Prediction of the Occurrence of Leprosy Reactions Based on Bayesian Networks

Rafael Saraiva de Andrade Rodrigues
Pontificia Universidade Catolica do Parana
Eduardo Ferreira José Heise
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Luis Felipe Hartmann
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Guilherme Eduardo Rocha
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Marcia Olandoski
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Mariane Martins de Araújo Stefani
Universidade Federal de Goiás Instituto de Patologia Tropical e Saúde Pública: Universidade Federal de Goias Instituto de Patologia Tropical e Saude Publica
Ana Carla Pereira Latini
Lauro de Souza Lima Institute: Instituto Lauro de Souza Lima
Cleverson Teixeira Soares
Instituto Lauro de Souza Lima
Andrea Belone
Lauro de Souza Lima Institute: Instituto Lauro de Souza Lima
Patrícia Sammarco Rosa
Lauro de Souza Lima Institute: Instituto Lauro de Souza Lima
Maria Araci de Andrade Pontes
Dona Libânia Dermatology Consultants
Heitor de Sá Gonçalves
Dona Libânia Dermatology Centre
Rossilene Cruz
Tropical Dermatology and Venerology Alfredo da Matta Foundation Amazonas
Maria Lúcia Fernandes Penna
Epidemiology and Biostatistics Department, Federal University Fluminense
Deborah Ribeiro Carvalho
Pontifical Catholic University of Parana: Pontificia Universidade Catolica do Parana
Vinicius Medeiros Fava
McGill University
Samira Bührer-Sékula
Federal University of Goiás Institute of Tropical Pathology and Public Health: Universidade Federal de Goias Instituto de Patologia Tropical e Saude Publica
Gerson Oliveira Penna
University of Brasilia: Universidade de Brasilia
Claudia Maria Cabral Moro
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Julio Cesar Nievola
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana
Marcelo Távora Mira (m.mira@pucpr.br)
Pontifícia Universidade Católica do Paraná: Pontificia Universidade Catolica do Parana https://orcid.org/0000-0002-8087-9400

Research Article

Keywords: Leprosy, Leprosy Reactions, Bayesian Networks, Artificial Intelligence

Posted Date: January 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1135008/v1

Abstract

Background: Leprosy reactions (LR) are severe episodes of intense activation of the host inflammatory response, of uncertain etiology, today the leading cause of permanent nerve damage in leprosy patients. Several genetic and non-genetic risk factors for LR have been described; however, there are limited attempts to combine this information in order to estimate the risk of a leprosy patient to develop LR. Here we present an artificial intelligence (AI)-based system able to estimate risk of LR using clinical, demographic and genetic data.

Methods: The study includes four datasets from different regions of Brazil, totalizing 1,450 leprosy patients followed prospectively for at least two years to assess the occurrence of LR. Data mining using WEKA software was performed following a two-step protocol to select the variables included in the AI system, based on Bayesian Networks and developed using the NETICA software.

Results: Analysis of the complete database resulted in a system able to estimate LR-risk with 82.7% accuracy, 79.3% sensitivity, and 86.2% specificity. When using only databases for which host genetic information associated with LR was included, the performance increased to up to 87.7% accuracy, 85.7% sensitivity, and 89.4% specificity.

Conclusion: We produced an easy-to-use, online, free-access system that allows the identification of leprosy patients at high risk of developing LR. Risk assessment of LR for individual patients may detect candidates close monitoring, with potential positive impact upon the prevention of permanent disabilities, the quality of life of the patients, as well as upon leprosy control programs.

Background

Leprosy is a chronic, disabling infectious disease caused by *Mycobacterium leprae* (*M. leprae*) [1-3] that affects 202,000 new individuals worldwide every year, with most cases concentrated in India and Brazil [4]. In the classical Ridley & Jopling (R&J) classification system, tuberculoid (TT) and lepromatous (LL) leprosy occupy opposite ends of a continuous disease spectrum that includes three borderline forms (BT, BB and BL) [5]. The TT+BT and BB+BL+LL cases roughly correspond to paucibacillary (PB) and multibacillary (MB) leprosy, according to the treatment-oriented World Health Organization (WHO) classification scheme, respectively [4, 6, 7]. Today, it is widely accepted that exposure to *M. leprae* is necessary but not sufficient to develop leprosy. Different sets of host gene variants mediate susceptibility to leprosy in three different stages [8]: (i) controlling infection *per se*, that is, the disease regardless of its clinical presentation; (ii) defining the clinical form of disease after the infection is established; and (iii) outlining the risk of developing leprosy reactions (LR) [9, 10].

Leprosy reactions are characterized by an intense and sudden (re)activation of the host inflammatory response that may be diagnosed concomitantly with leprosy, during or even after treatment [2, 11-13]. Upon diagnosis, LR requires immediate medical attention to prevent irreversible nerve damage, motor disability and permanent anatomical deformities. In 2019, 5.35% of newly detected leprosy cases worldwide presented grade-2 disabilities at diagnosis [4], often due to the occurrence of LR. Cohort studies estimate that, during the course of leprosy, 16 to 56% of the patients will develop irreversible nerve damage, again, mainly due to the occurrence of reactional episodes [14-17]. Over the past years, advances in genetic research improved our understanding of the molecular basis of leprosy pathogenesis and several host genetic variations have been implicated in the control of LR episodes [18-20].

There are two major types of LR of distinct clinical presentation: type-1 (T1R) and type-2 reaction (T2R). T1R affects 10-30% of leprosy patients and occur mostly within, but not limited to, the first two years after leprosy diagnosis [21, 22]. Known risk factors for T1R are: (i) borderline clinical groups BT-BL [23]; (ii) age of leprosy onset, with older individuals being at higher risk [24, 25]; (iii) positive bacillary index [26]; (iv) an increased number of lesions at leprosy diagnosis [27, 28]; (v) detection of *M. leprae* DNA in biopsies of lesions [25]; and (vi) genetic/genomic studies have identified association between T1R and genes *TLR1* [29], *TLR2* [30], *TLR3* [31], *TLR7* [31], *TLR10* [31], *NRAMP1/SCLC11A1* [32], *VRD* [33], *NOD2* [34], *TNFSF15*/*TNFSF8* [3, 35], lncRNA *ENSG00000235140* [36], *LRRK2* [20], and *PRKN* [20].

