Modeling using clinical examination indicators predicts interstitial lung disease among patients with rheumatoid arthritis

Yao Wang [1,2] , Wuqi Song [1] , Jing Wu [1] , Zhangming Li [3] , Fengyun Mu [4] , Yang Li [5] , He Huang [5] , Wenliang Zhu [6] , Fengmin Zhang [Corresp. 1]

[1] Department of Microbiology, Wu Lien-Teh Institute, Harbin Medical University, Harbin, Heilongjiang Province, China

[2] Department of Microbiology and Immunology, School of Basic Medical Sciences, Heilongjiang University of Chinese Medicine, Harbin, Heilongjiang Province, China

[3] Department of Pharmacy Administration, Harbin Medical University, Harbin, Heilongjiang Province, China

[4] Department of Laboratory Medicine, The Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, China

[5] Department of Rheumatology, The Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, China

[6] Institute of Clinical Pharmacology, The Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, China

Corresponding Author: Fengmin Zhang
Email address: fengminzhang@ems.hrbmu.edu.cn

Interstitial lung disease (ILD) is a severe extra-articular manifestation of rheumatoid arthritis (RA) that is well-defined as a chronic systemic autoimmune disease. A proportion of patients with RA-associated ILD (RA-ILD) develop pulmonary fibrosis (PF), resulting in poor prognosis and increased lifetime risk. We investigated whether routine clinical examination indicators (CEIs) could be used to identify RA patients with high PF risk. A total of 533 patients with established RA were recruited in this study for model building and 32 CEIs were measured for each of them. To identify PF risk, a new artificial neural network (ANN) was built, in which inputs were generated by calculating Euclidean distance of CEIs between patients. Receiver operating characteristic curve analysis indicated that the ANN performed well in predicting the PF risk (Youden index = 0.436) by only incorporating four CEIs including age, eosinophil count, platelet count, and white blood cell count. A set of 218 RA patients with healthy lungs or suffering from ILD and a set of 87 RA patients suffering from PF were used for independent validation. Results showed that the model successfully identified ILD and PF with a true positive rate of 84.9% and 82.8%, respectively. The present study suggests that model integration of multiple routine CEIs contributes to identification of potential PF risk among patients with RA.

1   **Modeling using clinical examination indicators predicts interstitial lung disease among**

2   **patients with rheumatoid arthritis**

3   Wang Yao[1,2], Song Wuqi[1], Wu Jing[1], Li Zhangming[3], Mu Fengyun[4], Li Yang[5], Huang He[5], Zhu

4   Wenliang[6], Zhang Fengmin[1,#]

5   [1]Department of Microbiology, Wu Lien-Teh Institute, Harbin Medical University, Harbin,

6   Heilongjiang Province, China

7   [2]Department of Microbiology and Immunology, School of Basic Medical Sciences, Heilongjiang

8   University of Chinese Medicine, Harbin, Heilongjiang Province, China

9   [3]Department of Pharmacy Administration, Harbin Medical University, Harbin, Heilongjiang

10   Province, China

11   [4]Department of Laboratory Medicine, the Second Affiliated Hospital of Harbin Medical

12   University, Harbin, Heilongjiang Province, China

13   [5]Department of Rheumatology, the Second Affiliated Hospital of Harbin Medical University,

14   Harbin, Heilongjiang Province, China

15   [6]Institute of Clinical Pharmacology, the Second Affiliated Hospital of Harbin Medical University,

16   Harbin, Heilongjiang Province, China

17   [#]Corresponding author: Zhang Fengmin, Ph.D., 157 Baojian Road, Harbin, Heilongjiang Province,

18   150081, China; Phone/Fax: 86-451-8668-9576, Email: fengminzhang@ems.hrbmu.edu.cn.

19 **Abstract**

20 Interstitial lung disease (ILD) is a severe extra-articular manifestation of rheumatoid arthritis (RA)

21 that is well-defined as a chronic systemic autoimmune disease. A proportion of patients with RA-

22 associated ILD (RA-ILD) develop pulmonary fibrosis (PF), resulting in poor prognosis and

23 increased lifetime risk. We investigated whether routine clinical examination indicators (CEIs)

24 could be used to identify RA patients with high PF risk. A total of 533 patients with established

25 RA were recruited in this study for model building and 32 CEIs were measured for each of them.

26 To identify PF risk, a new artificial neural network (ANN) was built, in which inputs were

27 generated by calculating Euclidean distance of CEIs between patients. Receiver operating

28 characteristic curve analysis indicated that the ANN performed well in predicting the PF risk

29 (Youden index = 0.436) by only incorporating four CEIs including age, eosinophil count, platelet

30 count, and white blood cell count. A set of 218 RA patients with healthy lungs or suffering from

31 ILD and a set of 87 RA patients suffering from PF were used for independent validation. Results

32 showed that the model successfully identified ILD and PF with a true positive rate of 84.9% and

33 82.8%, respectively. The present study suggests that model integration of multiple routine CEIs

34 contributes to identification of potential PF risk among patients with RA.

