The Detection of hospitalized patients at risk of testing positive to multi-drug resistant bacteria using MOCA-I, a rule-based "white-box" classification algorithm for medical data

Julie Jacques, Helene Martin-Huyghe, Justine Lemtiri-Florek, Julien Taillard, Laetitia Jourdan, Clarisse Dhaenens, David Delerue, Arnaud Hansske, Valérie Leclercq

The Detection of hospitalized patients at risk of testing positive to multi-drug resistant bacteria using MOCA-I, a rule-based "white-box" classification algorithm for medical data

JULIE JACQUES* (1,6), HELENE MARTIN-HUYGHE (3,5), JUSTINE LEMTIRI-FLOREK (3,4), JULIEN TAILLARD (2), LAETITIA JOURDAN (6), CLARISSE DHAENENS (6), DAVID DELERUE (2), ARNAUD HANSSKE (7), VALERIE LECLERCQ (3)

AUTHOR AFFILIATIONS

1. Lille Catholic University, Faculté de Gestion, Economie et Sciences
2. Alicante SARL, Seclin, France
3. Lille Catholic hospitals, Infection control department, Lille Catholic University, KASHMIR, Lille, France
4. CH Valenciennes, Intensive Care Department , F-59322 Valenciennes, France.
5. CH Arras, Pharmacy Department, Arras, France.
6. Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
7. Lille Catholic hospitals, IT System Department, Lille Catholic University, KASHMIR, Lille, France

*** corresponding author :**

Julie Jacques

60 boulevard Vauban

CS 40109

59016 Lille Cedex

France

julie.jacques@univ-catholille.fr

+33 (0)3 59 56 79 11

Keywords : Data Mining; Supervised Machine Learning; Information Systems; Infection Control; Drug Resistance, Multiple; Electronic Health Records

WordCount : 2970

Abstract

Background: Multi-drug resistant (MDR) bacteria are a major health concern. In this retrospective study, a rule-based classification algorithm, MOCA-I (Multi-Objective Classification Algorithm for Imbalanced data) is used to identify hospitalized patients at risk of testing positive for multidrug-resistant (MDR) bacteria, including Methicillin-resistant Staphylococcus aureus (MRSA), before or during their stay.

Methods: Applied to a data set of 48,945 hospital stays (including known cases of carriage) with up to 16,325 attributes per stay, MOCA-I generated alert rules for risk of carriage or infection. A risk score was then computed from each stay according to the triggered rules. *Recall* and *precision* curves were plotted.

Results: The classification can be focused on specifically detecting high risk of having a positive test, or identifying large numbers of at-risk patients by modulating the risk score cut-off level. For a risk score above 0.85, *recall* (sensitivity) is 62% with 69% *precision* (confidence) for MDR bacteria, *recall* is 58% with 88% *precision* for MRSA. In addition, MOCA-I identifies 38 and 21 cases of previously unknown MDR and MRSA respectively.

Conclusions: MOCA-I generates medically pertinent alert rules. This classification algorithm can be used to detect patients with high risk of testing positive to MDR bacteria (including MRSA). Classification can be modulated by appropriately setting the risk score cut-off level to favor specific detection of small numbers of patients at very high risk or identification of large numbers of patients at risk. MOCA-I can thus contribute to more adapted treatments and preventive measures from admission, depending on the clinical setting or management strategy.

1. Introduction

Multi-drug resistant (MDR) bacteria, including methicillin-resistant *Staphylococcus aureus* (MRSA), are a major health concern because MDR infections are very difficult to treat and can have significant medical impact, potentially leading to a fatal outcome if not treated appropriately. It is thus crucial to limit the diffusion of MDR bacteria. In hospital, this means identifying patients who are carriers or infected with MDR bacteria so precautionary measures can be instituted [1] from admission.

In hospital, the infection control (IC) team receives information about MDR status from many sources and is responsible for ensuring that colonized or infected patients receive adapted care. For instance, the hospital bacteriology laboratory may alert the IC team whenever a test sample is positive for MDR bacteria. This enables the team to identify large numbers of MDR patients but misses others because such tests are not always ordered. Patients who had a positive test before admission might also be missed. Care units complete initial screening for MDR and inform the IC team using alert systems that recall available data on current and prior history of MDR colonization or infection.

