GRALENIA: Antimicrobial Resistance Management based on Natural Language and Artificial Intelligence

Cristóbal Bernardo-Castiñeira[1], Germán Bou[2], Manuel Campos[3], Bernardo Cánovas-Segura[3], Sergio Figueiras Gómez[1], Carlos Gómez-Rodríguez[4], Enrique Míguez Rey[2] and Jesús Vilares[4]

[1] *Bahía Software SLU, Rúa das Hedras 4, L-1, Ames, 15895, A Coruña, Spain*
[2] *Instituto de Investigación Biomédica de A Coruña (INIBIC), Hospital Teresa Herrera, 15006, A Coruña, Spain*
[3] *Medical Informatics and Artificial Intelligence Laboratory (MedAI Lab), University of Murcia, 30100, Murcia, Spain*
[4] *CITIC - Universidade da Coruña, Depto. de Ciencias de la Computación y Tecnologías de la Información, Campus de Elviña, 15071, A Coruña, Spain*

Abstract

The objective of GRALENIA project is to develop a multidisciplinary, comprehensive and interoperable platform incorporating artificial intelligence algorithms and natural language processing techniques to improve the management of antimicrobial resistance (AMR) and reduce the impact of antimicrobial- or antibiotic-resistant microorganisms (aka *superbugs*) in hospitals.
GRALENIA is supported through a Red.es grant for research and development projects in artificial intelligence and other digital technologies and their integration into value chains.

Keywords

Natural language processing, artificial intelligence, machine learning, deep learning, cloud services, healthcare, antimicrobial resistance, superbug, industrial research project

1. Introduction

Over the last two decades, antimicrobial resistance (AMR) has become a threat to public health systems worldwide. World Health Organization (WHO) considers this problem as one of its priorities [1]. An example of AMR is the resistance of certain pathogenic bacteria to antibiotics, which causes some treatments based on antibiotic prescription not to work, thus resulting in the appearance of serious clinical complications and a considerable increase in healthcare costs. In Europe alone, AMR causes 33,000 deaths per year and a loss of €1.5 billion in terms of additional treatment and social costs [2].

Furthermore, the uncontrolled transmission of such superbugs[1] between patients and healthcare workers poses a serious risk to the healthcare system. Appropriate use of antimicrobials is very complicated because of the complexity of infectious diseases and the spread of antibiotic resistance. Because of this, the Spanish National Plan for Antibiotic Resistance [3] has established the implementation of programs for optimizing the use of antibiotics in hospitals, the so-called Antimicrobial Stewardship Programs (ASP), that carried out by multidisciplinary teams of specialists, the ASP teams [4].

There is a wealth of information that can be exploited to manage this serious problem, but the analysis of all this information is a very complex process. In many cases, such analysis is done manually, which results in a work overload for professionals, thus increasing hospital costs or, failing that, worsening health care work.

Artificial intelligence (AI) algorithms can improve interventions and decision making by ASP and healthcare-associated infection (HCAI) surveillance teams. The information held by the different services of a hospital is susceptible to be exploitable by AI algorithms for the development of intelligent AMR and HCAI control strategies, and for improving evidence-based decision making. However, the exploitation of such data by AI is very complex due to the heterogeneity of the data and its frequent lack of standardization and structuring. This fact has hindered the development of AI models for dealing with AMR and superbug infections, which are generally still in a laboratory/research phase.

In this context, GRALENIA constitutes an ambitious industrial research project that seeks to develop solutions based on the use of AI and natural language processing (NLP) techniques to exploit this

SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics, June 19-20, 2024, A Coruña, Spain

cristobal.bernardo@bahiasoftware.es (C. Bernardo-Castiñeira); german.bou.arevalo@sergas.es (G. Bou); manuelcampos@um.es (M. Campos); bernardocs@um.es (B. Cánovas-Segura); sergio.figueiras@bahiasoftware.es (S. Figueiras); carlos.gomez@udc.es (C. Gómez-Rodríguez); enrique.miguez.rey@sergas.es (E. Míguez); jesus.vilares@udc.es (J. Vilares)

0000-0003-0411-3056 (C. Bernardo-Castiñeira); 0000-0001-8837-0062 (G. Bou); 0000-0002-5233-3769 (M. Campos); 0000-0002-0777-0441 (B. Cánovas-Segura); 0009-0003-1680-9269 (S. Figueiras); 0000-0003-0752-8812 (C. Gómez-Rodríguez); 0000-0003-2941-1834 (J. Vilares)

[1] *Superbugs* is a name given to harmful bacteria that have acquired resistance to one or more of the antibiotics used to treat them.

information and provide innovative solutions to the AMR challenge.