Introduction

Rheumatoid arthritis (RA) is a common chronic systemic autoimmune disorder mainly characterized by joint inflammation. Apart from articular tissue, multiple other tissues and organs may be involved in the pathological process of RA. Indeed, extra-articular manifestations (EAMs) have become the main cause of the morbidity and mortality of patients with RA [1,2]. Among the recognized EAMs, interstitial lung disease (ILD) follows cardiac manifestations [3] as the second contributor to the excess mortality (10% to 20%) of patients with RA [4]. Compared with the median survival of 9.9 years for RA alone, patients with RA-associated ILD (RA-ILD) have been reported to have poor prognosis with a median survival of 2.6 years [5]. However, the situation might be worse if pulmonary fibrosis (PF) is also confirmed. In a clinical study focusing on fibrotic interstitial pneumonia, Solomon and colleagues verified that RA patients with fibrotic ILD had worse survival than those with non-fibrotic ILD [6], indicating that PF is an independent risk for mortality in RA.

PF causes the aggressive deterioration of lung function and leads to poor prognosis of RA. Unfortunately, fibrotic ILD, especially the subtype usual interstitial pneumonia, still lacks targeted therapy [7]. This leads to worse outcomes in such patients. Nevertheless, early identification of patients with high PF risk would definitely benefit individual RA patient management, in which joint participation of multidisciplinary doctors has been suggested for driving treatment decisions [8]. In the clinical context, any decision making is dependent on diversified clinical examinations including a large number of clinical examination indicators (CEIs). The collection of numerous

55 CEIs comprehensively reflects the current pathophysiological condition of the patient. In this

56 study, we hypothesized that integration of the CEIs may reveal the potential risk of an individual

57 patient suffering from PF. If the assumption is valid, early risk assessment may be made on

58 admission. To validate the rationality and feasibility of this hypothesis, we performed a

59 retrospective study, in which 620 patients with established RA were included and their clinical

60 examination results were retrieved from the electronic medical records system of the hospital. A

61 novel artificial neural network (ANN) was considered for abstracting and integrating significant

62 information related to PF risk from CEIs. Especially, rather than the traditional ANN and neural

63 network cascade previously described [9-11], the new ANN built here was thought to be a

64 derivative information integration system, in which inputs were generated by calculating the

65 Euclidean distances of CEIs between patients. In conclusion, such an effort aims to provide a

66 viable clinical approach to identify patients with high PF risk and facilitate implementation of early

67 preventive interventions.

68

69 **Methods**

70 **Ethical statement**

71 This study is a retrospective study that was approved by the Ethics Committee of Harbin Medical

72 University (HMU) (Approval number: HMUIRB20150028) and was carried out in accordance

73 with the Declaration of Helsinki.

74 **Patients**

75 The electronic medical record systems of the first and second affiliated hospitals of HMU were

76 used to access the medical records of hospitalized patients that were clinically diagnosed with RA.

77 For each patient, high-resolution computed tomography (HRCT, Discovery CT 750 HD, GE

78 Medical Systems, LLC., Waukesha, WI, USA) was performed to examine whether the

79 complication of ILD or PF was present. In the second affiliated hospital of HMU, 32 CEIs were

80 retrieved from the hospital's electronic system of medical records (form January 1, 2013 to

81 December 31, 2015). All the CEIs were categorized into three classes. The three classes were basic

82 information (2 items), routine blood test (22 items), and routine urine test (8 items). In the first

83 affiliated hospital of HMU, four CEIs including age, eosinophil count (EO), platelet count (PLT),

84 and white blood cell count (WBC) were retrieved from the hospital's electronic system of medical

85 records (from January 1, 2014 to December 31, 2015). If a patient's clinical examination was

86 inadequate, he or she was not taken into consideration for inclusion as a subject. It should be

87 especially noted that there were no other reasons for rejecting subject inclusion, such as age or

88 gender.

89 **Identification of ILD-associated CEIs**

90 The software MedCalc v15.8 (MedCalc, Mariakerke, Belgium) was used to perform receiver

91 operating characteristic (ROC) curve analysis and calculate the Youden index for each CEI. The

92 Youden index was the sum of sensitivity and specificity minus 1 as defined previously [12]. This

93 effort aimed to investigate whether a CEI can be used as a marker to distinguish patients suffering

94 from ILD from those with healthy lungs. It should be noted that only the RA patients with healthy

95   lungs and those suffering from ILD were included in the calculation of the Youden index and the

96   subsequent construction of networks and models. ILD-associated CEIs were identified only when

97   the area under the ROC curve (AUC) was significantly larger than 0.5 ($P < 0.05$), and were retained

98   for further integration by ANN.