Different expert groups have proposed specific screening rules for hospital patients, e.g. the alert system described for MRSA by Evans et al. [2]. Data mining techniques can also be used to generate alert rules automatically. There has been a large volume of work dedicated to medical data mining [3], including identification of patients at risk of contagious infection [4] or risks factors for MDR or MRSA harbouring [5–8]. In our case, we used MOCA-I (Multi-Objective Classification Algorithm for Imbalanced data), a rule-based classification algorithm adapted to specificities of medical data [9]. First, MOCA-I is able to process binary or qualitative data (ordered or not) with more than 15,000 variables, when the previously presented approaches deal with a small number of variables ($n \leq 50$) [4–8]. This eliminates the need for data filtering and the risk of setting aside useful information. Secondly, an interesting feature of MOCA-I is its capacity to manage highly imbalanced data sets. According to the MDR 2014 report from RAISIN (French Nosocomial Infection Warning and Surveillance Investigation Network), the incidence density of MRSA is 0.27 per 1000 hospitalization days. This explains why many classical data mining algorithms fail [10]. Finally, MOCA-I is a white box classification algorithm, in opposition to state-of-the-art machine learning techniques such as Neural Networks and Random Forest. This is consistent with November 2018 CCNE (French National Consultative Ethics Committee)' recommendations about AI and health, suggesting to use AI approaches that the care team can criticize or challenge [11]. However new approaches started to emerge recently [12] that allows to explain the decision given by black-box models, that could be very useful in the future.

The approach proposed in the present work is also novel in that it assigns a risk score to each patient. This score can then be used to adapt the number of patient files to investigate as a function of available resources and the probability of detecting MDR carriage or infection.

The main purpose of the present work is to apply MOCA-I to a large-volume real-life data set in order to assess its capacity to identify patients at risk of testing positive to MDR. A secondary objective is to determine the medical pertinence of the alert rules and the ranking generated by the system. Two use cases are envisaged for the rules obtained. Retrospectively to identify coding errors or missed patients. Prospectively to create a questionnaire with relevant questions to ask incoming patients.

2. Materials and Methods

2.1. Data set

2.1.1. Data set elaboration

The data used for this retrospective study was obtained from the annual activity records of the 750-bed Lille Catholic Hospitals (St-Philibert and St-Vincent-de-Paul hospitals, Lille - France) in 2013, which represents 48,945 hospital stays, all units combined. During this period, the IC (Infection Control) team identified 340 stays concerning patients who were tested positive for MDR before or during their stay, including 128 for MRSA. Our preliminary work focused on MRSA, which has long been used in France as a marker for nosocomial diseases. Seeing the promising results, we have extended to MDR requiring additional precautionary measures as recommended by the French Hospital Hygiene Society [13]: *MRSA*, *enterobacteria BLSE*, *pseudomonas aeruginosa* and *acinetobacter baumannii*.

In all, up to 16,279 attributes were identified for each stay. However, the average number of recorded attributes per stay was 26.9. Possible attributes and their distribution are summarized in Table 1. They include ICD-10 codes [14] and risk factors widely described in the literature [5,6,15].

In addition to ICD 10 codes, the corresponding ICD hierarchy was recovered. For example, for a patient coded E10 – insulin-dependent diabetes, codes E10-E14 – diabetes mellitus and E00-E90 – endocrine, nutrition, and metabolic diseases were also attributed. We did the same for antibiotics with the corresponding hierarchy using ATC index[1]. We then added information relative to medical services: duration of stay in each unit and inter-unit referrals. Regarding antibiotics, we extracted those listed in the pharmacopeia of the corresponding hospital. When one of the antibiotics was mentioned for a patient using the drug's proprietary name or the international non-proprietary name, or a close approximation thereof (e.g. amoxicilin instead of amoxicillin), it was added to the patient's data set. If an antibiotic was found in a report or letter up to 6 months before the start of the stay, the number of days between this report and the first day of the stay was also added. Such reports were also searched to find elements indicating a patient's poor nutritional status including specific words (undernutrition, poor nutritional status) or blood test results (albumin, CRP, pre-albumin).

We also used the patient's address to identify nursing home residents. Specific terms using the name of specific nursing or retirement homes were taken to indicate the patient resided in a nursing home. The method described by Jaro-Winkler [16] to identify an address equivalent to a nursing home address was also applied.

[1] WHO Collaborating Centre for Drug Statistics Methodology: http://www.whocc.no/atcdd/

Table 1. Possible attributes for each hospital stay.

Data	**Number of possible attributes**
ICD-10 + ICD10 hierarchy	15,702
For each hospital unit: number of days spent during stay	81
Inter-unit referral combinations	315
Antibiotic mentioned in a pre-hospital report + antibiotic hierarchy	87
Antibiotic mentioned in a pre-hospital report during the 3 months before the stay + antibiotic hierarchy	85
Poor nutritional status	1
Residence in a nursing home	1
Calendar status : stay during holiday season, week-end, eve of public holiday,…	7
Total	16,279

2.1.2. Dividing the dataset

We followed the same protocol as Tandan et al. [17], we divided the data set into training data and test data to detect overfitting. Table 2 gives insights about the data set repartition. Two-thirds of the set, i.e. 32,192 stays included 235 stays involving known MDR patients of which 85 were MRSA. The other third of the data set, i.e. 16,753 stays included 105 stays concerning MDR patients (43 MRSA patients).

