Candida* colonization.

As shown in Supplementary Table S2 (for PLR with norm L1), Supplementary Table S3 (for PLR with norm L2), and Table 2 (for random forest), all classifiers showed the best performance in terms of TSS, fulfilling the condition $sensitivity \geq specificity \geq 0.60$, around $\tau=0.1$. Overall, the random forest classifier (with $\tau=0.1$) achieved the best performance, with a sensitivity of 0.98 and a specificity of 0.65 on the training set (TSS = 0.63) and a sensitivity of 0.74 and specificity 0.57 on the test set (TSS = 0.31). Fagan's nomograms also showed a better performance of random forest over PLR models in impacting post-test probabilities of candidemia in both the training and test sets (Supplementary Figure S2).

Overall, 1165 episodes (of which 177 were candidemia, 15%) out of the 12,483 episodes of the complete sample had available BDG and PCT values. To increase the stability of feature selection, a stratified 10-fold cross-validation for three different shuffles of the training set was performed, and the best set of hyperparameters was selected as the one that maximized the TSS. Stratified cross-validation allows the maintenance of the same percentage of positive samples in the training and validation sets. At each iteration, we evaluated the optimal value of the threshold

τ , which was 0.1 in most cases. Overall, 23/30 models met the condition $sensitivity \geq specificity \geq 0.60$ thus, we considered feature ranks only from those 23 models. To select features to be used in combination with BDG and PCT in the subgroup of patients with these tests available, we first counted the number of times each feature was assigned within the first 18 positions (i.e. the positions including all ranks of at least one feature; see Supplementary Figure S3). The final feature ranks resulting from this count are listed in Supplementary

Table 1. Descriptive univariable comparisons between candidemia and bacteremia in training set (performed before imputation of missing values).

| Features* | Candidemia episodes <i>n</i> =892 (10.2%) | Bacteremia episodes <i>n</i> =7846 (89.8%) | <i>p</i> |
|---|--|---|-------------------|
| Male sex, <i>n</i> (%) | 517 (58.0%) | 4615 (58.8%) | 0.6465 |
| <i>Candida</i> colonization, <i>n</i> (%) | 262/656 (39.9%) | 846/4293 (19.7%) | <0.0001 |
| Age, in years | 73 [61–81] | 71 [59–80] | 0.0075 |
| Basophil cell count, in cells $\times 10^9/L$ | 0.01 [0.00–0.04] | 0.01 [0.00–0.04] | 0.0253 |
| Eosinophil cell count, in cells $\times 10^9/L$ | 0.07 [0.00–0.20] | 0.03 [0.00–0.10] | <0.0001 |
| Lymphocyte cell count, in cells $\times 10^9/L$ | 0.80 [0.49–1.25] | 0.76 [0.40–1.23] | 0.0666 |
| Monocyte cell count, in cells $\times 10^9/L$ | 0.44 [0.23–0.70] | 0.44 [0.20–0.70] | 0.4728 |
| Neutrophil cell count, in cells $\times 10^9/L$ | 7.05 [4.50–10.49] | 7.24 [3.71–11.70] | 0.7894 |
| Hematocrit, in % | 28.9 [26.2–32.4] | 30.3 [26.4–35.1] | <0.0001 |
| Hemoglobin, in g/L | 95 [86–107] | 99 [87–116] | <0.0001 |
| White cell count, in cells $\times 10^9/L$ | 8.90 [6.35–12.98] | 9.38 [5.41–14.10] | 0.8405 |
| Red cell count, in cells $\times 10^9/L$ | 3.3 [3.0–3.7] | 3.4 [3.0–4.0] | <0.0001 |
| Platelet count, in cells $\times 10^9/L$ | 200 [117–299] | 170 [90–256] | <0.0001 |
| aPTT, in sec | 35.10 [30.80–42.15] | 33.80 [29.80–39.90] | <0.0001 |
| INR | 1.20 [1.10–1.35] | 1.16 [1.06–1.32] | <0.0001 |
| Prothrombin time, as activity % | 76 [63–88] | 80 [65–93] | <0.0001 |
| Uric acid, in mg/dL | 3.67 [2.37–5.67] | 4.11 [2.69–5.97] | 0.0014 |
| ALP, in U/L | 122.00 [79.25–193.00] | 100.00 [70.00–161.00] | <0.0001 |
| ALT, in U/L | 28 [16–52] | 27 [16–52] | 0.5294 |
| AST, in U/L | 27 [17–43] | 25 [16–46] | 0.4539 |
| Total bilirubin, in mg/dL | 0.63 [0.41–1.11] | 0.70 [0.45–1.27] | 0.0014 |
| Creatinine, in mg/dL | 1.0 [0.7–1.6] | 1.0 [0.7–1.6] | 0.0479 |
| GGT, in U/L | 80.5 [38.0–178.0] | 65.0 [29.0–165.0] | 0.0002 |
| LDH, in U/L | 242 [194–328] | 232 [180–319] | 0.0062 |
| Urea, in mg/dL | 57 [35–102] | 51 [34–86] | 0.0008 |
| Glucose, in mg/dL | 124 [102–152] | 120 [99–152] | 0.0615 |
| Albumin, in g/L | 25.5 [21.6–30.9] | 27.1 [22.5–33.0] | <0.0001 |
| CRP, in mg/dL | 92.45 [43.48–150.75] | 98.40 [44.03–165.00] | 0.0412 |
| Total proteins, in g/L | 61.0 [55.9–68.2] | 62.4 [56.0–69.0] | 0.0498 |