The rest of the paper is organized as follows. Section 2 describes the project, its objectives, the lines of work involved in its development, the participant teams and their contribution. Section 3 presents the general architecture of the GRALENIA platform and describes the modules that comprise it, with the exception of the IA module, to be described in Section 4. Finally, Section 5 summarizes and closes the paper.

2. The Project

As a whole, the GRALENIA project[2,3] aims to drive and accelerate digital transformation in healthcare, specifically in the area of AMR and superbug management. Hospital information systems contain a lot of information from various specialties directly or indirectly related to AMR occurrences, such as the possible inappropriate use of antibiotics (dosage information can be provided by the Pharmacy service) or diagnostic and characterization tests for superbugs (provided by the Microbiology service). Nevertheless, much of this information is not interrelated or standardized, making it difficult to exploit it automatically and to make decisions considering all the available information.

2.1. Objectives

The GRALENIA project pursues the development of a multidisciplinary, comprehensive and interoperable platform that incorporates AI algorithms and NLP techniques to improve the digital management of AMR, and to reduce the impact of superbugs in hospitals. Starting from this general objective, the project has the following specific objectives:

1. The integration and standardization of relevant clinical data held by different hospital services. This allows its ulterior automatic processing.
2. The development of a base infrastructure for automated annotation of unstructured clinical documents using NLP techniques.
3. The prediction of the risk of superbug emergence in the hospital, and the identification of groups of patients with high susceptibility to superbug infection.
4. To improve the management and visualization of integrated data for AMR surveillance.

GRALENIA platform is being developed to be compatible with the data infrastructure of the A Coruña University Hospital Complex (CHUAC). This allows AI models to be designed using real-world information, thus providing the hospital with a digital solution to improve decision making regarding an existing problem.

To this end, GRALENIA not only focuses on R&D tasks for developing predictive models, but also provide tools for structuring and standardizing both structured and unstructured information from various information silos. This makes it possible to integrate it as a whole, in a structured way, into a common data model on the platform. In turn, GRALENIA enables easy data visualization and exploitation by AI algorithms.

2.2. Work Lines

To achieve these ambitious objectives, considerable R&D effort is needed. Some of the main research lines involved in this project are:

1. To determinate which type of information, directly or indirectly related to AMR, is potentially exploitable.
2. To design an NLP-based system for the automatic extraction of AMR-related information from unstructured clinical documents.
3. Definition and development of advanced AI models for AMR-related risk prediction.
4. Validation of model results in a real-world environment.

In addition, it is also necessary to provide the system with the necessary mechanisms so that it can access information from the different hospital services relevant for AMR management.

2.3. Consortium

These work lines are developed by a consortium of four participants.

Bahía Software.[4] Coordinating team and industrial partner of the consortium. Founded by professionals with experience in the IT and Health & Healthcare sectors, the commercial activity of the company is mainly focused on e-Health, with important presence on e-Administration, industry and banking too. The company is recognized as an Innovative SME and as a national reference in medical image, tumor committees and surgical block management.

LYS Group (CITIC-UDC). The second organization involved in the project is the Language and Information Society Group (LYS)[5] of the ICT Research Center (CITIC)[6] of the University of A Coruña (UDC). LYS is an interdisciplinary research group formed by professors and researchers in AI and Linguistics, with extensive experience in the fields of NLP and Computational Linguistics. CITIC has an outstanding activity in the transfer of research results to society and industry. Based on its experience and expertise, the contribution of the LYS research team to the project will focus on the development of a base infrastructure for the automatic annotation of unstructured clinical information:

- The NLP-based clinical Information Extraction (IE) system for processing the

[2] Spanish acronym for *Gestión de Resistencias a Antimicrobianos basada en Lenguaje Natural e Inteligencia Artificial* (Antimicrobial Resistance Management based on Natural Language and Artificial Intelligence).