Leprosy T2R mainly affects patients classified within the BB-LL range [14, 37]. Patients presenting bacterial index higher than 4+ in skin smears are at increased risk for T2R [38, 39]. There is a wide variation in the prevalence of T2R in different geographic and endemic settings. In Brazil, approximately 37% of BL and LL cases develop T2R, while in India, Nepal and Thailand, the proportion is between 19-26% [40]. A prospective study involving BL and LL patients from India followed for 11 years showed that less than 10% of the individuals who developed T2R had a single episode, whereas 62% had chronic T2R [22]. In Ethiopia, 63% of leprosy cases had more than one T2Repisode, while 37% had a single event [41]. Host genetics also seems to play a major role controlling the occurrence of T2R, and genes *C4B* [42], *TLR1* [43], *NRAMP1*/*SCLC11A1* [32], *NOD2* [34, 35] and *IL6* [2, 35] have been implicated as important molecular players.

Currently, one of the challenges of translational medicine is to systematize the analysis of a large amount of patient data to predict a specific outcome. In addition, scientific results from basic research are often difficult to translate to the daily medical practice. Artificial Intelligence (AI) methods seek to systematically address a set of information to provide a base for decision-making. Of particular interest in health care, Bayesian Networks (BN) are among the most successful techniques in processing and unraveling the relationship between a large number of variables, with risk estimation being the final outcome [44].

There are several BN-based systems created using medical data, developed for different purposes and applied to several diseases [45-47], including leprosy [44, 48-55]. However, few initiatives aim to systematize a large amount of existing information of distinct nature with the purpose of estimating the risk of occurrence of a particular event. In the context of leprosy, the creation of a simple to use and flexible platform to predict the risk of LR based on patient data may be helpful to minimize the consequences of such aggressive events. Moreover, such tool would have potential to contribute to improve

leprosy control initiatives and public health systems. Here we present an AI system designed to predict the risk of a leprosy patient to develop LR using a complete or partial dataset of clinical, demographic, and genetic patient data.

Materials And Methods

Population samples

This study was performed using four pre-existing data sets obtained from previous research initiatives of different/independent designs and contexts. The first database included in the study consisted of 409 leprosy patients diagnosed at the Reference Center for Diagnosis and Therapy located in Goiania, central-western Brazil, between February 2006 and March 2008, originally used for the genetic study that identified association between T2R and variants of the *IL6* gene. A complete description of the Goiania population has been published elsewhere [2]. Later, the Goiania population was used for an expanded investigation involving a larger number of candidate genes that detected association between T1R and variants of the gene *TNFSF8* [3]. Finally, association between T1R and lncRNA *ENSG00000235140* [36] and *LRRK2* (unpublished data) was also detected in the Goiania sample. Two additional databases were composed of 533 patients recruited at the Dermatological Center Dona Libânia, Fortaleza, northeast Brazil, and 137 patients diagnosed with leprosy at the Fundação Alfredo da Matta, Manaus, north Brazil. Enrolment of these two population samples was performed under a single protocol of a clinical study described previously [56] and conducted by the Tropical Medicine Center of the University of Brasília between March 2007 and February 2012. Finally, a fourth database consisted of 371 patients diagnosed with leprosy at the Instituto Lauro de Souza Lima, Bauru, southeast Brazil, between March 2008 and January 2013, originally for a genetic study that detected association between leprosy and variants of the *TLR1* [57] and *NOD2* [58] genes. For all databases, leprosy diagnosis/classification was defined after detailed dermatological and neurological examination by specialized leprologists, complemented by bacilloscopy and histopathology of skin lesion. All cases were classified following the R&J scheme [5]. Patients were followed up for at least 2 years since diagnosis to monitoring LR occurrence. Controls were leprosy affected patients who did not presented LR at the time of initial diagnosis or during follow-up.

All patients were treated for leprosy according to WHO MDT guidelines and for LR with the appropriate therapy. All subjects were evaluated for an extensive list of clinical, socioeconomic, and demographic information. This study was approved by the Brazilian Committee for Ethics in Research (CONEP) (protocol 1.722.447). All patients signed an informed consent to participate in the original study; for patients <18 years old, the informed consent was signed by one of the parents or the legal guardian.

Variable selection

The four databases included in this study were composed by clinical and laboratorial parameters, most of it obtained for descriptive, epidemiological purposes not related to the occurrence of LR. Each one of the databases was subjected individually to a two-step, unbiased process aiming to identify those variables exerting the highest impact upon risk of LR, thus, to be included in the system, as follows:

a. Frequency, redundancy and grouping

The first selection step consisted of removing variables with low frequency (less than 15%) of occurrence and/or that were mutually correlated (redundant), consequently capturing the same information. In the case of redundant variables, the most frequent was selected to capture the information of the set.

b. Data mining

Data mining is one of the main stages of the knowledge extraction process from large databases, also known as KDD – Knowledge Discovery in Databases [59]. This AI method is defined as the process of discovering patterns in data to generate useful information for decision-making [60]. WEKA (Waikato Environment for Knowledge Analysis) is an open source program with a collection of algorithm implementations of various data mining techniques, as pre-processing, classification, clustering and visualization [61]. In this study, WEKA was used in the second step of variable selection with the objective of identifying those hierarchically important for LR occurrence in the population samples. The variables were selected using the C4.5 algorithm, which creates a decision tree and identifies the most relevant and non-redundant variants in the process, thus reducing the number of attributes. The C4.5 selection is made according to the gain ratio, which is a normalization of the information gain, a parameter based on the entropy measure (originating from information theory) closely related to the maximum likelihood estimations (MLE), and usually used to make inferences about parameters of the underlying probability distribution from a given dataset [62-66].

The two-step variable selection process was continuously validated by four dermatologists/leprologists with extensive experience in the area.

Finally, two datasets contributed with genotypic information: Goiania for genes *IL6, TNFSF8*, *LRRK2* and *ENSG00000235140*, and Bauru for *TLR1* and *NOD2*, all previously studied in these population samples.

System development

A Bayesian Network (BN) is a graphical model of the posterior conditional probability distribution of an outcome variable based on evidence. It contains nodes that represent the random variables, and links between pairs of nodes, which represent the causal relationship of these nodes, together with a conditional probability distribution in each of the nodes. From the definition, one can deduce that any joint probability distribution may be represented by a Bayesian network, which shows its modeling power: any deterministic model is a particular case of a probabilistic model, and any probabilistic model may be represented as a Bayesian network [66, 67].

The system was created as a BN using Shell NETICA (Norsys Software Corporation) [68] with a dynamic interface that is customized considering the amount of variables in the database. The system was designed to operate with complete or partial information, which is of critical importance considering the translational bias of the proposal and the fact that several leprosy centers may not have access to all the information included, particularly the molecular genetic data. The system loads a spread sheet in which columns and lines refer to the variables and records, respectively. Each variable (columns) is related to one node of the BN. The variables are made up of demographic, clinical, laboratorial, and genetic data (markers). For each one of the databases, two groups were formed randomly to create the network: the test file, with 30% of patients, and the training file with 70% of patients, both stored in an Excel file format.