Table 2. Training and test data characteristics repartition

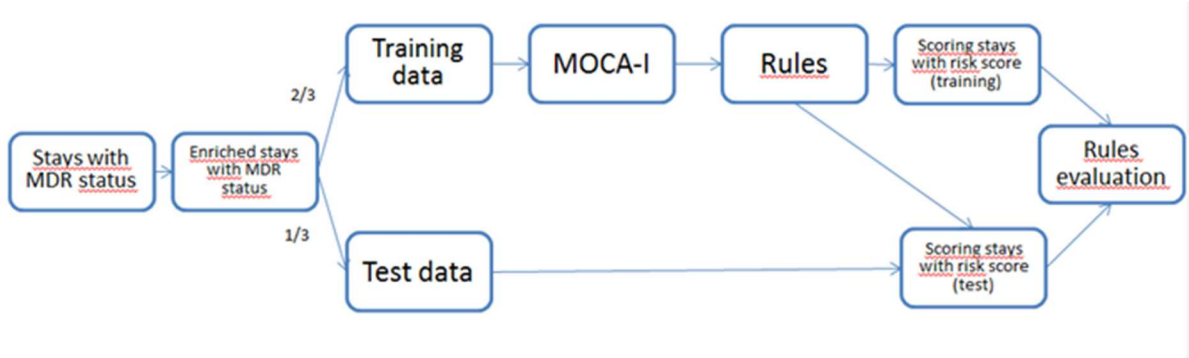
Characteristics	Training		Test	
	n(tra)	%	n(tst)	%
Total	32192		16753	
Age				
<=21	7051	21,90%	3740	22,32%
22-64	14093	43,78%	7356	43,91%
>=65	11048	34,32%	5657	33,77%
Carriage or infection status				
MDR	235	0,73%	105	0,63%
MRSA	85	0,26%	43	0,26%
Antibiotics				
3 months before admission	3444	10,70%	1740	10,39%
During stay	3965	12,32%	2040	12,18%
Other Informations				
At least 1 stay >= 48 hours in the last 3 months	4142	12,87%	2111	12,60%
Referral from nursing home	712	2,21%	339	2,02%
Poor nutritional status	1780	5,53%	942	5,62%
Sex				
Male	14632	45,45%	7581	45,25%
Female	17560	54,55%	9172	54,75%

2.2. Rules generation and exploitation

The process used to score risk by mining rules is illustrated in Figure 1. The process begins with the recording of 48,945 hospital stays, 340 of which were identified as MRSA, including 128 for MRSA. Each stay was then enriched with the information items described in section 2.1.1. As indicated in section 2.1.2, two-thirds of the data were used as training data to generate rules, one third was only used for evaluation and will be referred as test data. The classification algorithm – MOCA-I –produced 103 rules for predicting a risk of MDR positive test and 198 for MRSA. A score was attributed to each rule as a function of its performance level on training data (based on the F-measure corresponding to the mean recall and precision harmonic) [18]. Once evaluated, the rules generated were applied to both training data and test data. Stays that triggered at least one rule were considered at risk of a positive testing (407 stays for MDR and 200 for MRSA). At this step, the scoring system assigned a risk score to each stay as a function of the rules triggered.

MOCA-I algorithm is stochastic: different runs could give different rules. We already assessed the robustness of MOCA-I in a previous work [9]. The objective here was to check if this robustness is maintained on a real use case. Results are therefore presented for 10 runs of MOCA-I.

Figure 1. Scoring risk using existing stays and classification.



2.3. Risk estimation

The rules generated by MOCA-I may overlap, the mean being 6 rules triggered by stay for MDR carriage and 9 for MRSA carriage. When a single stay triggered several rules, a weighted score was generated as proposed by Zhang et al. [19]. Jaccard weighting that can be used to measure the degree of similarity between two rules as a function of the concerned populations, as suggested by Iglesia et al. [20], was applied. Thus stays triggering rules targeting almost the same population were assigned lower scores than stays triggering rules targeting different populations.

2.4. Model Evaluation

Several assessment criteria are used in the medical community [18] most using the number of true/false positives and true/false negatives. In this work, we will consider as positive a patient identified by MOCA-I at risk of testing positive to MDR or MRSA. The IC team retrospectively checked the files of patients detected by our approach. If they indicated that the patient was at risk of MDR or MRSA, the stay was counted as a true positive, if not the stay was counted as a false positive.

Sometimes the IC team is not informed that a stay had a positive test to MDR or MRSA bacteria. Thus there is no way to determine with certainty the number of true negatives: the number of patient files to be checked would have been too large. Next, we experimentally determine a cut-off beyond which the gain in precision and recall is small compared to the number of patients identified.

