

A Wide Database for Future Studies Aimed at Improving Early Recognition of Candidemia

Sara MORA^{a,1}, Daniele Roberto GIACOBBE^{b,c}, Chiara RUSSO^{b,c}, Elia DIANA^d,
Alessio SIGNORI^b, Luca CARMISCIANO^b,
Matteo BASSETTI^{b,c} and Mauro GIACOMINI^a

^aDepartment of Informatics, Bioengineering, Robotics and System Engineering

^bDepartment of Health Sciences, University of Genoa, Genoa, Italy

^cSan Martino Polyclinic Hospital – IRCCS, Genoa, Italy

^dSchool of Medicine, University of Genoa, Genoa, Italy

Abstract. Invasive candidiasis is associated with high morbidity and mortality in critically ill patients, i.e. patients admitted to *Intensive Care Units* (ICUs) or in surgical wards. There are no clinical signs or specific symptoms and even though early diagnosis risk scores and rapid tests are available, none of such strategies has an equally-optimal level of sensitivity and specificity. In the era of *Electronic Health Records* (EHRs), several clinical studies exploited *Machine Learning* (ML) models and large database of features to improve the diagnosis accuracy. The main aim of this work is to build a wide dataset which can be exploited to apply ML models to further improve the early recognition of candidemia at the bedside of patients with compatible signs and symptoms.

Keywords. Candidemia, Bacteremia, Early Diagnosis, Relational Database, EHR

1. Introduction

Invasive candidiasis is associated with high morbidity and mortality in critically ill patients, i.e. patients admitted to ICUs or in surgical wards. Candidemia is the fourth most frequent healthcare-associated *bloodstream infection* according to some estimates [1,2]. Key points in the management of candidemia are the early recognition and the consequent rapid initiation of an appropriate antifungal therapy. The problem with early recognition is that the clinical presentation of candidemia is indistinguishable from that of bacteremia. Early diagnosis through risk scores and/or rapid tests is possible, but further improvement in diagnostic accuracy may be done for example by exploiting the use of ML algorithms on large database of features extracted from EHRs [3]. The aim of this paper is to present the database structure and automatic data transfer system that will be used to collect pseudonymous and high-quality data of patients with episodes of candidemia or bacteremia. The scope is to create a big sample (thousands of episodes) that will be used together with ML methods to further improve the early recognition of candidemia at the bedside of patients with compatible signs and symptoms.

¹ Corresponding Author, Sara Mora, DIBRIS University of Genoa, Italy; E-mail: sara.mora@edu.unige.it.

2. Methods

The complete list of parameters included in the protocol of data collection was approved by the Liguria ethics committee (71/2020-10323), after a process of pseudonymization. The protocol included two main types of data: Laboratory exams results and Microbiological culture results. So, to organize all the specific information in a document that could be securely transferred from the hospital database to the target database, it is necessary to use standards, in particular *Clinical Document Architecture Release 2* (CDA R2) by HL7 (*Health Level 7*), according to Italian rules. The system, exploited to collect the pseudonymous clinical and laboratory data reused some modules developed for the Liguria HIV Network [4], according to the service-oriented approach [5].

3. Results

The developed database structure allowed the storing of all the relevant information about: the patient (sex and age); the hospitalization (ward and the date of start of the patient's staying in each ward); information about each single parameter of the laboratory exams results and of the microbiological cultures. The designed and developed architecture is composed by: a *Windows Console Application* (client), it reads data from the hospital database and organizes them into a CDA R2; a main *Windows Communication Foundation* (WCF) service, that receives the CDA R2, it validates the structure of the document and then it extracts and stores data into the target database. This type of structure can be easily adapted and extended to be used in multicentric studies, it only requires having more than one client application (one for each hospital). At present in the first preliminary test a random selection of 50 patients has been considered and data from 57 clinical features were stored.

4. Conclusions

This manuscript's aim is to present the architecture developed to automatically collect into a research target database pseudonymous high-quality data of patients with episodes of candidemia or bacteremia admitted to the university hospital in Genoa.

References

- [1] Bassetti M. et al. Diagnosis and treatment of candidemia in the intensive care unit. *Seminars in respiratory and critical care medicine*. Vol. 40. No. 04. Thieme Medical Publishers, 2019.
- [2] Ripoli A. et al. Personalized machine learning approach to predict candidemia in medical wards. *Infection* 2020;48(5): 749-759. <https://doi.org/10.1007/s15010-020-01488-3>
- [3] AlZunitan M. et al. Predicting Community-Onset Candidemia in an Academic Medical Center Using Machine Learning. *Infection Control & Hospital Epidemiology*. 2020;41(S1): s355-s355.
- [4] Giannini B, et al. From Liguria HIV Web to Liguria Infectious Diseases Network: How a Digital Platform Improved Doctors' Work and Patients' Care. *AIDS Research and Human Retroviruses* 2018;34: 239-240.
- [5] Gazzarata R. et al. A SOA-based platform to support clinical data sharing. *Journal of healthcare engineering* 2017 (2017). <https://doi.org/10.1155/2017/2190679>