The performance of the system was assessed by its accuracy, sensitivity, specificity, and both negative and positive predictive values. Predictive values were estimated using the prevalence of occurrence of reversal reactions observed for the studied population samples. The feature importance was measured also using theF_1score, which is the harmonic mean between positive predictive value (PPV) and sensitivity. The F_1 score was calculated accordingly to the equation $F_1 score = 2 * \frac{PPV * sensitivity}{PPV + sensitivity}$ using Python 3.7.9.

Results

Table 1 summarizes information on age, gender, and clinical form of leprosy according to the R&J classification system for T1R, T2R and control groups of all population samples.

Table 1
Distribution of sex, age at diagnosis and clinical type of disease of leprosy affected individuals with T1R, T2R and controls in each population sample.

	Patients, No. (%)														
	Goiania			**Fortaleza**			**Manaus**			**Bauru**			**Combined**		
Age, Years (Mean ± SD)	44.63 ± 16.67			45.15 ± 14.25			40.00 ± 15.39			59.00 ± 18.04			48.00 ± 17.29		
Sex															
Male	234 (57.1)			352 (66.0)			100 (72.9)			258 (69.5)			944 (65.1)		
Female	175 (42.9)			181 (34.0)			37 (27.1)			113 (30.5)			506 (34.9)		
Ridley &Jopling Classification	**Controls**	**T1R**	**T2R**	**Controls**	**T1R**	**T2R**	**Controls**	**T1R**	**T2R**	**Controls**	**T1R**	**T2R**	**Controls**	**T1R**	**T2R**
TT	22	0	0	28	0	0	16	0	0	34	0	0	100	0	0
BT	124	79	0	164	24	0	36	4	0	18	30	0	342	137	0
BB	16	29	3	12	14	0	2	3	0	27	27	1	57	73	4
BL	26	46	8	47	71	66	12	28	10	12	20	33	97	165	117
LL	28	0	28	33	0	68	5	0	16	66	0	102	132	0	214
I	0	0	0	6	0	0	5	0	0	1	0	0	12	0	0
Proportion per Group	52.9	37.6	9.5	54.4	20.5	25.1	55.5	25.5	19.0	42.6	20.8	36.6	51.0	25.9	23.1
Total	409			533			137			371			1450		

Legend: BB: Borderline Borderline; BL: Borderline Lepromatous; BT: Borderline Tuberculoid; I: Indeterminate Leprosy; LL: Lepromatous Leprosy; SD: Standard Deviation; TT: Tuberculoid Leprosy; T1R: Type-1 Reaction; T2R: Type-2 Reaction.

Our strategy for variable selection led to the inclusion of 34 demographic, clinical, laboratorial, and genetic parameters **(Supplementary Table 1)** related to the occurrence of LR in the population samples (Table 2).

Table 2
Demographical, clinical, laboratorial and genetic variables selected in the study.

Data	Variable Information
Socio-demographic	Sex
	Age group
	Ethnicity
Clinical	Multidrug therapy
	First signs and symptoms
	Ridley-Jopling classification
	Number of skin lesions
	Type of lesion
	Color of lesion
	Sensibility testing
Laboratory	Bacilloscopic index
	Histological index
	PGL-1
Genetic	*IL6* markers (4)
	NOD2 marker (1)
	TLR1 markers (2)
	TNFSF8 markers (4)
	*ENSG00000235140*markers (4)
	LRRK2 markers (3)
Family History	First degree[b]
	Second degree[c]
	Contact[d]
Legend:	
[a] Self-Report in years since noticing the early signs and symptoms of leprosy.	
[b] Father, mother, child and sibs affected by leprosy.	
[c] Cousins, nephews, uncles/aunts, grandparents and grandchildren affected by leprosy.	
[d] Close household contact affected by leprosy.	

Since the initial set of variables was not the same across the four datasets – thus, the variables selected by the two-step process and validated by the specialists were not necessarily the same – the prediction system was designed to include all variables selected in each population sample. Detailed information about the distribution of the included variables across the four different datasets is available at **Supplementary Table 2**.

The risk-prediction system was developed to enable the use of each one of the four databases individually as reference, as well as to use a single, combined dataset, thus allowing for customization and facilitating the inclusion of new data sets. The system – named SEPAREH (from Portuguese: *Sistema Especialista Para Avaliação de Risco de Estado Reacional em Hanseníase*; in English: Specialist System for Evaluation of Risk of Occurrence of Reactional States in Leprosy) is designed to present a friendly graphical user interface (Figure 1), which allows the primary care professional to use it intuitively. Variation of the patient's risk of developing one of the two types of LR is shown in real time, as each available clinical and/or genetic data is included in the interface. The platform can be accessed for free at https://orfeu.ppgia.pucpr.br:7200/home.[1]

The overall sensitivity and specificity of the system, as estimated using the combined dataset of 1,450 patients, was 79.3% (95% CI 73.9 – 84.7) and 86.2% (95% CI 81.6 – 90.8), respectively. Accuracy reached 82.7% (95% CI 79.2 – 86.3), and positive and negative predicted values were 85.1% (95% CI 80.2 – 90.1) and 80.6% (95% CI 75.5 – 85.7), respectively.

In order to assess the importance of each of the variables individually, modeling was carried out after removing one at the time, and the impact on system performance was measured through changes in sensitivity, specificity and F1 **(Suppl. Table 1)**. As summarized on Figure 2, the three attributes exerting the highest impact were R&J classification, combined genetic markers, and histological index.

Interestingly, the highest estimates of accuracy, sensitivity, specificity and both negative and positive predictive values were observed for the Bauru and the Goiania datasets, for which genotypic data was available, even higher than what was observed for the combined dataset of much larger sample size (the only exception being the positive predictive value for Bauru: 82.7% vs. 85.1% for the combined dataset) (Table 3).

Table 3
Results obtained for each population sample.