Consequently, we excluded assessment criteria based on true negatives, so *specificity* and *ROC curve* could not be retained. Conversely, the number of false positives could be determined with certainty above the cut-off value.

Therefore, we evaluated model performance using 2 metrics. *Recall* is defined as the proportion of known cases identified by the system: true positives / (true positives + false negatives). *Precision* is defined as the proportion of cases identified by the system: true positives / (true positives + false positives).

Since our approach assigns a risk score to certain stays, we can visualize the progression of *recall* and *precision* as a function of score. We can thus determine experimentally the cut-off level best suited to the detection of stays at risk.

3. Results

Figure 2 present respectively the *recall* and the *precision* as a function of the risk score for test sets, for both MDR and MRSA. *Recall* (respectively *Precision*) and the number of patients screened are plotted as a function the risk score. For a given score, patients above the cut-off level are considered at risk for MDR or MRSA.

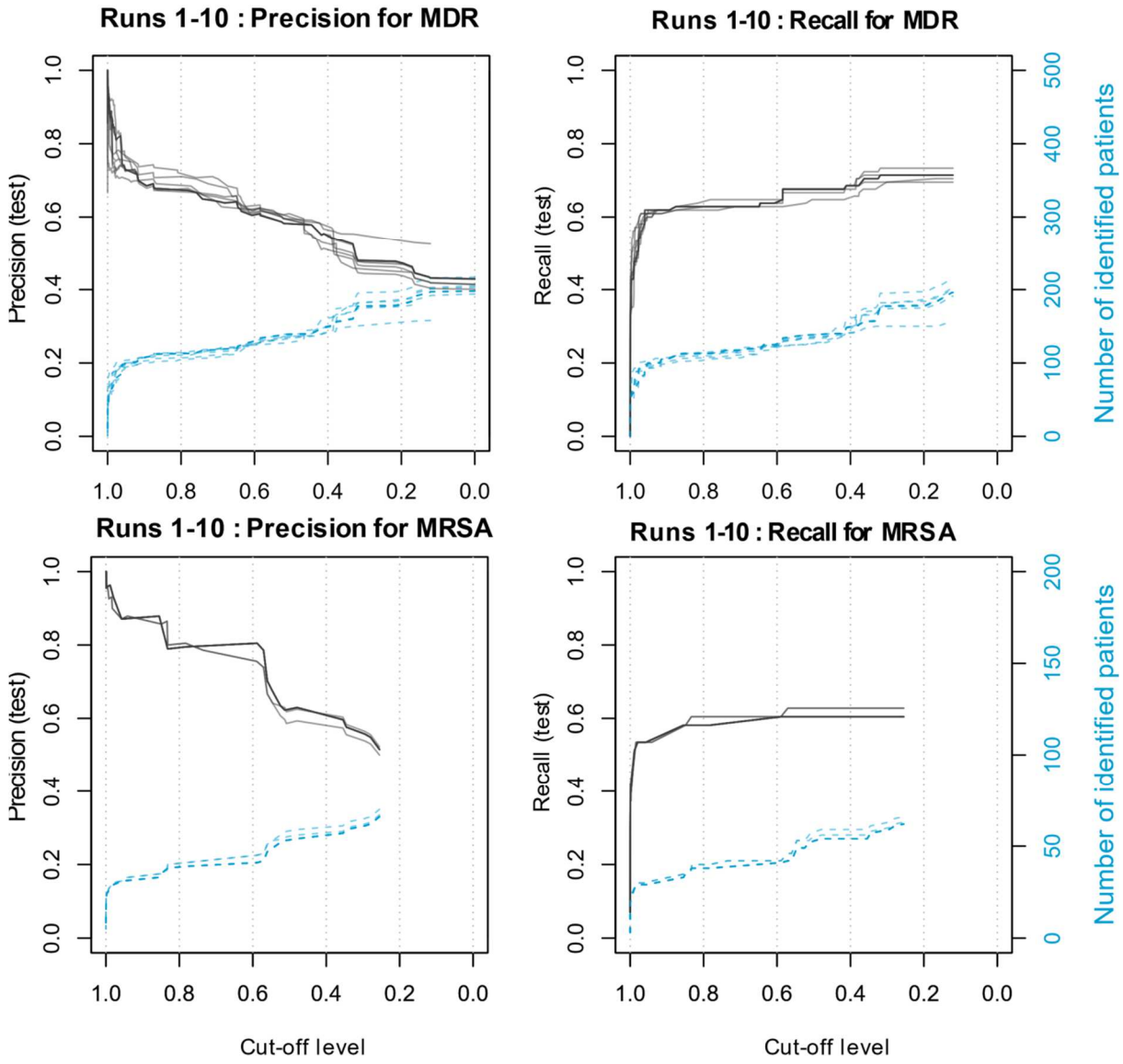
3.1. Recall

Using the test data set, the yield is at best 72% *recall* for MDR. The plot shows that starting with a score slightly below 0.85, the gain in *recall* achieved by lowering the cut-off level is at best 10% despite the addition of a supplementary 100 patients screened as at risk.