Results are presented as median [interquartile range (IQR)] unless otherwise indicated. Statistically significant *p* values ($p < 0.05$) are reported in bold text. aPTT: activated partial thromboplastin time; INR: international normalized ratio; ALP: alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; GGT: gamma-glutamyl transferase; LDH: lactate dehydrogenase; CRP: C-reactive protein. *Denominator for frequencies provided in case of missing values (excluded from the denominator).

Table 2. Evaluation of the diagnostic performance for candidemia of random forest classifier on both the training set and the test set as the threshold τ changes from 0 to 0.5 (with step 0.05).

| | | Random forest | | | | | | | | | | |
|--------------|-----------|---------------|-------|--------------|-------|--------|---------|----------|-------|--------|--------|--------|
| τ | | 0 | 0.05 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 | 0.4 | 0.45 | 0.5 |
| Training set | TSS | 0 | 0.260 | 0.626 | 0.644 | 0.504 | 0.428 | 0.332 | 0.267 | 0.184 | 0.126 | 0.092 |
| | Precision | 0.010 | 0.909 | 0.919 | 0.920 | 0.932 | 0.941 | 0.936 | 0.931 | 0.924 | 0.919 | 0.916 |
| | NPV | 1 | 0.997 | 0.996 | 0.967 | 0.947 | 0.939 | 0.929 | 0.923 | 0.915 | 0.909 | 0.906 |
| | Accuracy | 0.102 | 0.341 | 0.681 | 0.902 | 0.937 | 0.940 | 0.932 | 0.925 | 0.917 | 0.911 | 0.907 |
| | F1-score | 0.019 | 0.402 | 0.744 | 0.909 | 0.931 | 0.930 | 0.916 | 0.905 | 0.890 | 0.878 | 0.871 |
| | SP | 0 | 0.267 | 0.647 | 0.922 | 0.985 | 0.998 | 1 | 1 | 1 | 1 | 1 |
| | SE | 1 | 0.993 | 0.979 | 0.722 | 0.519 | 0.430 | 0.332 | 0.267 | 0.184 | 0.126 | 0.092 |
| | LR- | 0 | 0.025 | 0.033 | 0.302 | 0.488 | 0.571 | 0.668 | 0.733 | 0.816 | 0.874 | 0.908 |
| | LR+ | 1 | 1.354 | 2.774 | 9.256 | 34.223 | 211.103 | 2603.605 | Inf. | Inf. | Inf. | Inf. |
| | Test set | TSS | 0 | 0.163 | 0.305 | 0.229 | 0.134 | 0.093 | 0.068 | 0.032 | 0.021 | 0.022 |
| Precision | | 0.010 | 0.879 | 0.869 | 0.853 | 0.846 | 0.848 | 0.855 | 0.856 | 0.865 | 0.879 | 0.880 |
| NPV | | 1 | 0.965 | 0.950 | 0.923 | 0.911 | 0.907 | 0.904 | 0.901 | 0.900 | 0.900 | 0.899 |
| Accuracy | | 0.102 | 0.310 | 0.583 | 0.817 | 0.866 | 0.886 | 0.895 | 0.897 | 0.898 | 0.899 | 0.899 |
| F1-score | | 0.019 | 0.367 | 0.664 | 0.833 | 0.855 | 0.861 | 0.861 | 0.856 | 0.854 | 0.855 | 0.852 |
| SP | | 0 | 0.240 | 0.566 | 0.869 | 0.943 | 0.973 | 0.988 | 0.995 | 0.998 | 0.999 | 0.999 |
| SE | | 1 | 0.924 | 0.739 | 0.360 | 0.191 | 0.120 | 0.081 | 0.037 | 0.023 | 0.023 | 0.013 |
| LR- | | 0 | 0.315 | 0.462 | 0.736 | 0.858 | 0.904 | 0.931 | 0.968 | 0.979 | 0.978 | 0.988 |
| LR+ | | 1 | 1.216 | 1.702 | 2.747 | 3.373 | 4.437 | 6.479 | 7.681 | 11.286 | 19.751 | 21.945 |

Values in bold text refer to the best performance on the training set satisfying the predefined clinical condition (see text). TSS: true skill statistics; NPV: negative predictive value; SP: specificity; SE: sensitivity; LR-: negative likelihood ratio; LR+: positive likelihood ratio.