Population Sample	Two-by-Two Contingency				Results	95% CI
Combined		**Control**	**LR**	**Total**	**Sensitivity =** 79.3%	73.9% - 84.7%
	Control	187	45	232	**Specificity =** 86.2%	81.6% - 90.8%
	LR	30	172	202	**PVP =** 85.1%	80.2% - 90.1%
	Total	217	217	434	**PVN =** 80.6%	75.5% - 85.7%
					Accuracy = 82.7%	79.2% - 86.3%
Goiania		**Control**	**LR**	**Total**	**Sensitivity =** 85.7%	76.5% - 94.9%
	Control	59	8	67	**Specificity =** 89.4%	82.0% - 96.8%
	LR	7	48	55	**PVP =** 87.3%	78.5% - 96.1%
	Total	66	56	122	**PVN =** 88.0%	80.3% - 95.8%
					Accuracy = 87.7%	81.9% - 93.5%
Bauru		**Control**	**LR**	**Total**	**Sensitivity =** 82.7%	72.4% - 93.0%
	Control	51	9	60	**Specificity =** 85.0%	76.0% - 94.0%
	LR	9	43	52	**PVP =** 82.7%	72.4% - 93.0%
	Total	60	52	112	**PVN =** 85.0%	76.0% - 94.0%
					Accuracy = 83.9%	77.1% - 90.7%
Fortaleza		**Control**	**LR**	**Total**	**Sensitivity =** 78.1%	68.6% - 87.6%
	Control	62	16	78	**Specificity =** 71.3%	61.8% - 80.8%
	LR	25	57	82	**PVP =** 69.5%	59.5% - 79.5%
	Total	87	73	160	**PVN =** 79.4%	70.5% - 88.4%
					Accuracy = 74.3%	67.6% - 81.1%
Manaus		**Control**	**LR**	**Total**	**Sensitivity =** 77.8%	58.6% - 97.0%
	Control	18	4	22	**Specificity =** 78.3%	61.4% - 95.1%
	LR	5	14	19	**PVP =** 73.7%	53.9% - 93.5%
	Total	23	18	41	**PVN =** 81.8%	65.7% - 97.9%
					Accuracy = 78.0%	65.4% - 90.7%

Legend: LR: Leprosy Reactions; PPV: Positive Predictive Value; NPV: Negative Predictive Value; CI: Confidence Interval.

[1]The access to the platform is limited to HTTPS protocol. In case of difficulty to access the platform, please certify whether HTTPS is being used.

Discussion

Leprosy as an outcome of contact with its causative agent is controlled by multiple environmental and socioeconomic factors and innate characteristics of both host and pathogen. The specific contribution of each of these factors for the total risk of developing leprosy and its endophenotypes is widely unknown. Today, LRs are a major cause of disabilities associated with leprosy; thus, predicting patients at higher risk of developing LR at the time of leprosy diagnosis may help prevent permanent neural impairment. However, an accurate estimate of this risk demands the analysis of a very complex set of variables, which is difficult – if not impossible – to be performed by an unassisted primary health care professional. Here we present an easy-to-use, flexible, and automated system able to identify leprosy patients at increased risk of developing LR, based on clinical, socio-economical, laboratorial and

genetic data. Patients at high risk are candidates for close monitoring during and after treatment, aiming to a prompt management of these aggressive events, minimizing the likelihood of permanent disabilities. Our platform translates basic scientific data into a direct application that will have an immediate impact on the quality of life of leprosy patient as well as on the effectiveness of leprosy control programs.

The three features that exerted the highest impact upon the performance of the system were the R&J classification, histological index, and the combined effect of the genetic markers (Figure 2). The R&J class is a well-accepted major risk factor for the occurrence of reversal reactions [9, 14, 22, 23, 37, 40, 41]. As expected, simulations confirm that patients in the tuberculoid pole of the spectrum tend to have a higher chance of developing no reversal reaction (98%~ when the classification is TT). As clinical form moves towards borderline, the probability of a T1R raise from <1 – 53%~ when the classification is BB and, finally, patients at the lepromatous pole have higher risk of developing T2R – more specifically 61%~ when the classification is LL.

The second top-three parameter impacting upon the system is the histological index. An index equals to 2+ increases the risk of T1R to 56%~; and values higher than 5+ shifts the risk towards T2R – 45%~ when histological index is 6+. This behavior is expected since an increase in the histological index is highly correlated with a higher bacterial load, and consequently a move towards the lepromatous pole of the disease. Histological index higher than 5+ is also a well-known risk factor to develop T2R [38, 39].

Finally, genetic data seems to be of critical importance to improve the performance of the system (Figure 2), which suggests that the understanding of true, exact nature of LR depend on the description of the underlying genetic mechanisms.

We are aware of the limitations of the study: we have had limited access to genetic information across the population samples. The inclusion of genotypic data for additional, known LR susceptibility genes would likely have a positive impact on the performance of the system. In addition, the heterogeneity of the databases, originally obtained for independent studies of distinct designs, prevented a comprehensive analysis of the performance of the system, which we understand that was yet quite remarkable, likely due to the ability of Bayesian systems to estimate risk using all available – even if partial – information. This is important considering that not all leprosy centers across the globe will have access to molecular data of all the patients; in these cases, the platform can still be useful for estimating risk of LR using only the clinical/laboratorial and demographic data with fair sensitivity and specificity, as observed for the Fortaleza and Manaus datasets (Table 3). For a comprehensive evaluation and refining of the system, datasets enrolled prospectively with these specific purposes will be necessary.

Declarations

Ethics approval and consent to participate

The study was approved by the ethics committee of the Pontifical Catholic University of Paraná, reference number 1.722.447. All the patients signed a declaration of consent to participate in this study.

Availability of data and materials

All data used in the development of SEPAREH is available upon request directly to the corresponding author.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This work was supported by the Araucaria Foundation (Grant #41617.433.32610.10092013); the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Leprosy Research Initiative (LRI)/Turing Foundation, grant ID# 704.16.31. MTM is a Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) productivity (PQ) researcher level 2, grant #304368/2018-0. MMAS is under a research fellowship grant from the Brazilian Research Council/CNPq (Grant # 311986-2019-6).

Authors' contributions

RSAR, EFJH, DRC, CMCM, JCN contributed with the definition of the AI-based protocol and data modeling. LFH, GER and EFJH developed the on-line platform. MMAS contributed to recruitment and clinical characterization of the patients of the Tropical Pathology and Public Health Institute, Goiania, Goiás. ACPL, CTS, AB, and PSR contributed to recruitment and clinical characterization of the patients of the Lauro de Souza Lima Institute, Bauru, São Paulo. MAAP, HSG contributed to recruitment and clinical characterization of the patients of the Dona Libânia Dermatology Centre, Fortaleza, Ceará. RCSC contributed to recruitment and clinical characterization of the patients of the Alfredo da Matta Foundation, Manaus, Amazonas. MO contributed with the statistical analysis. VMF contributed to the generation of the original genetic data. SBM, GOP and MLFP contributed to the coordination of the original study under which the population samples were recruited and characterized. RSAR, EFJH, CMCM, JCN and MTM helped the draft of the manuscript. MTM is the principal investigator, the main responsible for the study design and execution, and provided senior supervision throughout the study. All authors read and approved the final version of the manuscript.