Consequently, the files of all patients whose score was ≥0.85, i.e. 380 patients (229 already known to the IC team and 151 new patients identified by the system), were reviewed. This cut-off corresponds to 62% *recall* with the test data.

Recall is better for MDR than MRSA. At best, MRSA *recall* reaches 61%, with a 0.5 cut-off score. As a compromise, the files of patients with a predictive score ≥0.85, which corresponded to 58% MRSA *recall*, were reviewed.

Figure 2. Precision (left panel) and Recall (right panel) on test (black lines for 10 runs) and number of identified patients (blue dashed lines for 10 runs) for hospital stays involving MDR (upper panel) or MRSA (lower panel).



3.2. Precision

For MDR *precision* is very high when the cut-off is ≥0.95, ranging from 80% to 100% with the test data. The lowest *precision* is 47%.

For MRSA, *precision* declines less rapidly, by plateaus, with a first plateau at ≥90% for a 0.9 cut-off score. A second plateau appears at 80% (cut-off score 0.7).

3.3. Mean recall and mean precision by score cut-off

The mean values for *recall* and *precision* are presented in Table 3 as a function of the cut-off scores determined from the plots in Figure 2. The number of new patients screened as at risk and the number of false positives are presented in the last two columns of Table 3.

Table 3. Mean and standard deviation recall and precision for different cut-off scores, with the number of patients identified as at risk for MDR or MRSA.

Risk	Score cut-off	Recall		Precision		Number of at risk patients identified			
		training	test	training	test	Total	Known	New	False

									positives
MDR	0.90	69 ±0.89%	62 ±0.50%	75 ±1.54%	71 ±1.66%	352 ±9.8	226 ±2.6	35 ±1.7	92 ±7.1
	0.85	69 ±0.67%	62 ±0.47%	71 ±2.03%	69 ±1.71%	380 ±11.1	228 ±2.1	38 ±0.8	113 ±9
	0.5	72 ±0.20%	67 ±1.07%	60 ±1.08%	59 ±1.05%	463 ±9.4	239 ±1.2	39 ±0.3	186 ±8.8
	0.2	75 ±0.48%	71 ±0.72%	49 ±0.49%	47 ±1.37%	598 ±9.2	251 ±1.2	39 ±0	308 ±8.5
MRSA	0.90	70 ±1.06%	58 ±0%	89 ±1,26%	88 ±0%	116 ±3	83 ±1.2	20 ±0.8	13 ±1.3
	0.85	72 ±0.24%	58 ±0%	87 ±2.31%	88 ±0%	123 ±4.7	86 ±1	21 ±1.2	16 ±2.7
	0.5	84 ±0%	61 ±0.99%	69 ±2.27%	64 ±2.44%	184 ±6.7	97 ±0.4	27 ±0	60 ±6.3
	0.25	84 ±0.48%	61 ±0.95%	59 ±2.67%	54 ±1.92%	218 ±9.8	97 ±0.8	28 ±0	93 ±9.2

3.4. Alert rules

Table 4 presents a selection of the 103 rules generated by the system for MDR and the 198 rules generated for MRSA. The rules having a good *precision* or *recall* are presented, but also those containing interesting risk factors.

Table 4. Selected rules generated by MOCA-I to detect risk of MDR and MRSA.

Prediction	**Rule (as combination of factors)**	**Precision**	**Recall**	**Precision**	**Recall**
		training		**test**	
MDR	**U88** - Agent resistant to multiple antibiotics **T80-T88** - Complications of surgical and medical care, not elsewhere classified **Age** > 75	78.8%	8.6%	33,3%	0.95%
	U88 - Agent resistant to multiple antibiotics **B95-B98** - Bacterial, viral and other infectious agents	73.5%	38.9%	70,5%	29.5%
	B95-B98 - Bacterial, viral and other infectious agents Spent between 31 and 90 days in intensive care	64.7%	4.7%	100%	4.8%
	L891 - Stage II decubitus ulcer **Z290** - isolation	38.9%	8.9%	25%	3.8%
	Found antibiotic of group ‘Carbapenems‘ (ATC **J01DH**) in one medical report during the stay	37.5%	10.2%	50%	11.4%
	Found antibiotic of group ‘imipenem and enzyme inhibitor ‘ (ATC **J01DH51**) in one medical report during the stay	34.1%	8%	47.4%	8.6%
MRSA	**E43** - Unspecified severe protein-energy malnutrition **J152** - Pneumonia due to staphylococcus **U00-U99** - Codes for special purposes	91.6%	12.9%	100%	4.7%