Table S4. Starting from the one with the highest rank, we progressively added each feature to BDG and PCT and trained a random forest model on the subgroup of patients with available BDG and PCT results. The diagnostic performances for candidemia of the random forest classifier, trained at different steps (i.e. with a progressively higher number of features in addition to BDG and PCT), in a set of 807 episodes (i.e. episodes with non-missing BDG and PCT values in the training set) are shown in [Supplementary Figure S4](#). As shown in the figure, the best TSS values, provided that the condition $sensitivity \geq specificity \geq 0.60$ was satisfied, were mostly obtained for thresholds τ between 0.15 and 0.2. Therefore, we decided to improve the search for the best subset of features by exploring thresholds τ between 0.15 and 0.2. As shown in [Supplementary Figure S5](#) and [Supplementary Table S5](#), the condition $sensitivity \geq specificity \geq 0.60$ was satisfied for thresholds τ between 0.15 and 0.18 (excluding subsets with a low number of features). Among these, we selected 0.175 as the optimal threshold τ , as it showed the highest average TSS value (see [Supplementary Table S5](#)), and from a number of features equal to 9–12 onwards the performance measures reached an almost flat trend (in line with the saturation of the model). With a number of features $N=12$, the TSS value reached a value greater than 0.95, and $N=12$ was thus considered for further analysis.

In addition to BDG and PCT, the 12 features included in the model were eosinophil count, platelet count, neutrophil cell count, hematocrit, uric acid, monocyte cell count, hemoglobin, urea, albumin, lymphocyte cell count, white cell count, and prothrombin.

[Figure 3](#), [Supplementary Figure S6](#), [Supplementary Table S6](#), and [Table 3](#) show the performances of the random forest classifier when trained in the subgroup of patients with available BDG and PCT values considering the following features: (i) only BDG and PCT (*BDG-PCT*); (ii) all available features (*BDG-PCT+All features*); (iii) BDG, PCT, and the additional 12 features (*BDG-PCT + 12 best features*); and (iv) only the additional 12 features without BDG and PCT (*12 best features*).

As shown in [Figure 3](#), the performance measures of the classifier (*BDG-PCT + 12 best features*) were the highest in most cases; however, as reported in [Table 3](#), statistically significant differences were observed when the model (*BDG-PCT + 12 best features*) was compared with the model (*12 best features*), but not when the model (*BDG-PCT + 12 best features*) was compared with model (*BDG-PCT*). Compared to the model (*BDG-PCT + 12 best features*), the performance measurements of the model (*BDG-PCT+All features*) tended to be worse, supporting feature selection as a reasonable way to obtain a more parsimonious and less expensive model for training.

Discussion

In this study, a random forest classifier performed better than penalized logistic regression for the early diagnosis of candidemia after training on laboratory and microbiological features in a large dataset of 12,483 episodes of candidemia or bacteremia.

In a sample of 295 patients hospitalized in medical wards with candidemia and bacteremia, Ripoli et al.