99   **Data preprocessing and model integration of ILD-associated CEIs**

100   ILD-associated CEIs were normalized into a 0 to 1 number before further analysis as previously

101   described [9]. The Intelligent Problem Solver (IPS) tool in the software STATISTICA Neural

102   Networks (SNN, Release 4.0E; Statsoft, Tulsa, OK, USA) was applied to construct a radial basis

103   function (RBF)-ANN model to investigate the effect of CEI integration on ILD association. In this

104   study, the model was simply named as ANN I. The model output then underwent normalization

105   processing and Youden index calculation to investigate whether an obvious association with ILD

106   status was still present after CEI integration. The holdout cross-validation method was applied for

107   preliminary validation of the model as IPS randomly divided the patients into three subsets

108   (training set, verification set, and testing set) in a 2:1:1 ratio. Thus, one-quarter of all the patients

109   did not participate in model building and were used for model testing. The IPS calculated

110   correlation coefficients for the training set (R_{Tr}) and the testing set (R_{Te}). The two correlation

111   coefficients measured the correlation between model output and status of ILD. Similar values of

112   R_{Tr} and R_{Te} indicates good generalization ability of the model.

113   **Euclidean distance calculation and construction of patient–patient similarity network**

114   For any two patients, we calculated their Euclidean distance in an *n*-dimensional space, in cases in

115   which ILD-associated CEIs were re-defined as space coordinates. The value of n was the sum of

116   ILD-associated CEIs. Following the clustering algorithm proposed by Rodriguez and Laio [13],

117   we established a patient–patient similarity network (PPSN) by using the network data visualization

118   software Cytoscape v2.8.3 (Institute of Systems Biology, Seattle, WA, USA) [14].

119   **Model integration of derivative information of patients**

120   Construction of the ANN I model and the calculation of Euclidean distance of CEIs between

121   patients was used to obtain derivative information for each patient. For example, four items of

122   derivative information could be obtained for patient i as follows: first, we divided the other patients

123   except patient i into m mutually exclusive divisions of nearly equal size according to the magnitude

124   of their distances to patient i in the n-dimensional space of ILD-associated CEIs. Thus, the four

125   items of derivative information might be the ANN I output of patient i, mean ANN I output of

126   patients in a given division, mean Euclidean distance to the patients in the division, and actual

127   proportion of RA-ILD among patients in the division. The four derivative information items were

128   imported in a new RBF-ANN as inputs to predict patients' potential PF risk, in the same way as

129   the ANN I. In order to distinguish it from ANN I, the new ANN was named as ANN II. We further

130   investigated the effect of different patient groupings on the performance of ANN II. In this study,

131   the division size m was assigned a value of 5, 10, 15 or 20. For example, if we divided patients

132   into five divisions, we could obtain five candidates of ANN II. The Youden index calculation was

133   then used to identify the best model as ANN II.

134   **Model validation and performance evaluation**

135   For ANN II, the 10-fold cross-validation method was used for model validation as previously

136   described [10]. Briefly, all the patients were randomly divided into 10 mutually-exclusive sets of

137   nearly equal size. Next, nine were selected for model training and one was used for model

138   validation. The above procedure was repeated 10 times to allow each of the 10 patient sets to be

139   independently used for validation.

140   To investigate whether the ANN models I and II identify patients with high PF risk, we

141   performed ROC curve analysis on ILD-associated CEIs and outputs of ANN I and ANN II using

142   MedCalc v15.8. Besides AUC, we also recorded the values of sensitivity, specificity, Youden

143   index, and calculated diagnostic odds ratio (DOR) at the optimal cut-off point. The Youden index

144   was the sum of sensitivity and specificity minus 1 as defined previously [12]. According to the

145   definition of DOR in previous studies [15,16], it was calculated as follows:

146
$$DOR = \frac{sensitivity \times specificity}{(1 - sensitivity) \times (1 - specificity)}$$

147   In addition, a set of 87 RA patients suffering from PF (the second affiliated hospital of HMU)

148   was used for independent evaluation of each ILD-associated CEI and ANN models I and II. A

149   further set consisting of 72 RA patients with healthy lungs and 146 RA patients suffering ILD (the

150   first affiliated hospital of HMU) was used for external validation of ANN model II. The true

151   positive rate (TPR) was calculated when the optimal cut-off point in the ROC curve was used as

152   the discriminant threshold. TPR was calculated as follows:

153
$$TPR = \frac{TP}{TP + FN} \times 100\%$$

154   where TP and FN are abbreviations of true positive and false negative.

155   **Statistical analysis**

156   MedCalc v15.8 was used to perform pairwise comparisons of the ROC curves based on the

157   methodology of DeLong et al. [17]. Differences were considered as statistically significant when

158     $P < 0.05$.

159

160   # Results

161   **Patients**

162   A total of 838 patients with RA were included in this study. Among them, 620 patients from the

163   second affiliated hospital of HMU were subjected to complete clinical examination (Tables S1 and

164   S2). HRCT examination verified the complication of PF in 87 of the 620 patients. In addition, 169

165   patients were identified as having healthy lungs and 364 were diagnosed as complicated with ILD.

166   For 218 patients with RA from the first affiliated hospital of HMU, only four CEIs were retrieved

167   from the hospital's electronic system of medical records (Table S3), among which 72 were

168   identified as having healthy lungs and 146 were diagnosed as complicated with ILD.