	J152 - Pneumonia due to staphylococcus **U00-U99** - Codes for special purposes	76.9%	15.6%	100%	11.6%
	U801 - Methicillin resistant agent **Z290** – Isolation **B95-B98** - Bacterial, viral and other infectious agents	73.5%	47.6%	74.1%	46.5%
	U801 - Methicillin resistant agent **Z290** – Isolation	71.7%	51.6%	74.1%	46.5%
	U801 - Methicillin resistant agent **E00-E90** - Endocrine, nutritional and metabolic diseases	64.4%	43.7%	77.3%	39.5%

4. Discussion

These results demonstrate that with a cut-off of 0.85, MOCA-I recalls in average 62% of patients who have a positive test for MDR and 58% for MRSA. Moreover, it identifies 38 other patients suspected of MDR or MRSA (21) not known to the IC team. With a high cut-off (0.9) the screened patients are relevant since 71% (mean precision on test in Table 3) of the identified MDR patients had a positive test as 88% of the identified MRSA patients. Setting the cut-off score at 0.85 yields a new set of 380 suspected MDR patients and a new set of 124 suspected MRSA patients. These patients correspond to 0.78% and 0.25% of all hospital stays during the year under study. Lowering the cut-off gives 597 patients at risk for MDR (1.2% of stays) and 184 at risk for MRSA (0.38% of stays). The risk score cut-off can be modulated to adapt the number of suspected patients to the proposed care or prevention strategy being proposed. Limiting the number of suspected patients enables the implementation of costly measures, e.g. nasal and rectal screening or complementary contact precautions from admission, while awaiting laboratory results.

The system does not appear to be particularly susceptible to overfitting, despite a very slight increase in MRSA *recall* for lower cut-off levels (<0.80). Regarding the rules generated, it can be seen that certain rules yield higher results with the test data than with the training data, which would suggest only modest overfitting. For robustness, some runs yield 5%-10% variation in precision depending on the cut-off level. An interesting perspective would be to study whether a combination of the results of several runs (5 or 10) would be more efficient.

It seems to be easier to identify MRSA than MDR. Our hypothesis is that bacteria may be resistant to a common set of antibiotics yet be quite different, exhibiting different patterns of drug resistance. Moreover, the pathogenic effect of MDR bacteria can be quite variable, inducing, unlike MRSA infection or colonization, a wide range of clinical effects. This produces a complex data set that complexifies rule generation.

4.1. Identified risk factors

Risk factors can be extracted from the rules. Most of the risk factors thus identified are directly related to MDR or MRSA infections, e.g. bacteria codes, resistance codes, codes indicating isolation. Other identified risk factors are pertinent for IC practices. For instance, if a patient file indicates that carbapenems were used, the IC team makes a presumption of MDR because these drugs are used for the treatment of MDR infections, notably due to

extended spectrum beta-lactamase producing *Enterobactericeace*. Hospitalization in an intensive care unit for 31-90 days or the presence of a pressure sore are also recognized risk factors because they are observed in patients who are bedridden for long periods of time, a condition favouring bacterial infection and colonization [15,21]. The rules also identify metabolic diseases as a risk factor. These diseases group together factors such as diabetes mellitus for MRSA [8,15] or other factors such as poor nutritional status or obesity, for which we have not found a literature reference.

Certain risk factors are related to coding practices. Several rules do not specifically contain the code for MRSA (ICD-10 B956) but have neighboring codes (ICD-10 B95 to B98) that correspond to other bacteria. MOCA-I focused more on the neighboring codes because they produced more pertinent results. This would suggest that the coding process is not necessarily straightforward for practitioners who may often choose a code for bacteria other than MRSA.

It is also noteworthy that certain candidate risk factors are not present in the rules, e.g. use of antibiotics before the hospital stay, referral from a nursing home, or inter-unit referrals. For use of antibiotics before the stay, this might involve the data source (prior medical reports). A prior medical stay is found for only 10% of the patients in the full data set. After reviewing a sample of the data we found that the discharge report does not always mention all treatments delivered during the stay, especially if terminated before discharge. This attribute is thus probably not very useful for data mining. Regarding the other attributes selected at the beginning of the study such as referral from a nursing home, it could be caused by insufficient data for use in rules, or insufficient impact on MDR or MRSA for selection by the data mining algorithm.