Acknowledgments

We are grateful to the patients and staffs of the Reference Center for Diagnosis and Therapy, Goiania; Dermatological Center Dona Libânia, Fortaleza; Alfredo da Matta Foundation, Manaus; Instituto Lauro de Souza Lima, Bauru, for agreeing and their cooperation in this study. A special thanks to Dr. Cássio Ghidella, MD dermatologist from Rondonia, for the valuable support.

References

1. Mira MT, Alcais A, Van Thuc N, Thai VH, Huong NT, Ba NN, et al. Chromosome 6q25 is linked to susceptibility to leprosy in a Vietnamese population. Nature genetics. 2003 Mar;33(3):412-5. PubMed PMID: 12577057.
2. Sousa; ALM, Fava VM, Sampaio LH, Martelli CM, Costa MB, Mira MT, et al. Genetic and immunological evidence implicates interleukin 6 as a susceptibility gene for leprosy type 2 reaction. The Journal of infectious diseases. 2012 May 1;205(9):1417-24. PubMed PMID: 22459738.
3. Fava VM, Cobat A, Van Thuc N, Latini AC, Stefani MM, Belone AF, et al. Association of TNFSF8 regulatory variants with excessive inflammatory responses but not leprosy per se. The Journal of infectious diseases. 2015 Mar 15;211(6):968-77. PubMed PMID: 25320285.
4. Whorld Health Organization. Global leprosy (Hansen disease) update, 2019: time to step-up prevention initiatives. Weekly epidemiological record. 2020;95:417-40.
5. Ridley DS, Jopling WH. Classification of leprosy according to immunity. A five-group system. International journal of leprosy and other mycobacterial diseases : official organ of the International Leprosy Association. 1966 Jul-Sep;34(3):255-73. PubMed PMID: 5950347.
6. Croft RP, Nicholls PG, Steyerberg EW, Richardus JH, Cairns W, Smith S. A clinical prediction rule for nerve-function impairment in leprosy patients. Lancet. 2000 May 6;355(9215):1603-6. PubMed PMID: 10821364.
7. Whorld Health Organization. Guidelines for the Diagnosis, Treatment and Prevention of Leprosy. 2018:1-106.
8. Abel L, Dessein AJ. The impact of host genetics on susceptibility to human infectious diseases. Current opinion in immunology. 1997 Aug;9(4):509-16. PubMed PMID: 9287190.
9. Sauer ME, Salomao H, Ramos GB, D'Espindula HR, Rodrigues RS, Macedo WC, et al. Genetics of leprosy: expected and unexpected developments and perspectives. Clinics in dermatology. 2015 Jan-Feb;33(1):99-107. PubMed PMID: 25432815.
10. Jacobson RR, Krahenbuhl JL. Leprosy. Lancet. 1999 Feb 20;353(9153):655-60. PubMed PMID: 10030346.
11. Alter A, Grant A, Abel L, Alcais A, Schurr E. Leprosy as a genetic disease. Mammalian genome : official journal of the International Mammalian Genome Society. 2011 Feb;22(1-2):19-31. PubMed PMID: 20936290.
12. Sampaio LH, Stefani MM, Oliveira RM, Sousa AL, Ireton GC, Reed SG, et al. Immunologically reactive M. leprae antigens with relevance to diagnosis and vaccine development. BMC infectious diseases. 2011;11:26. PubMed PMID: 21269435. Pubmed Central PMCID: 3040138.
13. Walker SL, Lockwood DN. Leprosy type 1 (reversal) reactions and their management. Leprosy review. 2008 Dec;79(4):372-86. PubMed PMID: 19274984.
14. Britton WJ, Lockwood DN. Leprosy. Lancet. 2004 Apr 10;363(9416):1209-19. PubMed PMID: 15081655.
15. Reddy BN, Bansal RD. An epidemiological study of leprosy disability in a leprosy endemic rural population of Pondicherry (south India). Indian journal of leprosy. 1984 Apr-Jun;56(2):191-9. PubMed PMID: 6548489.
16. Girdhar M, Arora SK, Mohan L, Mukhija RD. Pattern of leprosy disabilities in Gorakhpur (Uttar Pradesh). Indian journal of leprosy. 1989 Oct;61(4):503-13. PubMed PMID: 2533603.
17. Zhang G, Li W, Yan L, Yang Z, Chen X, Zheng T, et al. An epidemiological survey of deformities and disabilities among 14,257 cases of leprosy in 11 counties. Leprosy review. 1993 Jun;64(2):143-9. PubMed PMID: 8341117.
18. Fava V, Orlova M, Cobat A, Alcais A, Mira M, Schurr E. Genetics of leprosy reactions: an overview. Memorias do Instituto Oswaldo Cruz. 2012 Dec;107 Suppl 1:132-42. PubMed PMID: 23283464.
19. Cambri G, Mira MT. Genetic Susceptibility to Leprosy-From Classic Immune-Related Candidate Genes to Hypothesis-Free, Whole Genome Approaches. Frontiers in immunology. 2018;9:1674. PubMed PMID: 30079069. Pubmed Central PMCID: 6062607.
20. Fava VM, Xu YZ, Lettre G, Van Thuc N, Orlova M, Thai VH, et al. Pleiotropic effects for Parkin and LRRK2 in leprosy type-1 reactions and Parkinson's disease. Proceedings of the National Academy of Sciences of the United States of America. 2019 Jul 30;116(31):15616-24. PubMed PMID: 31308240. Pubmed Central PMCID: 6681704.
21. Scollard DM, Smith T, Bhoopat L, Theetranont C, Rangdaeng S, Morens DM. Epidemiologic characteristics of leprosy reactions. International journal of leprosy and other mycobacterial diseases : official organ of the International Leprosy Association. 1994 Dec;62(4):559-67. PubMed PMID: 7868954.
22. Pocaterra L, Jain S, Reddy R, Muzaffarullah S, Torres O, Suneetha S, et al. Clinical course of erythema nodosum leprosum: an 11-year cohort study in Hyderabad, India. The American journal of tropical medicine and hygiene. 2006 May;74(5):868-79. PubMed PMID: 16687695.
23. Van Brakel WH, Khawas IB, Lucas SB. Reactions in leprosy: an epidemiological study of 386 patients in west Nepal. Leprosy review. 1994 Sep;65(3):190-203. PubMed PMID: 8942150.
24. Ranque B, Nguyen VT, Vu HT, Nguyen TH, Nguyen NB, Pham XK, et al. Age is an important risk factor for onset and sequelae of reversal reactions in Vietnamese patients with leprosy. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2007 Jan 1;44(1):33-40. PubMed PMID: 17143812.
25. Sousa AL, Stefani MM, Pereira GA, Costa MB, Rebello PF, Gomes MK, et al. Mycobacterium leprae DNA associated with type 1 reactions in single lesion paucibacillary leprosy treated with single dose rifampin, ofloxacin, and minocycline. The American journal of tropical medicine and hygiene.