169   **Identification of ILD-associated CEIs**

170   For each of the 32 CEIs (Table S1), we investigated any potential association with the status of

171   ILD. Compared with patients with healthy lungs, those RA patients suffering from ILD have an

172   implied greater PF risk [3,4]. Eight CEIs were identified as ILD-associated CEIs (Table 1). For an

173   ILD-associated CEI, higher Youden index suggested better effectiveness as diagnostic marker of

174   PF risk (Table 1). For instance, among the 32 CEIs, age was assigned the highest Youden index

175   or 0.301 (Figure 1A). This characteristic was highlighted in the pairwise comparisons of the ROC

176   curves as compared with EO, PLT or WBC (the methodology of DeLong et al., P < 0.001, Figure

177   1B). The Youden index calculation indicated that age and three blood CEIs (EO, PLT, and WBC)

178   were assigned the highest Youden indices, implying relatively closer association with ILD status

179   (Figure 1A). Consistent with this, the optimal ANN model effect was obtained by using the four

180   CEIs (age, EO, PLT, and WBC) as joint inputs of ANN I (Figure S1).

181   **Euclidean distance calculation for networking patients' similarity**

182   By refining the four ILD-associated CEIs (age, EO, PLT, and WBC) as coordinates in a four-

183   dimensional space, we calculated the Euclidean distance between any two patients and mapped a

184   PPSN for the 533 RA patients including 169 with healthy lungs and 364 complicated with ILD

185   (Figure 2). In spite of the application of different edge settings, a huge patient cluster, rather than

186   multiple scattered patient clusters, was always observed in the network.

187   **Application of ANN for integration of ILD-associated CEIs**

188   An RBF-ANN with 4-12-1 architecture, named ANN I, was constructed for integration of the four

189   CEIs that were related to PF status, namely age, EO, PLT, and WBC. Compared with single CEIs,

190   the outputs of ANN I had a better ability to identify patients with high PF risk (Youden index =

191   0.387, Table 2). Furthermore, we built a series of RBF-ANNs with 4-12-1 architecture by

192   calculating Euclidean distance among patients, distributing patients into divisions, and generating

193   derivative indicators (See **Methods**). Regardless of the size of the divisions, the ANN generated

194   by the division containing most similar patients showed the best effect in identifying potential PF

195   risks (Figure 3A). Compared with smaller grouping size, the five division method led to creation

196   of an optimal ANN model, namely ANN II (Youden index = 0.436, Table 2). A significant

197   difference was observed in AUC when ANNs I and II were compared (P = 0.025, Figure 3B),

198   suggesting the advantage of ANN II in identifying potential PF risk. The DOR of ANN I was 5.41

199   and that of ANN II was 6.88. When specificity was fixed as 0.80, the sensitivity of ANN I was

200   0.473 while that of ANN II was 0.591. These assessments indicated that ANN II was more

201   effective and sensitive than ANN I in identifying patients with high PF risk.

202   **Evaluation of the models**

203   The holdout cross-validation method was used for preliminary validation of the ANNs established

204   in this study. Two model generalization ability parameters R_{Tr} and R_{Te} were calculated using SNN

205   software. Similar values of R_{Tr} and R_{Te} indicated that the model built using the IPS tool had good

206   generalization ability. For ANN I, R_{Tr} and R_{Te} were 0.398 and 0.440, respectively, while for ANN

207   II, R_{Tr} and R_{Te} were 0.461 and 0.489, respectively. For ANN II in particular, the 10-fold cross-

208   validation method was applied for further model validation (Figure 3C). The AUC was 0.751,

209   implying effectiveness of ANN II in identifying potential PF risk among the patients. Furthermore,

210   an independent set of 87 RA patients complicated with established PF was used to explore whether

211   patients with PF could be identified by single ILD-associated CEIs, ANN I, or ANN II. Compared

212   with single ILD-associated CEIs, ANNs I and II had better recognition for these patients (TPR =

213   82.8%, Figure 4A). Consistent with this, a dot column chart visually showed that patients with

214   high PF risk and those that had been complicated with PF could be successfully identified with

215   higher sensitivity and specificity using ANN II (Figure 4B). A validation set of patients from the

216   first affiliated hospital of HMU was used for further assessing the model effect of ANN II (Table

217   S3, Figure 4C). The AUC was 0.792, implying the effectiveness of ANN II to identify ILD. A TPR

218   of 84.9% was calculated for identifying ILD by drawing a 2×2 contingency table of the validation

219   set (Table S4).

220

221   **Discussion**

222   Approximately 10% of patients with RA have ILD-related complications, leading to varying

223   degrees of functional and structural impairment of the lungs [18-20]. This demands clinical

224   management targeted to those patients with RA-associated ILD, especially those with RA-

225   associated fibrotic-ILD [21,22]. For this purpose, it is important to develop early identification of

226   patients with high risk of pulmonary complications [23,24]. In the present study, a global analysis

227   of 32 CEIs was performed to reveal clinical predictors that related to high risk of fibrotic-ILD. A

228   modified ANN model was built for CEI integration and detection of risk of fibrotic-ILD among

229   patients with RA.