This work opens up interesting new perspectives. The first would be to benefit from the capacity of MOCA-I data mining to deal with large numbers of columns to be coupled with a data extraction technique applied to medical reports. With MOCA-I, it would not be necessary to target only a few interesting medical concepts to be extracted from medical reports as proposed in similar work [4] since all the concepts could be screened. An extraction tool could be applied to reports written in French : the work of Tvardik et al. gives interesting results for the detection of hospital acquired infections in medical texts [22]. Another perspective would be to use an algorithm such as MOSC (Multi-Objective Sequence Classifier) [23]. Finally, some data mining work devoted to the detection of drug adverse effects has shown that the models obtained are site-dependent [24]. It would thus be interesting to determine whether the rules generated here would be valid in another institution.

5. Conclusions

MOCA-I is a classification algorithm capable of detecting hospital patients at risk of testing positive for MDR or MRSA bacteria. Applied to the annual data issuing from 48,945 hospital stays in our institution, MOCA-I identified a majority of known carriers or infected plus 39 supplementary patients for MDR 27 for MRSA. The screening rules generated by the system are medically pertinent. MOCA-I can be used at hospital admission to screen for additional patients at high risk of having a positive test for MDR, allowing the implementation of adapted treatments and preventive measures.

Authors' contributions

CD, LJ, JJ, JT: evaluation protocol design

JJ, JT: data collection, processing

CD, LJ, JJ, JT: system elaboration and calibration

HMH, VL: review of patient files, medical expertise, qualitative analysis of generated rules

JJ: first draft, system implementation and evaluation, results collection and analysis

All authors: rereading, approval of final manuscript

Acknowledgements

Sabrina Meniaoui and Amélie Vasseur for reviewing the patient files
Justine Lemtiri-Florek for her medical expertise for the first versions of the system
Laurene Norberciak for advice on the evaluation protocol

Statement on conflicts of interest

Julien Taillard and David Delerue are employed by Alicante, which is the company that publishes MOCA-I. Julie Jacques is a former employee of Alicante.

Summary table

What was already know on the topic	What this study added to our knowledge
• MDR and MRSA bacteria carriage or infection need adapted care • MOCA-I is efficient for machine learning on medical data (imbalance, uncertainty, volumetry : high number of columns)	• MOCA-I is also efficient in detecting hospitalized patients at risk to have a positive test to MDR bacteria or MRSA • Rules can be combined to obtain a score and rank patients according to their risk • The care team can use the score to adapt the number of patients to review in priority

References

[1] J.D. Siegel, E. Rhinehart, M. Jackson, L. Chiarello, 2007 Guideline for Isolation Precautions: Preventing Transmission of Infectious Agents in Health Care Settings, American Journal of Infection Control. 35 (2007) S65–S164. https://doi.org/10.1016/j.ajic.2007.10.007.

[2] R.S. Evans, C.J. Wallace, J.F. Lloyd, C.W. Taylor, R.H. Abouzelof, S. Sumner, K.V. Johnson, A. Wuthrich, S. Harbarth, M.H. Samore, CDC Prevention Epicenter Program, Rapid identification of hospitalized

patients at high risk for MRSA carriage, J Am Med Inform Assoc. 15 (2008) 506–512. https://doi.org/10.1197/jamia.M2721.
[3] D.D.P. Shukla, S.B. Patel, A.K. Sen, A Literature Review in Health Informatics Using Data Mining Techniques, (2014) 7.
[4] S. Gerbier-Colomban, Q. Gicquel, A.-L. Millet, C. Riou, J. Grando, S. Darmoni, V. Potinet-Pagliaroli, M.-H. Metzger, Evaluation of syndromic algorithms for detecting patients with potentially transmissible infectious diseases based on computerised emergency-department data, BMC Med Inform Decis Mak. 13 (2013) 101. https://doi.org/10.1186/1472-6947-13-101.
[5] S. Harbarth, H. Sax, C. Fankhauser-Rodriguez, J. Schrenzel, A. Agostinho, D. Pittet, Evaluating the probability of previously unknown carriage of MRSA at hospital admission, Am. J. Med. 119 (2006) 275.e15–23. https://doi.org/10.1016/j.amjmed.2005.04.042.
[6] E. Tacconelli, New strategies to identify patients harbouring antibiotic-resistant bacteria at hospital admission, Clin. Microbiol. Infect. 12 (2006) 102–109. https://doi.org/10.1111/j.1469-0691.2005.01326.x.
[7] C. Couderc, S. Jolivet, A.C.M. Thiébaut, C. Ligier, L. Remy, A.-S. Alvarez, C. Lawrence, J. Salomon, J.-L. Herrmann, D. Guillemot, Antibiotic Use and Staphylococcus aureus Resistant to Antibiotics (ASAR) Study Group, Fluoroquinolone use is a risk factor for methicillin-resistant Staphylococcus aureus acquisition in long-term care facilities: a nested case-case-control study, Clin. Infect. Dis. 59 (2014) 206–215. https://doi.org/10.1093/cid/ciu236.
[8] A. Reighard, D. Diekema, L. Wibbenmeyer, M. Ward, L. Herwaldt, Staphylococcus aureus nasal colonization and colonization or infection at other body sites in patients on a burn trauma unit, Infect Control Hosp Epidemiol. 30 (2009) 721–726. https://doi.org/10.1086/598681.
[9] J. Jacques, J. Taillard, D. Delerue, C. Dhaenens, L. Jourdan, Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced

data sets, Applied Soft Computing. 34 (2015) 705–720. https://doi.org/10.1016/j.asoc.2015.06.002.
[10] H. He, E.A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering. 21 (2009) 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
[11] Comité Consultatif National d'Ethique, AVIS 129 - Contribution du Comité consultatif national d'éthique à la révision de la loi de bioéthique 2018-2019, (2018). https://www.ccne-ethique.fr/sites/default/files/avis_129_vf.pdf.
[12] M.T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016: pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
[13] Société française d'hygiène hospitalière, Recommandations nationales - Prévention de la transmission croisée : précautions complémentaires contact, Hygiènes. XVII (2009) 60.
[14] ICD-10: International statistical classification of diseases and related health problems., World Health Organization, Geneva, 2011.
[15] H. Sax, S. Harbarth, G. Gavazzi, N. Henry, J. Schrenzel, P. Rohner, J.P. Michel, D. Pittet, Prevalence and prediction of previously unknown MRSA carriage on admission to a geriatric hospital, Age Ageing. 34 (2005) 456–462. https://doi.org/10.1093/ageing/afi135.
[16] W.E. Winkler, The State of Record Linkage and Current Research Problems, Statistical Research Division, U.S. Census Bureau, 1999.
[17] M. Tandan, M. Timilsina, M. Cormican, A. Vellinga, Role of patient descriptors in predicting antimicrobial resistance in urinary tract infections using a decision tree approach: A retrospective cohort study, International Journal of Medical Informatics. 127 (2019) 127–133. https://doi.org/10.1016/j.ijmedinf.2019.04.020.
[18] M. Ohsaki, H. Abe, S. Tsumoto, H. Yokoi, T. Yamaguchi, Evaluation of rule interestingness measures in medical knowledge discovery in databases, Artificial Intelligence in Medicine. 41 (2007) 177–196. https://doi.org/10.1016/j.artmed.2007.07.005.

[19] J. Zhang, J.W. Bala, A. Hadjarian, B. Han, Ranking Cases with Classification Rules, in: J. Fürnkranz, E. Hüllermeier (Eds.), Preference Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011: pp. 155–177. https://doi.org/10.1007/978-3-642-14125-6_8.
[20] B. de la Iglesia, A. Reynolds, V.J. Rayward-Smith, Developments on a Multi-objective Metaheuristic (MOMH) Algorithm for Finding Interesting Sets of Classification Rules, in: C.A. Coello Coello, A. Hernández Aguirre, E. Zitzler (Eds.), Evolutionary Multi-Criterion Optimization, Springer Berlin Heidelberg, 2005: pp. 826–840.
[21] A. Korytny, K. Riesenberg, L. Saidel-Odes, F. Schlaeffer, A. Borer, Bloodstream infections caused by multi-drug resistant Proteus mirabilis: Epidemiology, risk factors and impact of multi-drug resistance, Infect Dis (Lond). 48 (2016) 428–431. https://doi.org/10.3109/23744235.2015.1129551.
[22] N. Tvardik, I. Kergourlay, A. Bittar, F. Segond, S. Darmoni, M.-H. Metzger, Accuracy of using natural language processing methods for identifying healthcare-associated infections, International Journal of Medical Informatics. 117 (2018) 96–102. https://doi.org/10.1016/j.ijmedinf.2018.06.002.
[23] M. Vandromme, J. Jacques, J. Taillard, A. Hansske, L. Jourdan, C. Dhaenens, Extraction and optimization of classification rules for temporal sequences: Application to hospital data, Knowledge-Based Systems. 122 (2017) 148–158. https://doi.org/10.1016/j.knosys.2017.02.001.
[24] E. Chazard, G. Ficheur, S. Bernonville, M. Luyckx, R. Beuscart, Data mining to generate adverse drug events detection rules, IEEE Trans Inf Technol Biomed. 15 (2011) 823–830. https://doi.org/10.1109/TITB.2011.2165727.

Funding

This work was supported by internal funds from Lille Catholic hospitals, Lille Catholic University (GHICL - Groupement des Hôpitaux de l'Institut Catholique de Lille).
Part of this work was conducted within the framework of the CLINMINE ANR-13-TECS-0009 French project.

Patient Consent

This is a data-reuse study. During their stay, patients are informed that they give their implied consent for the re-use of their data for research and educational purposes. Patients may refuse this implied consent, in this case the concerned patients' files were removed from this study.