2007 Nov;77(5):829-33. PubMed PMID: 17984336.

26. Saunderson P, Gebre S, Byass P. Reversal reactions in the skin lesions of AMFES patients: incidence and risk factors. Leprosy review. 2000 Sep;71(3):309-17. PubMed PMID: 11105489.
27. van Brakel WH, Khawas IB. Nerve function impairment in leprosy: an epidemiological and clinical study--Part 2: Results of steroid treatment. Leprosy review. 1996 Jun;67(2):104-18. PubMed PMID: 8684251.
28. Kumar B, Dogra S, Kaur I. Epidemiological characteristics of leprosy reactions: 15 years experience from north India. International journal of leprosy and other mycobacterial diseases : official organ of the International Leprosy Association. 2004 Jun;72(2):125-33. PubMed PMID: 15301592.
29. Misch EA, Macdonald M, Ranjit C, Sapkota BR, Wells RD, Siddiqui MR, et al. Human TLR1 deficiency is associated with impaired mycobacterial signaling and protection from leprosy reversal reaction. PLoS neglected tropical diseases. 2008 May 7;2(5):e231. PubMed PMID: 18461142. Pubmed Central PMCID: 2330092.
30. Bochud PY, Hawn TR, Siddiqui MR, Saunderson P, Britton S, Abraham I, et al. Toll-like receptor 2 (TLR2) polymorphisms are associated with reversal reaction in leprosy. The Journal of infectious diseases. 2008 Jan 15;197(2):253-61. PubMed PMID: 18177245. Pubmed Central PMCID: 3077295.
31. Rego JL, de Lima Santana N, Machado PRL, Ribeiro-Alves M, de Toledo-Pinto TG, Castellucci LC, et al. Whole blood profiling of leprosy type 1(reversal) reactions highlights prominence of innate immune response genes. BMC infectious diseases. 2018 Aug 24;18(1):422. PubMed PMID: 30143000. Pubmed Central PMCID: 6109344.
32. Teixeira MA, Silva NL, Ramos Ade L, Hatagima A, Magalhaes V. [NRAMP1 gene polymorphisms in individuals with leprosy reactions attended at two reference centers in Recife, northeastern Brazil]. Revista da Sociedade Brasileira de Medicina Tropical. 2010 May-Jun;43(3):281-6. PubMed PMID: 20563497. Polimorfismos do gene NRAMP1 em individuos com reacoes hansenicas, atendidos em dois Centros de Referencia no Recife, nordeste do Brasil.
33. Sapkota BR, Macdonald M, Berrington WR, Misch EA, Ranjit C, Siddiqui MR, et al. Association of TNF, MBL, and VDR polymorphisms with leprosy phenotypes. Hum Immunol. 2010 Oct;71(10):992-8. PubMed PMID: 20650301. Pubmed Central PMCID: 2941523.
34. Berrington WR, Macdonald M, Khadge S, Sapkota BR, Janer M, Hagge DA, et al. Common polymorphisms in the NOD2 gene region are associated with leprosy and its reactive states. The Journal of infectious diseases. 2010 May 1;201(9):1422-35. PubMed PMID: 20350193. Pubmed Central PMCID: 2853728.
35. Fava VM, Sales-Marques C, Alcais A, Moraes MO, Schurr E. Age-Dependent Association of TNFSF15/TNFSF8 Variants and Leprosy Type 1 Reaction. Frontiers in immunology. 2017;8:155. PubMed PMID: 28261213. Pubmed Central PMCID: 5306391.
36. Fava VM, Manry J, Cobat A, Orlova M, Van Thuc N, Moraes MO, et al. A genome wide association study identifies a lncRna as risk factor for pathological inflammatory responses in leprosy. PLoS genetics. 2017 Feb;13(2):e1006637. PubMed PMID: 28222097. Pubmed Central PMCID: 5340414.
37. Eickelmann M, Steinhoff M, Metze D, Tomimori-Yamashita J, Sunderkotter C. Erythema leprosum--after treatment of Lepromatous Leprosy. Journal der Deutschen Dermatologischen Gesellschaft = Journal of the German Society of Dermatology : JDDG. 2010 Jun;8(6):450-3. PubMed PMID: 19922466.
38. Becx-Bleumink M, Berhe D. Occurrence of reactions, their diagnosis and management in leprosy patients treated with multidrug therapy; experience in the leprosy control program of the All Africa Leprosy and Rehabilitation Training Center (ALERT) in Ethiopia. International journal of leprosy and other mycobacterial diseases : official organ of the International Leprosy Association. 1992 Jun;60(2):173-84. PubMed PMID: 1522359.
39. Manandhar R, LeMaster JW, Roche PW. Risk factors for erythema nodosum leprosum. International journal of leprosy and other mycobacterial diseases : official organ of the International Leprosy Association. 1999 Sep;67(3):270-8. PubMed PMID: 10575406.
40. Kahawita IP, Lockwood DN. Towards understanding the pathology of erythema nodosum leprosum. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2008 Apr;102(4):329-37. PubMed PMID: 18313706.
41. Saunderson P, Gebre S, Byass P. ENL reactions in the multibacillary cases of the AMFES cohort in central Ethiopia: incidence and risk factors. Leprosy review. 2000 Sep;71(3):318-24. PubMed PMID: 11105490.
42. de Messias IJ, Santamaria J, Brenden M, Reis A, Mauff G. Association of C4B deficiency (C4B*Q0) with erythema nodosum in leprosy. Clinical and experimental immunology. 1993 May;92(2):284-7. PubMed PMID: 8485914. Pubmed Central PMCID: 1554813.
43. Schuring RP, Hamann L, Faber WR, Pahan D, Richardus JH, Schumann RR, et al. Polymorphism N248S in the human Toll-like receptor 1 gene is related to leprosy and leprosy reactions. The Journal of infectious diseases. 2009 Jun 15;199(12):1816-9. PubMed PMID: 19456232.
44. Girardi DR, Moro CM, Bulegon H. SeyeS - support system for preventing the development of ocular disabilities in leprosy. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2010;2010:6162-5. PubMed PMID: 21097149.
45. Belle A, Kon MA, Najarian K. Biomedical informatics for computer-aided decision support systems: a survey. TheScientificWorldJournal. 2013;2013:769639. PubMed PMID: 23431259. Pubmed Central PMCID: 3575619.
46. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics. 2008 Feb;77(2):81-97. PubMed PMID: 17188928.
47. Lu Z, Mitchell RM, Smith RL, Karns JS, van Kessel JA, Wolfgang DR, et al. Invasion and transmission of Salmonella Kentucky in an adult dairy herd using approximate Bayesian computation. BMC veterinary research. 2013;9:245. PubMed PMID: 24304969. Pubmed Central PMCID: 4235045.