230   In this study, the Youden index, instead of Spearman's rho, was applied to explore potential

231   associations between ILD status and CEIs without any ANN data transformation. The choice of

232   Youden index for screening disease-associated CEIs was based on the consideration that the

233   presence or absence of ILD belonged to a logical variable rather than a continuous variable. By

234   calculating the Youden index, four of the 32 CEIs (age, EO, PLT, and WBC) were identified as

235   having a relatively closer association with ILD status. This result implies multifactorial

236   involvement of pulmonary complications in RA. Among the four ILD-associated CEIs, age

237   showed the strongest association with ILD status (Youden index = 0.301). Older age might lead

238   to a greater risk of fibrotic ILD. This result was in line with a previous study performed by

239   Yilmazer and colleagues [25]. Their study suggested that age was an independent risk factor for

240   lung damage caused by RA-associated ILD. Obvious immunity system alterations and high

241   infection rate were observed in RA patients with ILD [26,27]. In the present study, EO, PLT, and

242   WBC were found to be significantly associated with RA-ILD. White blood cells are a large

243   category of immune system cells, which prevent damage to the body caused by foreign invaders

244   and infectious disease. Eosinophils are a type of white blood cell, while it has been traditionally

245   recognized that platelets are a type of blood cell which plays a central role in physiological

246   hemostasis and pathological thrombosis. Just recently, it was discovered that platelets also act as

247   inflammatory effector cells and are importantly implicated in pathological infectious and immune

248   responses of the lungs [28]. Taken together, abnormalities in the count of immune cells suggest

249   that the lungs of patients with RA-ILD might be subject to infection leading to further tissue

250   damage, such as pulmonary fibrosis. Obviously, aging reduces the ability of the immune system

251   to prevent disease [29]. However, further investigation will be necessary.

252   A PPSN was established by refining the four ILD-associated CEIs as coordinates in a four-

253   dimensional space and calculating Euclidean distance between any two patients. The existence of

254   a huge cluster of similar patients in the network implied a concentrated distribution of the majority

255   of patients based on the results of routine medical tests (Figure 2). Despite this, it was found that

256   subtle differences in CEIs contributed to heterogeneous local clustering of patients with healthy

257   lungs and those complicated with ILD in the PPSN, suggesting a possibility of joint application of

258   the four CEIs to identify risk of PF.

259   ANN is a universal machine learning method nowadays, which has been widely applied in

260   various areas of medicine, such as decision-making for neurosurgery and prostate cancer diagnosis

261   [30–32]. Application of ANN in medicine has been validated to facilitate disease risk detection

262   and medical decision-making. In our study, a decision-making system with novel ANN model

263   architecture was developed for the purpose of facilitating identification of high PF risk among

264   patients with RA. Our results confirmed that integrated processing of ILD-associated CEIs by the

265   derivative information integration system developed here more effectively identified RA patients

266   with high risk of PF, compared with data processing the same CEIs using a traditional ANN (Table

267   2). Construction of the system differed from the procedure used to build a traditional ANN,

268   requiring multiple operations, including ANN-based CEI integration, establishment of a Euclidean

269   distance-based PPSN, and patient derivative information extraction and integration using an ANN.

270   Although more complex, it was thought that the system would be feasible for clinical practice,

271   because only four routine CEIs were required as network inputs, the model architectures of ANNs

272   I and II were simple, and the Euclidean distance calculation was easy to perform by computer

273   programming.

274   In conclusion, our study for the first time investigated associations between routine CEIs and ILD

275   in patients with RA using a modified ANN system. The results contributed to new knowledge

276   regarding the identification of patients with high PF risk when routine CEIs were used. Integration

277   of CEIs in a mathematical model facilitated their application in clinical management of such

278   patients. Superior to the traditional ANN model, the developed system consisting of ANNs I and

279   II successfully identified patients at high risk of PF and those having PF among patients with RA

280   by co-considering meaningful associations between CEIs and ILD and similar patients in medical

281   testing. However, two limitations should be noted: our findings were obtained from a small

282   collection of RA patients and only 32 routine CEIs were investigated for their association with

283   ILD. Further research on a larger patient set is definitely needed to validate our results and more

284   CEIs should be considered for investigation.

285

References

287   1. Turesson C. Extra-articular rheumatoid arthritis. Curr Opin Rheumatol. 2013; 25(3):360–366.

288   2. Sihvonen S, Korpela M, Laippala P, Mustonen J, Pasternack A. Death rates and causes of death

289   in patients with rheumatoid arthritis: a population-based study. Scand J Rheumatol 2004;

290   33(4):221–227.

291   3. Brown KK. Rheumatoid lung disease. Proc Am Thorac Soc. 2007; 4(5):443–448.

292   4. Ingegnoli F, Lubatti C, Ingegnoli A, Boracchi P, Zeni S, Meroni PL. Interstitial lung disease

293   outcomes by high-resolution computed tomography (HRCT) in anti-Jo1 antibody-positive

294   polymyositis patients: a single centre study and review of the literature. Autoimmun Rev 2012;

295   11(5):335–340.