[3] gralenia.es (visited on February 2024).

[4] bahiasoftware.es (visited on February 2024).

[5] www.grupolys.org (visited on February 2024).

[6] citic.udc.es (visited on February 2024).

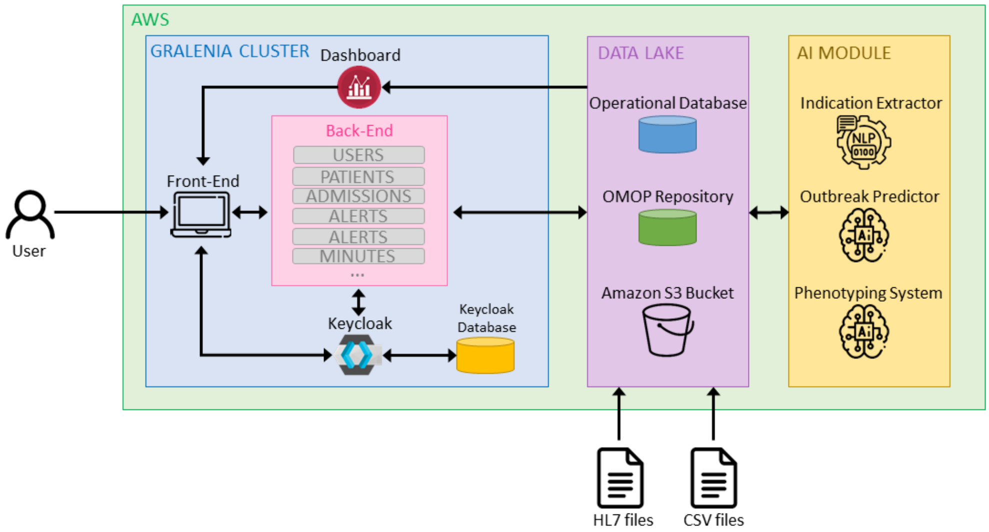

Figure 1: General architecture of the GRALENIA platform.

unstructured content of documents for their standardization and structuring.

- The guidelines and annotation tools necessary for the construction of the training corpus.
- Collection, analysis and adaptation of existing resources and tools necessary for the previous tasks.

AIKE Group. The Artificial Intelligence and Knowledge Engineering Group (AIKE) research group[7] of the University of Murcia has extensive experience in the development of IT solutions applied to the health field. Its main research lines focus on the development of intelligent systems for clinical application, data analysis and data mining in the health field, computational biology and medical informatics. Its contribution will be fundamental for the development of the predictive models of the platform:

- Machine learning (ML) and expert knowledge management applied to infections and treatments.
- Phenotyping of patients with nosocomial infections.
- AI visualization and explainability models applied to infection surveillance.

INIBIC-CHUAC.[8] The Institute of Biomedical Research of A Coruña (INIBIC) is a prestigious health research institute that integrates research groups of CHUAC, UDC and A Coruña Primary Care Service Area. INIBIC has leading researchers in the fields of clinical microbiology and AMR, where it is a national reference. Its main collaboration areas within the project are:

- Coordinating CHUAC services to identify and access those data repositories relevant to the project (Pharmacy, Virology, Epidemiology, etc.).
- Scientific-clinical consultancy. INIBIC oversees the results of the AI algorithms, the platform usability, the visualization tools, etc.
- Co-design of the end-user tools (healthcare professionals). This guarantees the usability of tools and their algorithms, as well as the highest possible added value.

In addition, INIBIC coordinates the management of access to data and infrastructures with the Information Systems Service of CHUAC.

3. General Architecture

Figure 1 shows the general architecture of the GRALENIA platform. All components run in a cloud environment. This allows us to have a cluster of virtual machines where all the applications and components shown in the blue block can be deployed. Specifically, the platform works on Amazon Web Services (AWS).[9] This solution has been chosen based on its scalability, flexibility, interoperability, cost efficiency, security and reliability.