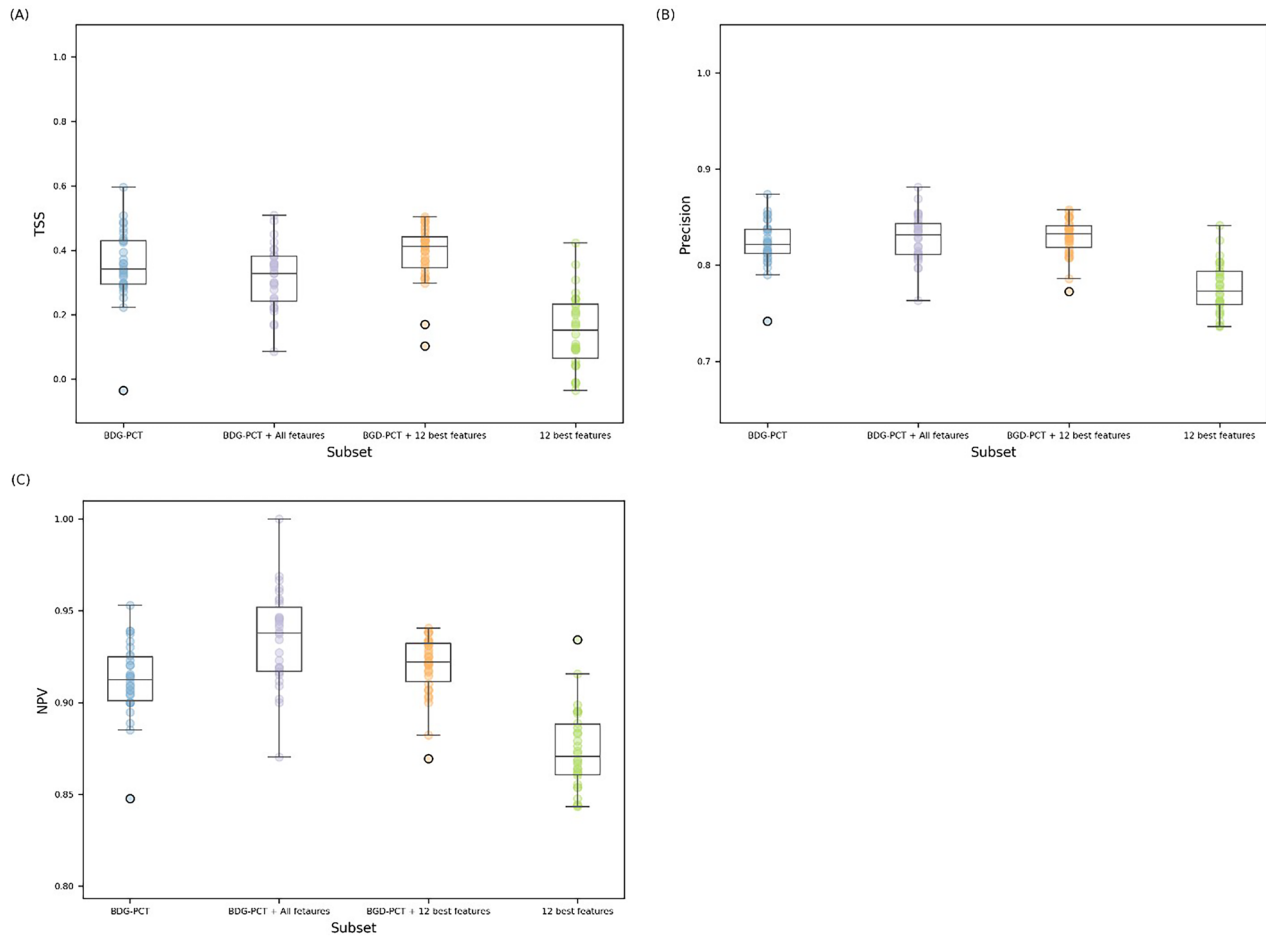


Figure 3. Classification performances of random forest trained with 4 different set of features (BDG-PCT, BDG-PCT+all features, BDG-PCT + 12 best features, 12 best features). Classification performances were evaluated in terms of: (A) true skill statistic (TSS), (B) Precision, and (C) negative predictive value (NPV). Each box plot shows results obtained on the validation set over the 10 folds of cross-validation and 3 shuffles (30 points in total).

Table 3. p -Values resulting from the statistical analysis performed with Friedman and Wilcoxon tests.

| Measure of performance | Friedman | Wilcoxon | | | | | |
|------------------------|-------------------|-----------------------------------|-----------------------------|-----------------------------|--|---|---|
| | Joint comparison | BDG-PCT vs BDG-PCT + All features | BDG-PCT vs 12 best features | BDG-PCT vs 12 best features | BDG-PCT + All features vs BDG-PCT + 12 best features | BDG-PCT + All features vs. 12 best features | BDG-PCT + 12 best features vs. 12 best features |
| TSS | <0.0001 | 0.1981 | 0.0266 | <0.0001 | 0.0043 | <0.0001 | <0.0001 |
| Precision | <0.0001 | 0.2801 | 0.0455 | <0.0001 | 0.8872 | <0.0001 | <0.0001 |
| NPV | <0.0001 | 0.0006 | 0.0190 | <0.0001 | 0.0120 | <0.0001 | <0.0001 |
| Accuracy | <0.0001 | <0.0001 | 0.8475 | 0.0015 | <0.0001 | <0.0001 | 0.0003 |
| F1-score | <0.0001 | <0.0001 | 0.8036 | 0.0001 | <0.0001 | <0.0001 | <0.0001 |
| SP | <0.0001 | <0.0001 | 0.2642 | 0.7303 | <0.0001 | <0.0001 | 0.6109 |
| SE | <0.0001 | <0.0001 | 0.0195 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |
| LR- | <0.0001 | 0.0007 | 0.0179 | <0.0001 | 0.0155 | <0.0001 | <0.0001 |
| LR+ | <0.0001 | <0.0001 | 0.5666 | <0.0001 | <0.0001 | 0.8236 | <0.0001 |

The analyses were performed to assess if the distribution of the measures of performance (obtained through a stratified ten-fold cross-validation for three shuffles, i.e. collecting each value of the metrics on each validation set) in the four subsets of features under consideration differed significantly. If the Friedman test for the joint comparison of the four subgroups failed for one measure of performance, each pair comparison between two subgroups was evaluated with the Wilcoxon test. TSS: true skill statistics; NPV: negative predictive value; SP: specificity; SE: sensitivity; LR-: negative likelihood ratio; LR+: positive likelihood ratio. Bonferroni correction was used to adjust for multiple testing: each pair comparison was considered significant if the relative p -value was less than $\alpha^* = 0.05/6 = 0.008$ (statistically significant p values are reported in bold text).