296   5. Bongartz T, Nannini C, Medina-Velasquez YF, Achenbach SJ, Crowson CS, Ryu JH, Vassallo

297   R, Gabriel SE, Matteson EL. Incidence and mortality of interstitial lung disease in rheumatoid

298   arthritis: a population based study. Arthritis Rheum 2010; 62(6):1583–1591.

299   6. Solomon JJ, Ryu JH, Tazelaar HD, Myers JL, Tuder R, Cool CD, Curran-Everett D, Fischer

300   A, Swigris JJ, Brown KK. Fibrosing interstitial pneumonia predicts survival in patients with

301   rheumatoid arthritis-associated interstitial lung disease (RA-ILD). Respir Med. 2013;

302   107(8):1247–1252. doi: 10.1016/j.rmed.2013.05.002.

303   7. Travis WD, Hunninghake G, King TE Jr, Lynch DA, Colby TV, Galvin JR, Brown KK, Chung

304   MP, Cordier JF, du Bois RM, Flaherty KR, Franks TJ, Hansell DM, Hartman TE, Kazerooni

305   EA, Kim DS, Kitaichi M, Koyama T, Martinez FJ, Nagai S, Midthun DE, Müller NL,

306   Nicholson AG, Raghu G, Selman M, Wells A. Idiopathic nonspecific interstitial pneumonia.

307   Report of an American Thoracic Society project. Am J Respir Crit Care Med 2008;

308   177(12):1338–1347.

309   8. Lake F, Proudman S. Rheumatoid arthritis and lung disease: from mechanisms to a practical

310   approach. Semin Respir Crit Care Med. 2014; 35(2):222–238. doi: 10.1055/s-0034-1371542.

311   9. Zhu W, Kan X. Neural network cascade optimizes microRNA biomarker selection for

312   nasopharyngeal   cancer   prognosis.   PLoS   One.   2014;   9(10):e110537.   doi:

313   10.1371/journal.pone.0110537.

314 10. Li Z, Li Y, Sun L, Tang Y, Liu L, Zhu W. Artificial neural network cascade identifies multi-

315   P450 inhibitors in natural compounds. PeerJ. 2015; 3:e1524. doi: 10.7717/peerj.1524.

316 11. Hou S, Wang J, Li Z, Wang Y, Wang Y, Yang S, Xu J, Zhu W. Five-descriptor model to

317   predict the chromatographic sequence of natural compounds. J Sep Sci. 2016; 39(5):864–872.

318   doi: 10.1002/jssc. 201501016.

319 12. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3:32-35. doi: 10.1002/1097-

320   0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO; 2–3.

321 13. Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks.

322   Science. 2014; 344(6191):1492–1496. doi: 10.1126/science.1242072.

323 14. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data

324   integration and network visualization. Bioinformatics. 2011; 27(3):431–432. doi:

325   10.1093/bioinformatics/btq675.

326 15. Böhning D, Holling H, Patilea V. A limitation of the diagnostic-odds ratio in determining an

327   optimal cut-off value for a continuous diagnostic test. Stat Methods Med Res. 2011; 20:541–

328   550. doi: 10.1177/0962280210374532.

329 16. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single

330   indicator of test performance. J Clin Epidemiol. 2003; 56:1129–1135. doi: 10.1016/S0895-

331   4356(03)00177-X.

332 17. DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more

333   correlated   receiver   operating   characteristic   curves: a   nonparametric   approach.

334   Biometrics,1998; 44:837–845.

335   18. Olson AL, Swigris JJ, Sprunger DB, Fischer A, Fernandez-Perez ER, Solomon J, Murphy J,

336   Cohen M, Raghu G, Brown KK. Rheumatoid arthritis-interstitial lung disease-associated

337   mortality. Am J Respir Crit Care Med. 2011;183(3):372–8. doi: 10.1164/rccm.201004-

338   0622OC.

339   19. Zou YQ, Li YS, Ding XN, Ying ZH. The clinical significance of HRCT in evaluation of

340   patients with rheumatoid arthritis-associated interstitial lung disease: a report from China.

341   Rheumatol Int. 2012;32(3):669–673. doi: 10.1007/s00296-010-1665-1.

342   20. Richman NC, Yazdany J, Graf J, Chernitskiy V, Imboden JB. Extraarticular manifestations of

343   rheumatoid arthritis in a multiethnic cohort of predominantly Hispanic and Asian patients.

344   Medicine (Baltimore). 2013;92(2):92–97. doi: 10.1097/MD.0b013e318289ce01.

345   21. Wells AU, Denton CP. Interstitial lung disease in connective tissue disease--mechanisms and

346   management. Nat Rev Rheumatol. 2014;10(12):728–739. doi: 10.1038/nrrheum.2014.149.