3.1. Front-End

The user interface consists of a web interface, allowing a simpler and more accessible use. This reduces the learning curve of its handling and, therefore, makes it more pleasant to use. Additionally, the front-end also acts as a display window and information control panel for the Dashboard module (see Section 3.4). The front-end works as another service running within the GRALENIA cluster.

3.2. Data Ingestion

The data ingestion process of the GRALENIA platform includes several (sub)processes:

1. Extracting the data from the various information silos of the hospital.
2. Transforming the data to adapt it to the required format and structure.

[7] www.um.es/aike/ (visited on February 2024).
[8] www.inibic.es (visited on February 2024).
[9] aws.amazon.com (visited on February 2024).

Table 1
Description of the back-end microservices

MICROSERVICE	FUNCIONALITY
USERS	Its functionality involves returning the list of users and reviewers.
PATIENTS	It returns the patient information.
ADMISSIONS	It manages everything related to admissions: filtering, updates, revisions, etc.
ALERTS	It manages both the configuration of alerts and the list of detected alerts.
MINUTES	It enables access to the minutes of the ASP group meetings.
RECOMMENDATIONS	Focused on managing the recommendations made on a patient.by the ASP group.
REPORTS	In charge of accessing the lists of medications, microorganisms, services, etc.
S3REPOSITORY	Microservice that connects to the Amazon S3 and allows both accessing to the information stored there and uploading new documents.

Due to CHUAC internal regulations, privacy and information security issues, and project deadlines, a data integration strategy through CSV data downloads has been chosen. Given the heterogeneous nature of both the information systems of the hospital and the variables required, specific extraction, anonymization and transformation processes have been developed to fit a common data model. The resulting data are dumped into CSV files to be ingested into the platform storage system.

However, our platform also contemplates the possibility of integration with hospital information systems using HL7 standards.[10]

3.3. Storage System

GRALENIA platform implements a data lake. A data lake is a data repository used to store large amounts of both structured and unstructured data in its original format. Unlike traditional database systems, that require data to conform to a predefined schema, a data lake allows the ingestion of data from various raw sources and in various formats (text, images, sensor data, logs, etc.). The use of this data lake allows data from different information silos to be stored centrally and securely.

Despite being a centralized repository, a three-layer design has been used:

1. **Landing zone**. This is the entry point of the platform. In this first layer the data from the different information silos is stored in raw form, without any type of transformation.
2. **Staging zone**. An intermediate layer where validations and transformations are performed to adapt the data to the data model of the platform.
3. **Refined zone**. The last layer of the data lake. This is where the already structured and curated data is stored so that it can be consumed by the systems and tools that require it.

In the case of the first two layers, the landing and staging zones, an Amazon Simple Storage Service (S3) bucket[11] is used. This is a versatile cloud object storage service that stands out for its scalability, security and accessibility. On the other hand, the last layer, the refined zone, uses the Amazon RDS service[12] to host two relational databases:

1. **An operational database** to feed the various functionalities of the platform. The operational database is organized around hospital admissions, which uniquely identify each patient stay in a hospital.
2. **An OMOP-CDM database** to perform observational data analysis. The Observational Medical Outcomes Partnership (OMOP)[13] is an open community standard, designed to provide a standardized way to represent data structure (i.e. a Common Data Model or CDM) and content (terminologies, vocabularies, coding scheme, etc.), and to enable efficient analyses that can produce reliable evidence.

3.4. Information Control Panel

The GRALENIA platform has a catalog of components oriented to the visualization of the information in an intuitive way so that it can be exploited by any user, from generalist profiles to specialists in data analysis. This catalog facilitates the correct assimilation of the data and, when integrated with the AI module, it also allows showing the results obtained by the algorithms.

The dashboard designed in GRALENIA makes it possible to quickly and easily analyze various Key Performance Indicators (KPI) potentially related to the emergence of superbugs in a hospital. The dashboard integrates a set of filters that can be configured by the user to run customized analyses such as, for example, studying the behavior of a specific superbug in a given time period and/or area of the hospital.