48. Souza WV, Barcellos CC, Brito AM, Carvalho MS, Cruz OG, Albuquerque MF, et al. [Empirical bayesian model applied to the spatial analysis of leprosy occurrence]. Revista de saude publica. 2001 Oct;35(5):474-80. PubMed PMID: 11723520. Aplicacao de modelo bayesiano empirico na analise espacial da ocorrencia de hanseniase.
49. Smith RL, Grohn YT. Use of Approximate Bayesian Computation to Assess and Fit Models of Mycobacterium leprae to Predict Outcomes of the Brazilian Control Program. PloS one. 2015;10(6):e0129535. PubMed PMID: 26107951. Pubmed Central PMCID: 4479607.
50. Crump RE, Medley GF. Back-calculating the incidence of infection of leprosy in a Bayesian framework. Parasites & vectors. 2015;8(1):534. PubMed PMID: 26490744. Pubmed Central PMCID: 4618872.
51. Joshua V, Mehendale S, Gupte MD. Bayesian model, ecological factors & transmission of leprosy in an endemic area of South India. The Indian journal of medical research. 2016 Jan;143(1):104-6. PubMed PMID: 26997022. Pubmed Central PMCID: 4822349.
52. Zhang X, Yuan Z, Ji J, Li H, Xue F. Network or regression-based methods for disease discrimination: a comparison study. BMC medical research methodology. 2016 Aug 18;16:100. PubMed PMID: 27538955. Pubmed Central PMCID: 4991108.
53. Wang N, Wang Z, Wang C, Fu X, Yu G, Yue Z, et al. Prediction of leprosy in the Chinese population based on a weighted genetic risk score. PLoS neglected tropical diseases. 2018 Sep;12(9):e0006789. PubMed PMID: 30231057. Pubmed Central PMCID: 6166985.
54. Gama RS, Souza MLM, Sarno EN, Moraes MO, Goncalves A, Stefani MMA, et al. A novel integrated molecular and serological analysis method to predict new cases of leprosy amongst household contacts. PLoS neglected tropical diseases. 2019 Jun;13(6):e0007400. PubMed PMID: 31181059. Pubmed Central PMCID: 6586366.
55. Tio-Coma M, Kielbasa SM, van den Eeden SJF, Mei H, Roy JC, Wallinga J, et al. Blood RNA signature RISK4LEP predicts leprosy years before clinical onset. EBioMedicine. 2021 Jun;68:103379. PubMed PMID: 34090257. Pubmed Central PMCID: 8182229.
56. Penna GO, Pontes MA, Cruz R, Goncalves Hde S, Penna ML, Buhrer-Sekula S. A clinical trial for uniform multidrug therapy for leprosy patients in Brazil: rationale and design. Memorias do Instituto Oswaldo Cruz. 2012 Dec;107 Suppl 1:22-7. PubMed PMID: 23283449.
57. Marques Cde S, Brito-de-Souza VN, Guerreiro LT, Martins JH, Amaral EP, Cardoso CC, et al. Toll-like receptor 1 N248S single-nucleotide polymorphism is associated with leprosy risk and regulates immune activation during mycobacterial infection. The Journal of infectious diseases. 2013 Jul;208(1):120-9. PubMed PMID: 23547143.
58. Sales-Marques C, Salomao H, Fava VM, Alvarado-Arnez LE, Amaral EP, Cardoso CC, et al. NOD2 and CCDC122-LACC1 genes are associated with leprosy susceptibility in Brazilians. Human genetics. 2014 Dec;133(12):1525-32. PubMed PMID: 25367361.
59. Witten IH, Frank E. Data Mining Practical Machine Learning Tools and Techniques. Elsevier, editor2005.
60. Bellazzi R, Ferrazzi F, Sacchi L. Predictive Data Mining in Clinical Medicine: A Focus on Selected Methods and Applications. WIREs Data Mining and Knowledge Discovery. 2011.
61. The University of Waikato. Weka 3: Data Mining Software in Java. https://wwwcswaikatoacnz/ml/weka/. 2018.
62. Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. . Mach Learn 1994;16:235–40.
63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. 2009;11(1):10-8.
64. Ruggieri S. Efficient C4.5. IEEE Transactions on Knowledge and Data Engineering. 2002;14(2):438-44.
65. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research. 2016;17:1-5.
66. Liu S, Ge Z, McGree J, Xie Y. Computational and Statistical Methods for Analysing Big Data with Applications. Academic Press. 2016:1-194.
67. Ben-Gal I. Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability. 2007:1-6.
68. Norsys Software Corp. Netica API Programmer's Library. Reference Manual. Version 4.18. 2010.