347   22. Mori S. Management of rheumatoid arthritis patients with interstitial lung disease: safety of

348   biological antirheumatic drugs and assessment of pulmonary fibrosis. Clin Med Insights Circ

349   Respir Pulm Med. 2015;9(Suppl 1):41–49. doi: 10.4137/CCRPM.S23288.

350   23. Moua T, Zamora Martinez AC, Baqir M, Vassallo R, Limper AH, Ryu JH. Predictors of

351   diagnosis and survival in idiopathic pulmonary fibrosis and connective tissue disease-related

352   usual interstitial pneumonia. Respir Res. 2014;15:154. doi: 10.1186/s12931-014-0154-6.

353   24. Giles JT, Danoff SK, Sokolove J, Wagner CA, Winchester R, Pappas DA, Siegelman S,

354   Connors G, Robinson WH, Bathon JM. Association of fine specificity and repertoire expansion

355   of anticitrullinated peptide antibodies with rheumatoid arthritis associated interstitial lung

356   disease. Ann Rheum Dis. 2014;73(8):1487–1494. doi: 10.1136/annrheumdis-2012-203160.

357   25. Yilmazer B, Gümüştaş S, Coşan F, İnan N, Ensaroğlu F, Erbağ G, Yıldız F, Çefle A.High-

358   resolution computed tomography and rheumatoid arthritis: semi-quantitative evaluation of

359   lung damage and its correlation with clinical and functional abnormalities. Radiol Med.

360   2016;121(3):181–9. doi: 10.1007/s11547-015-0590-5.

361   26. Papanikolaou IC, Boki KA, Giamarellos-Bourboulis EJ, Kotsaki A, Kagouridis K,

362   Karagiannidis N, Polychronopoulos VS. Innate immunity alterations in idiopathic interstitial

363   pneumonias and rheumatoid arthritis-associated interstitial lung diseases. Immunol Lett.

364   2015;163(2):179–86. doi: 10.1016/j.imlet.2014.12.004.

365   27. Zamora-Legoff JA, Krause ML, Crowson CS, Ryu JH, Matteson EL.   Risk of serious infection

366   in patients with rheumatoid arthritis-associated interstitial lung disease. Clin Rheumatol.

367   2016;35(10):2585–9. doi: 10.1007/s10067-016-3357-z.

368   28. Middleton EA, Weyrich AS, Zimmerman GA. Platelets in pulmonary immune responses and

369   inflammatory lung diseases. Physiol Rev. 2016;96(4):1211–1259. doi:

370   10.1152/physrev.00038.2015.

371   29. Weyand CM, Goronzy JJ. Aging of the immune system. mechanisms and therapeutic targets.

372   Ann Am Thorac Soc. 2016;13(Supplement_5):S422–S428. doi: 10.1513/AnnalsATS.201602-

373   095AW.

374 30. Hu X, Cammann H, Meyer HA, Miller K, Jung K, Stephan C. Artificial neural networks and

375   prostate cancer--tools for diagnosis and management. Nat Rev Urol. 2013;10(3):174–82. doi:

376   10.1038/nrurol.2013.9.

377 31. Sheikhtaheri A, Sadoughi F, Hashemi Dehaghi Z. Developing and using expert systems and

378   neural networks in medicine: a review on benefits and challenges. J Med Syst. 2014;38(9):110.

379   doi: 10.1007/s10916-014-0110-5.

380 32. Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural

381   networks in neurosurgery. J Neurol Neurosurg Psychiatry. 2015;86(3):251–256. doi:

382   10.1136/jnnp-2014-307807.

383

384 **Figure legends**

385 **Figure 1.** ILD-associated CEIs. **A.** Distribution of the 32 CEIs in Youden index value. The

386 numbers on the columns indicate the number of CEIs. Four CEIs were observed to have a Youden

387 index of more than 0.15. **B.** ROC curves of the four ILD-associated CEIs age, eosinophil count

388 (EO), blood platelet count (PLT), and white blood cell count (WBC).

389 **Figure 2.** PPSNs of different edges. **A.** PPSN of 1500 edges; **B.** PPSN of 2000 edges; **C.** PPSN of

390 2500 edges; **D.** PPSN of 3000 edges. Green dots represent patients with healthy lungs and red dots

391 indicate patients complicated with ILD. Edge between two dots that represent patients means a

392 short Euclidean distance between the two patients in a 4-dimensional space, in which the four ILD-

393 associated CEIs were re-defined as coordinates. Although 533 patients were included in the

394   calculation of Euclidean distance, only the shortest distances could be visualized as edges in PPSN.

395   For each edge setting, the average number of neighbors was also calculated.

396   **Figure 3.** Optimization and evaluation of ANN II. **A.** Distribution of the Youden index values of

397   ANN models built by dividing all the patients with healthy lungs and those complicated with ILD

398   into 5, 10, 15, or 20 groups, respectively. The first ANN models of different division sizes 5D1,

399   10D1, 15D1, and 20D1 are highlighted as larger dots. The position of the dotted line is 0.387, and

400   values of Youden index were calculated using ANN I. **B.** Comparison of ANNs I and II in ROC

401   curves. A significant difference in AUCs was found ($P = 0.025$). **C.** The 10-fold cross-validation

402   (10FCV) result of ANN II.