also previously assessed the performance of a random forest classifier for the prediction of candidemia versus controls with bacteremia (matched according to the time at risk elapsed from admission to bloodstream

infection), showing 84% sensitivity and 91% specificity after training based on 42 clinical features [30]. In another study including 501 patients with candidemia and 2000 controls without candidemia, a random

forest classifier based on clinical and laboratory values achieved 90% sensitivity and 72% specificity [29]. There are some important differences between the approaches employed in these studies and ours. First, the authors of the two studies described above included in their training some well-known clinical and microbiological predictors of candidemia (e.g. previous broad-spectrum antibiotic therapy, total parenteral nutrition, central venous catheters, and *Candida* score), whereas we relied only on laboratory and microbiological variables. While on the one hand the inclusion of clinical features can improve predictive ability, manual collection of clinical variables from electronic medical records (EMRs) is a highly time-consuming task that frequently precludes the collection of large samples. A possible solution to this limitation would be to automatically extract data about clinical features, frequently organized in free text, from the EMRs, in order to build large datasets without the need for (and beyond the possibility of) manual collection. Our group is currently working on the use of natural language processing (NLP) models to achieve this aim, although the applicability of similar algorithms for the extraction of common clinical features employed for predicting candidemia is still preliminary. On the other hand, it could not be excluded that training ML models only on large datasets of automatically extracted laboratory and microbiological features could improve current diagnostic algorithms for early candidemia, considering that it had long been very unusual to exploit very large datasets for this specific purpose. In this regard, although trained on laboratory markers usually considered nonspecific for candidemia in clinical practice, the best classifier (random forest) in our study achieved 74% sensitivity and 57% specificity in the test sets. Certainly, this is not sufficient to employ the developed algorithm alone for a reliable diagnosis of candidemia in clinical practice; however, as a hypothesis-generating result, it suggests further exploration of the possible role of non-specific markers in improving the diagnostic performance of other more specific markers and/or clinical variables for candidemia.

For the above reason, by exploiting what was learned from all included episodes (i.e. feature selection was based on the performance of random forest in the entire dataset), we assessed whether the developed algorithm based on non-specific markers could improve the diagnostic performance of BDG and PCT in the subgroup of patients with available values for these two markers. In this regard, although the addition of selected features to BDG and PCT almost consistently showed better median performance measures when compared with BDG and PCT alone, the

differences were not statistically significant. Consequently, although encouraging, our results do not currently support this practice. However, based on the results themselves (numerically better values, albeit not statistically significant), the following two hypotheses could merit exploration: (i) whether other ML classifiers exploiting a more articulated relationship between nonspecific markers, for example, neural networks, might lead to a better diagnostic performance of nonspecific laboratory markers, which in turn significantly improves the diagnostic performance of BDG and PCT alone (although with all the issues connected to the explainability of neural-network-based models); and (ii) whether the addition of specific and non-specific clinical features in the training of the classifiers could further and consistently improve current diagnostic algorithms for candidemia. Of note, the issue of explainability also applies to random forest models, although in our study we attempted to partly explain the model by limiting collection to laboratory and microbiological variables, and through the ranking in the feature selection process.

This study had some important limitations that should be acknowledged. First, as reported above, we did not include clinical features in our large, automatically extracted dataset, because their accurate extraction from EMRs free text through NLP algorithms is still under study. However, rather than a limitation, we consider this as a future development of our approach. A second limitation was the large number of missing values, which precluded the evaluation of the impact of sequential variations over days of laboratory values on the prediction of candidemia. Nonetheless, we extracted all the results of the tests performed in clinical practice; thus, missing values do not reflect the inability to collect information, but rather, a true lack of data. Because this is the rule rather than the exception in real-life practice, any developed ML model should be able to consider and deal with this issue. A third important limitation, connected to the previous one, is that BDG and PCT values were available for a few events; thus, the additional contribution (to the diagnostic performance of BDG and PCT) of the developed predictive model based on less specific markers was assessed only in the subgroup of events with available BDG and PCT values. However, again, this stresses the unavoidability of missing data in real practice and supports our tentative training of the classifier in a larger dataset without BDG and PCT information, and then to exploit feature ranking results (obtained from the large dataset) in the subgroup of patients with available BDG and PCT values. A fourth limitation is the

imbalance between candidemia and bacteremia episodes (10% vs. 90%). In this regard, even though there are various techniques to balance a dataset, e.g. augmentation, undersampling, oversampling such as Synthetic Minority Over-sampling TEchnique (SMOTE), in the design phase of the study we decided to stratify the train-test split, keeping the original percentages of the episodes, due to the consequent search for the value of threshold τ to meet the desired clinical condition.

In conclusion, random forest classifiers trained on large datasets of automatically extracted laboratory and microbiological data show the potential to improve current diagnostic algorithms for the early diagnosis of candidemia, although they did not significantly improve the isolated diagnostic performance of BDG and PCT in our study. Further development of our approach through the exploration of additional ML classifiers and the implementation of clinical features automatically extracted from the EMRs by NLP algorithms could help achieve crucial improvements.