3.5. Microservices (Back-End)

The back-end is the part of the platform that runs on the GRALENIA servers. We have decided to divide the functionalities of the system into several microservices. Thus, for each task of the platform, there is a single microservice with that single functionality. These microservices are described in Table 1.

[10] www.hl7.org/implement/standards/ (visited on February 2024).
[11] aws.amazon.com/s3/storage-classes/ (visited on February 2024).
[12] aws.amazon.com/rds/ (visited on February 2024).
[13] www.ohdsi.org/data-standardization/ (visited on February 2024).

3.6. Security, Management and Support Layer

In order to guarantee the security of a system that handles such sensitive data as this one, it is essential to use secure user and role control tools. For this purpose, the system uses Keycloak,[14] an open source and widely used identity and access management (IAM) tool. Keycloak is deployed as another service running on the GRALENIA cluster.

4. AI Module

The purpose of the AI module is to identify the triggers for the occurrence of multi-resistances. To this end, it integrates three systems: (1) an NLP-based IE system for clinical texts; (2) a system for predicting the risk of superbug outbreaks in different areas of the hospital; and (3) a system for identifying patient profiles with a high susceptibility to superbug infection.

4.1. Indication Extractor

The NLP subsystem of the IA module consists of an IE system to identify and extract relevant clinical information on free text sources such as electronic health records (EHR), nursing notes, etc. [5, 6]. In this case, we are interested in mentions that may indicate the possible existence or susceptibility of the patient to the infections under analysis. Those indicators are classified according to their typology (sign, symptom, etc.). The extracted information is then structured, homogenized and validated to complement that information obtained from other sources of the hospital. Finally, all this collected data feeds, in the form of features, the rest of subsystems of the AI module (outbreak prediction and phenotyping).

The general architecture of the indication extractor system is based on a pipeline, which gives it great flexibility. Two different pipeline configurations are provided:

1. A low-resource rule-based approach [7] that relies on a fuzzy pattern matching process and term expansion [8].
2. A state-of-the-art but resource-demanding approach based on transformers [9]. In the absence of a gold standard to train the system, the output of the previous (and simpler) pattern-based approach is used as a silver standard.

With the international market in mind, both approaches have been implemented for Spanish and English languages.

The IE system has been *dockerized*. for its final deployment to facilitate its integration with the rest of components of the platform.

4.2. Outbreak Predictor

The goal of this second component focuses on predicting the risk of outbreaks in different areas of the hospital at different times in the future [10, 11]. The scope of the project has limited the analysis to the most prevalent superbugs (*MRSA*, *Klebsiella ESBL* and *Klebsiella carbapenemase*), as well as to the most vulnerable areas of the hospital, those where an outbreak of these characteristics could have catastrophic consequences: the intensive care unit and the recovery room. For each superbug, two approaches have been, although a possible third approach (based on time series) is also currently under consideration.

The first approach consists on predicting the risk of occurrence of at least one superbug. Classification algorithms are used for this purpose. A battery of alternatives including logistic regression, decision trees, random forest, support vector machines (SVM), boosting and artificial neural network (ANN) based algorithms are being evaluated.

The second approach corresponds to the prediction of the number of cases of superbug contagion. In this case, due to the discrete nature of the target variable, Poisson regression models are used.

4.3. Phenotyping System

The objective of this third and final system is to identify patient profiles with high susceptibility to trigger a superbug outbreak.

For the phenotyping algorithms, an approach based on subgroup discovery has been selected [12]. These mixed descriptive-predictive models aim to obtain several sets of phenotypes to describe a concept, but also different in form so that they can show different perspectives of the database and, possibly, some of them have a clinical interpretation.

5. Conclusions

As a whole, GRALENIA constitutes an industrial research project with high added value for the healthcare professionals of the A Coruña University Hospital Complex (CHUAC), who are facing such complex problems as antimicrobial resistance (AMR) and superbug management. GRALENIA project lays the groundwork for extending the technologies developed within it not only to other hospital services, but also to other hospitals and health services.

It is also worth noting that many of the tools and techniques based on artificial intelligence developed within the project (e.g. predictive and spatial models, natural language processing tools, etc.) lay the foundations for their adaptation to other clinical areas, further boosting the digital transformation of the healthcare sector.