Figures

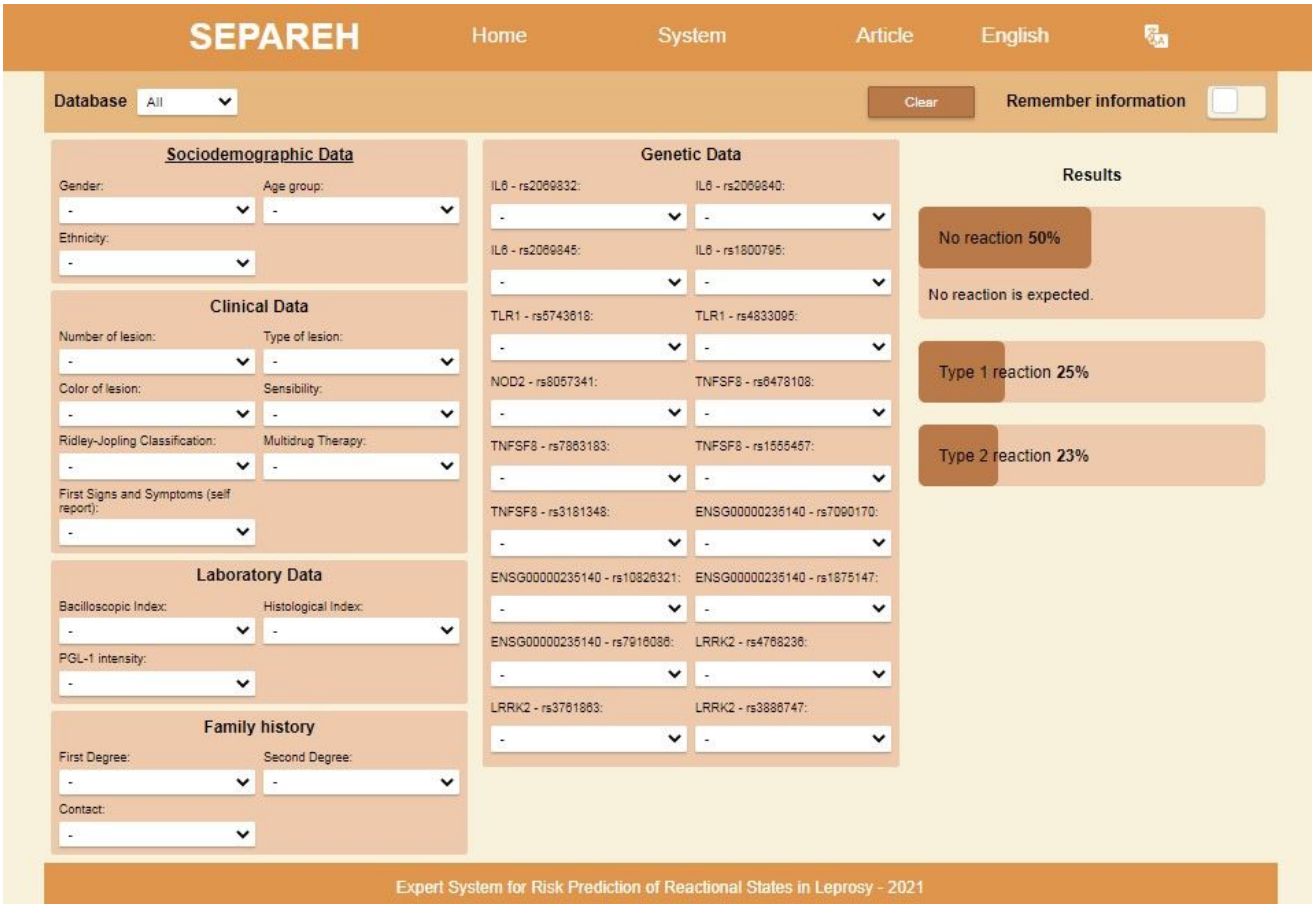

Figure 1

System Interface

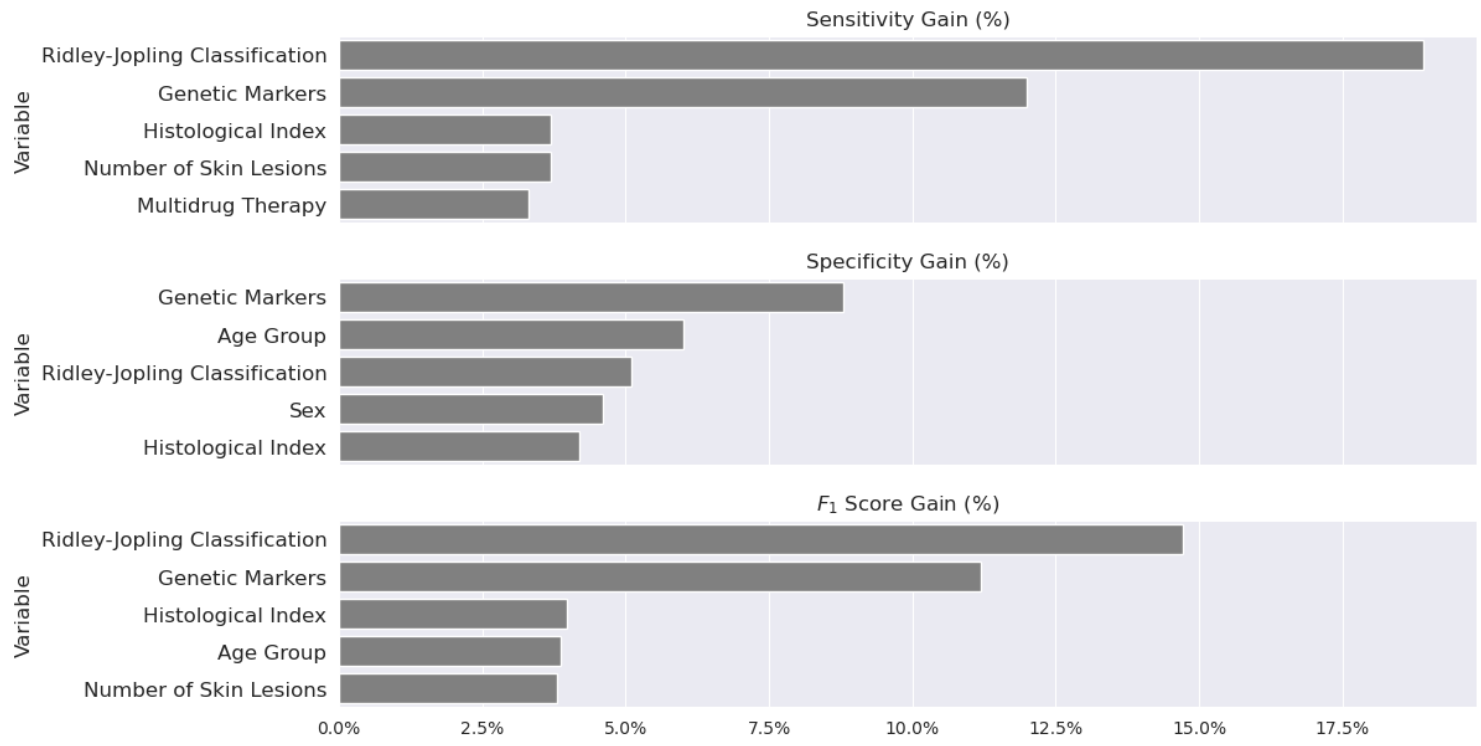

Figure 2

Top 5 most important features measured in relative gain using sensitivity, specificity, and the F1 score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- JTransMedTablesS1andS2.docx