Author contributions

Conceptualization: D.R.G.; methodology: D.R.G., C.M., S.M., S.G., C.C., A.S., S.P., M.T.; formal analysis and investigation: C.M., S.M., S.G.; data collection: C.R., G.B., A.L., C.M., S.M., A.V.; writing—original draft preparation, D.R.G., C.M., S.M.; writing—review and editing, D.R.G., C.M., S.M., S.G., C.R., G.B., A.L., A.V., M.M., M.T., S.P., A.S., A.D.B., A.M., C.C., M.G., M.B.; supervision: D.R.G., M.M., A.D.B., A.M., S.P., C.C., M.G., M.B.

All authors have read and approved the final version of the manuscript. All authors agree to be accountable for all aspects of this study.

Disclosure statement

Daniele Roberto Giacobbe is a section editor of *Annals of Medicine*. He had no role in the editorial process and decisions regarding the present article. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, and Gilead Italia and speaker/advisor fees from Pfizer, Menarini, and Tillotts Pharma. Outside the submitted work, Matteo Bassetti received funding for scientific advisory boards, travel, and speaker honoraria from Angelini, Astellas, BioMérieux, Cidara, Gilead, Menarini, MSD, Pfizer, Shionogi, Tetrphase, and Nabriva. Anna Marchese reports investigator-initiated grants from Gilead, Italy. The other authors have no conflicts of interest to declare.

Funding

The AUTO-CAND project was supported by Pfizer Global Medical Grants (GMG) for General Research [Project Tracking Number 69511763]. The funder had no role in the study design, data collection and analysis, decision to publish, and preparation of the manuscript.

ORCID

Daniele Roberto Giacobbe  <http://orcid.org/0000-0003-2385-1759>

Cristina Marelli  <http://orcid.org/0000-0003-1510-1819>

Mauro Giacomini  <http://orcid.org/0000-0001-5646-2034>

Data availability statement

The data presented in this study will be available from the corresponding author upon reasonable request, provided all regulatory and privacy requirements are fulfilled.

References

- [1] Bouza E, Munoz P. Epidemiology of candidemia in intensive care units. *Int J Antimicrob Agents*. 2008;32 Suppl 2(Suppl 2):1–12. doi: [10.1016/S0924-8579\(08\)70006-2](https://doi.org/10.1016/S0924-8579(08)70006-2).
- [2] Bounoux ME, Kac G, Aegerter P, et al. Candidemia and candiduria in critically ill patients admitted to intensive care units in France: incidence, molecular diversity, management and outcome. *Intensive Care Med*. 2008;34(2):292–299. doi: [10.1007/s00134-007-0865-y](https://doi.org/10.1007/s00134-007-0865-y).
- [3] Bassetti M, Righi E, Analdi F, et al. A multicenter study of septic shock due to candidemia: outcomes and predictors of mortality. *Intensive Care Med*. 2014;40(6):839–845. doi: [10.1007/s00134-014-3310-z](https://doi.org/10.1007/s00134-014-3310-z).
- [4] Bassetti M, Giacobbe DR, Vena A, et al. Incidence and outcome of invasive candidiasis in intensive care units (ICUs) in Europe: results of the EUCANDICU project. *Crit Care*. 2019;23(1):219. doi: [10.1186/s13054-019-2497-3](https://doi.org/10.1186/s13054-019-2497-3).
- [5] Wisplinghoff H, Bischoff T, Tallent SM, et al. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin Infect Dis*. 2004;39(3):309–317. doi: [10.1086/421946](https://doi.org/10.1086/421946).
- [6] Pappas PG, Lionakis MS, Arendrup MC, et al. Invasive candidiasis. *Nat Rev Dis Primers*. 2018;4(1):18026. doi: [10.1038/nrdp.2018.26](https://doi.org/10.1038/nrdp.2018.26).
- [7] Bassetti M, Giacobbe DR, Vena A, et al. Diagnosis and treatment of candidemia in the intensive care unit. *Semin Respir Crit Care Med*. 2019;40(4):524–539. doi: [10.1055/s-0039-1693704](https://doi.org/10.1055/s-0039-1693704).
- [8] Garey KW, Rege M, Pai MP, et al. Time to initiation of fluconazole therapy impacts mortality in patients with candidemia: a multi-institutional study. *Clin Infect Dis*. 2006;43(1):25–31. doi: [10.1086/504810](https://doi.org/10.1086/504810).
- [9] Bassetti M, Molinari MP, Mussap M, et al. Candidaemia in internal medicine departments: the burden of a rising problem. *Clin Microbiol Infect*. 2013;19(6):E281–E284. doi: [10.1111/1469-0691.12155](https://doi.org/10.1111/1469-0691.12155).
- [10] Johnson MD, Lewis RE, Dodds Ashley ES, et al. Core recommendations for antifungal stewardship: a statement of the mycoses study group education and research consortium. *J Infect Dis*. 2020;222(Suppl 3):S175–S198. doi: [10.1093/infdis/jiaa394](https://doi.org/10.1093/infdis/jiaa394).
- [11] Giacobbe DR, Signori A, Tumbarello M, et al. Desirability of outcome ranking (DOOR) for comparing diagnostic tools and early therapeutic choices in patients with suspected candidemia. *Eur J Clin Microbiol Infect Dis*. 2019;38(2):413–417. doi: [10.1007/s10096-018-3441-1](https://doi.org/10.1007/s10096-018-3441-1).
- [12] Yera H, Sendid B, Francois N, et al. Contribution of serological tests and blood culture to the early diagnosis