Acknowledgements

This work is partially funded by the GRALENIA project (reference 2021/C005/00150055) supported by the Spanish Ministry of Economic Affairs and Digital Transformation, the Spanish Secretariat of State for

[14] www.keycloak.org (visited on February 2024).

Digitization and Artificial Intelligence, Red.es and by the NextGenerationEU funding.

Research on CITIC is also financed by the Xunta de Galicia through the collaboration agreement between the Ministry of Culture, Education, Vocational Training and Universities and the Galician universities to reinforce the System's research centers University of Galicia (CIGUS)

We would also like to thank Dr. José María Gutiérrez Urbón (CHUAC Pharmacy service), Dr. Laura Gutiérrez Fernández (CHUAC Infectious Diseases unit), Daniel Llamas Gómez (current Chief Information Officer-CIO, A Coruña-Cee Health District, SERGAS) and Guillermo Vázquez González (former CIO) for their valuable contributions to the project.

References

[1] World Health Organization, Global Action Plan: Antimicrobial Resistance, WHO, Geneva, 2025. ISBN 9789241509763. URL: https://www.who.int/publications/i/item/9789241509763 (visited on February 2024).

[2] European Commission Public Health, EU Action on Antimicrobial Resistance, 2024. URL: https://ec.europa.eu/health/antimicrobial-resistance/eu-action-antimicrobial-resistance_en (visited on February 2024).

[3] Plan Nacional frente a la Resistencia a los Antibióticos (PRAN), 2024. URL: https://www.resistenciaantibioticos.es (visited on February 2024).

[4] J. Rodríguez-Baño, J.R. Paño-Pardo, L. Alvarez-Rocha *et al.*, "Programas de optimización de uso de antimicrobianos (PROA) en hospitales españoles: documento de consenso GEIH-SEIMC, SEFH y SEMPSPH", Enfermedades Infecciosas y Microbiología Clínica, 30.1 (2012): 22.e1–22.e23. DOI: 10.1016/j.eimc.2011.09.018.

[5] K. Kreimeyer, M. Foster, A. Pandey et al., "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review", Journal of Biomedical Informatics, 73:14–29, 2017. DOI: 10.1016/j.jbi.2017.07.012.

[6] A. Névéol, H. Dalianis, S. Velupillai et al., "Clinical Natural Language Processing in languages other than English: opportunities and challenges", Journal of Biomedical Semantics, 9:12, 2018. DOI: 10.1186/s13326-018-0179-8.

[7] G. Napolitano, C. Fox, R. Middleton et al., "Pattern-based information extraction from pathology reports for cancer registration", Cancer Causes Control, 21:1887–1894, 2010. DOI: 10.1007/s10552-010-9616-4.

[8] F. Prado-Valiño, R. Santos-Rios, C. Gómez-Rodríguez & J. Vilares, "Prototype of an Entity Recognition System for Antimicrobial Resistance Data Management", in: Proceedings of the 6th XoveTIC Conference (XoveTIC 2023), A Coruña, Spain, *forthcoming*.

[9] X. Yang, J. Bian, W. R. Hogan & Y. Wu, "Clinical concept extraction using transformers", Journal of the American Medical Informatics Association (JAMIA), 27(12):1935–1942, 2020. DOI: 10.1093/jamia/ocaa189.

[10] N. Peiffer-Smadja, T.M. Rawson, R. Ahmad et al., "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications", Clinical Microbiology and Infection, 26(5):584–595, 2020. DOI: 10.1016/j.cmi.2019.09.009.

[11] C.F. Luz, M. Vollmer, J. Decruyenaere et al., "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies", Clinical Microbiology and Infection, 26(10):1291–1299, 2020. DOI: 10.1016/j.cmi.2020.02.003.

[12] A. Lopez-Martinez-Carrasco, H.M. Proença, J.M. Juarez et al., "Novel Approach for Phenotyping Based on Diverse Top-K Subgroup Lists", in: Artificial Intelligence in Medicine, AIME 2023, vol. 13897 of Lecture Notes in Computer Science, Springer, Cham, 2023. DOI: 10.1007/978-3-031-34344-5_6.