- of systemic candidiasis. *Eur J Clin Microbiol Infect Dis*. 2001;20(12):864–870. doi: [10.1007/s100960100629](https://doi.org/10.1007/s100960100629).
- [13] White PL, Archer AE, Barnes RA. Comparison of non-culture-based methods for detection of systemic fungal infections, with an emphasis on invasive *Candida* infections. *J Clin Microbiol*. 2005;43(5):2181–2187. doi: [10.1128/JCM.43.5.2181-2187.2005](https://doi.org/10.1128/JCM.43.5.2181-2187.2005).
- [14] Wei S, Wu T, Wu Y, et al. Diagnostic accuracy of *Candida albicans* germ tube antibody for invasive candidiasis: systematic review and meta-analysis. *Diagn Microbiol Infect Dis*. 2019;93(4):339–345. doi: [10.1016/j.diagmicrobio.2018.10.017](https://doi.org/10.1016/j.diagmicrobio.2018.10.017).
- [15] Walker B, Powers-Fletcher MV, Schmidt RL, et al. Cost-effectiveness analysis of multiplex PCR with magnetic resonance detection versus empiric or blood culture-directed therapy for management of suspected candidemia. *J Clin Microbiol*. 2016;54(3):718–726. doi: [10.1128/JCM.02971-15](https://doi.org/10.1128/JCM.02971-15).
- [16] Rouze A, Loidant S, Poissy J, et al. Biomarker-based strategy for early discontinuation of empirical antifungal treatment in critically ill patients: a randomized controlled trial. *Intensive Care Med*. 2017;43(11):1668–1677. doi: [10.1007/s00134-017-4932-8](https://doi.org/10.1007/s00134-017-4932-8).
- [17] Raineri SM, Cortegiani A, Vitale F, et al. Procalcitonin for the diagnosis of invasive candidiasis: what is the evidence? *J Intensive Care*. 2017;5(1):58. doi: [10.1186/s40560-017-0252-x](https://doi.org/10.1186/s40560-017-0252-x).
- [18] Posteraro B, Tumbarello M, De Pascale G, et al. (1,3)-beta-d-Glucan-based antifungal treatment in critically ill adults at high risk of candidaemia: an observational study. *J Antimicrob Chemother*. 2016;71(8):2262–2269. doi: [10.1093/jac/dkw112](https://doi.org/10.1093/jac/dkw112).
- [19] Posteraro B, De Pascale G, Tumbarello M, et al. Early diagnosis of candidemia in intensive care unit patients with sepsis: a prospective comparison of (1→3)-beta-D-glucan assay, *Candida* score, and colonization index. *Crit Care*. 2011;15(5):R249. doi: [10.1186/cc10507](https://doi.org/10.1186/cc10507).
- [20] Paphitou NI, Ostrosky-Zeichner L, Rex JH. Rules for identifying patients at increased risk for candidal infections in the surgical intensive care unit: approach to developing practical criteria for systematic use in antifungal prophylaxis trials. *Med Mycol*. 2005;43(3):235–243. doi: [10.1080/13693780410001731619](https://doi.org/10.1080/13693780410001731619).
- [21] Ostrosky-Zeichner L, Sable C, Sobel J, et al. Multicenter retrospective development and validation of a clinical prediction rule for nosocomial invasive candidiasis in the intensive care setting. *Eur J Clin Microbiol Infect Dis*. 2007;26(4):271–276. doi: [10.1007/s10096-007-0270-z](https://doi.org/10.1007/s10096-007-0270-z).
- [22] Mikulska M, Magnasco L, Signori A, et al. Sensitivity of serum Beta-D-Glucan in candidemia according to *Candida* species epidemiology in critically ill patients admitted to the Intensive Care Unit. *J Fungi (Basel)*. 2022;8(9):921. doi: [10.3390/jof8090921](https://doi.org/10.3390/jof8090921).
- [23] Mikulska M, Giacobbe DR, Furfaro E, et al. Lower sensitivity of serum (1,3)-beta-d-glucan for the diagnosis of candidaemia due to *Candida parapsilosis*. *Clin Microbiol Infect*. 2016;22(7):646.e5–646.e8. doi: [10.1016/j.cmi.2016.05.020](https://doi.org/10.1016/j.cmi.2016.05.020).
- [24] Martinez-Jimenez MC, Munoz P, Valerio M, et al. *Candida* biomarkers in patients with candidaemia and bacteraemia. *J Antimicrob Chemother*. 2015;70(8):2354–2361. Augdoi: [10.1093/jac/dkv090](https://doi.org/10.1093/jac/dkv090).
- [25] Leon C, Ruiz-Santana S, Saavedra P, et al. A bedside scoring system ("Candida score") for early antifungal treatment in nonneutropenic critically ill patients with *Candida* colonization. *Crit Care Med*. 2006;34(3):730–737. doi: [10.1097/01.CCM.0000202208.37364.7D](https://doi.org/10.1097/01.CCM.0000202208.37364.7D).
- [26] Giannella M, Paolucci M, Roncarati G, et al. Potential role of T2Candida in the management of empirical antifungal treatment in patients at high risk of candidaemia: a pilot single-Centre study. *J Antimicrob Chemother*. 2018;73(10):2856–2859. doi: [10.1093/jac/dky247](https://doi.org/10.1093/jac/dky247).
- [27] Giacobbe DR, Mikulska M, Tumbarello M, et al. Combined use of serum (1,3)-beta-D-glucan and procalcitonin for the early differential diagnosis between candidaemia and bacteraemia in intensive care units. *Crit Care*. 2017;21(1):176. doi: [10.1186/s13054-017-1763-5](https://doi.org/10.1186/s13054-017-1763-5).
- [28] Arendrup MC, Andersen JS, Holten MK, et al. Diagnostic performance of T2Candida among ICU patients with risk factors for invasive candidiasis. *Open Forum Infect Dis*. 2019;6(5):ofz136. doi: [10.1093/ofid/ofz136](https://doi.org/10.1093/ofid/ofz136).
- [29] Yoo J, Kim SH, Hur S, et al. Candidemia risk prediction (CanDETEC) model for patients with malignancy: model development and validation in a single-center retrospective study. *JMIR Med Inform*. 2021;9(7):e24651. doi: [10.2196/24651](https://doi.org/10.2196/24651).
- [30] Ripoli A, Sozio E, Sbrana F, et al. Personalized machine learning approach to predict candidemia in medical wards. *Infection*. 2020;48(5):749–759. doi: [10.1007/s15010-020-01488-3](https://doi.org/10.1007/s15010-020-01488-3).
- [31] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019;20(5):e262–e273. doi: [10.1016/S1470-2045\(19\)30149-4](https://doi.org/10.1016/S1470-2045(19)30149-4).
- [32] Giacobbe DR, Signori A, Del Puente F, et al. Early detection of sepsis With machine learning techniques: a brief clinical perspective. *Front Med (Lausanne)*. 2021;8:617486. doi: [10.3389/fmed.2021.617486](https://doi.org/10.3389/fmed.2021.617486).
- [33] Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317–1318. doi: [10.1001/jama.2017.18391](https://doi.org/10.1001/jama.2017.18391).
- [34] Mora S, Giacobbe DR, Russo C, et al. A wide database for future studies aimed at improving early recognition of candidemia. *Stud Health Technol Inform*. 2021;281:1081–1082.
- [35] Giacobbe DR, Mora S, Signori A, et al. Validation of an automated system for the extraction of a wide dataset for clinical studies aimed at improving the early diagnosis of candidemia. *Diagnostics (Basel)*. 2023;13(5):961. doi: [10.3390/diagnostics13050961](https://doi.org/10.3390/diagnostics13050961).
- [36] Kim SH, Yoon YK, Kim MJ, et al. Risk factors for and clinical implications of mixed candida/bacterial bloodstream infections. *Clin Microbiol Infect*. 2013;19(1):62–68. Jandoi: [10.1111/j.1469-0691.2012.03906.x](https://doi.org/10.1111/j.1469-0691.2012.03906.x).
- [37] ECDC. Healthcare-Associated Infections Acquired in Intensive Care Units. Annual Epidemiological Report for 2017 [cited 2023 Jun 24]. https://www.ecdc.europa.eu/sites/default/files/documents/AER_for_2017-HAI.pdf
- [38] AUTO-CAND source code. <https://github.com/AUTOCAND/AUTO-CAND.git>
- [39] Heymans MW, Twisk JWR. Handling missing data in clinical research. *J Clin Epidemiol*. 2022;151:185–188. doi: [10.1016/j.jclinepi.2022.08.016](https://doi.org/10.1016/j.jclinepi.2022.08.016).
- [40] Faisal S, Tutz G. Multiple imputation using nearest neighbor methods. *Inf Sci*. 2021;570:500–516. doi: [10.1016/j.ins.2021.04.009](https://doi.org/10.1016/j.ins.2021.04.009).
- [41] Fagan TJ. Letter: nomogram for bayes theorem. *N Engl J Med*. 1975;293(5):257.