403   **Figure 4.** Model evaluation using an independent patient set of 87 RA patients with PF and an

404   external validation set of 218 RA patients. **A.** Comparison of single ILD-associated CEIs, ANNs

405   I and II in identifying patients with PF. **B.** Scatter plots of ANN II outputs for patients with healthy

406   lungs and those complicated with ILD or PF. The position of the dotted line is 0.622, the optimal

407   ROC curve cut-off point of ANN II. **C.** The external validation result of ANN II.

408

409   **Supplementary information**

410   **Table S1.** Clinical examination results for the 620 patients included in this study.

411   **Table S2.** Distribution of the 32 clinical examination indicators used in this study.

412   **Table S3.** Clinical examination results for the patients included in the validation set (n = 218).

413   **Table S4.** ILD-associated CEIs.

414   **Figure S1.** Optimization of ANN I inputs. The optimal combination of four CEIs is highlighted

415   with a bigger dot indicating the highest Youden index. EO: eosinophil count; PLT: blood platelet

416   count; WBC: white blood cell count. For optimization of ANN I inputs, each of the eight ILD-

417   associated CEIs was selected alone as ANN I input and Youden index was calculated for the output

418   of each ANN model. After that, the CEI with the highest Youden index was obtained and the other

419   CEIs were added by the same rule until all of the eight CEIs were simultaneously used as inputs.

Table 1. Comparison of ILD-associated CEIs in identifying patients with high PF risk.

CEI: clinical examination indicator; AUC: area under the ROC curve; SE: sensitivity; SP: specificity; $SE_{SP = 0.8}$: sensitivity at fixed specificity = 0.8; DOR: diagnostic odd ratio. EO: eosinophil count; PLT: blood platelet count; WBC: white blood cell count NEUT: neutrophil count; U-SG: urine specific gravity; U-WBC: white blood cell count in urine; U-WBCH: white blood cell (high power field) in urine.

1   **Table 1.** Comparison of ILD-associated CEIs in identifying patients with high PF risk.

CEI	AUC	Youden index	SE	SP	$SE_{SP = 0.8}$	DOR
Age	0.710	0.301	0.74	0.56	0.47	3.63
EO	0.562	0.164	0.64	0.53	0.25	1.96
PLT	0.569	0.165	0.55	0.62	0.26	1.95
WBC	0.587	0.173	0.45	0.73	0.32	2.15
NEUT	0.577	0.133	0.85	0.28	0.26	2.26
U-SG	0.580	0.135	0.93	0.20	0.30	3.57
U-WBC	0.562	0.138	0.78	0.36	0.28	1.97
U-WBCH	0.561	0.123	0.78	0.34	0.28	1.85

2   CEI: clinical examination indicator; AUC: area under the ROC curve; SE: sensitivity; SP: specificity;

3    $SE_{SP = 0.8}$: sensitivity at fixed specificity = 0.8; DOR: diagnostic odd ratio. EO: eosinophil count; PLT:

4   blood platelet count; WBC: white blood cell count NEUT: neutrophil count; U-SG: urine specific

5   gravity; U-WBC: white blood cell count in urine; U-WBCH: white blood cell (high power field) in urine.

Table 2. Comparison of models in identifying patients with high PF risk.

CEI: clinical examination indicator; AUC: area under the ROC curve; SE: sensitivity; SP: specificity; $SE_{SP=0.8}$: sensitivity at fixed specificity = 0.8; DOR: diagnostic odd ratio; 5D1, 10D1, 15D1, and 20D1 were the first ANN models created by dividing all the patients with healthy lungs and those complicated with ILD into 5, 10, 15, or 20 groups, respectively.

1   **Table 2.** Comparison of models in identifying patients with high PF risk.

Model	AUC	Youden index	SE	SP	$SE_{SP = 0.8}$	DOR
ANN I	0.734	0.387	0.772	0.615	0.473	5.41
5D1	0.767	0.436	0.791	0.645	0.591	6.88
10D1	0.758	0.411	0.731	0.681	0.563	5.80
15D1	0.746	0.407	0.780	0.627	0.489	5.96
20D1	0.736	0.403	0.805	0.598	0.500	6.14

2   CEI: clinical examination indicator; AUC: area under the ROC curve; SE: sensitivity; SP: specificity;

3    $SE_{SP = 0.8}$: sensitivity at fixed specificity = 0.8; DOR: diagnostic odd ratio; 5D1, 10D1, 15D1, and 20D1

4   were the first ANN models created by dividing all the patients with healthy lungs and those complicated

5   with ILD into 5, 10, 15, or 20 groups, respectively.

Figure 1

Figure 1. ILD-associated CEIs.

A. Distribution of the 32 CEIs in Youden index value. The numbers on the columns indicate the number of CEIs. Four CEIs were observed to have a Youden index of more than 0.15. **B.** ROC curves of the four ILD-associated CEIs age, eosinophil count (EO), blood platelet count (PLT), and white blood cell count (WBC).
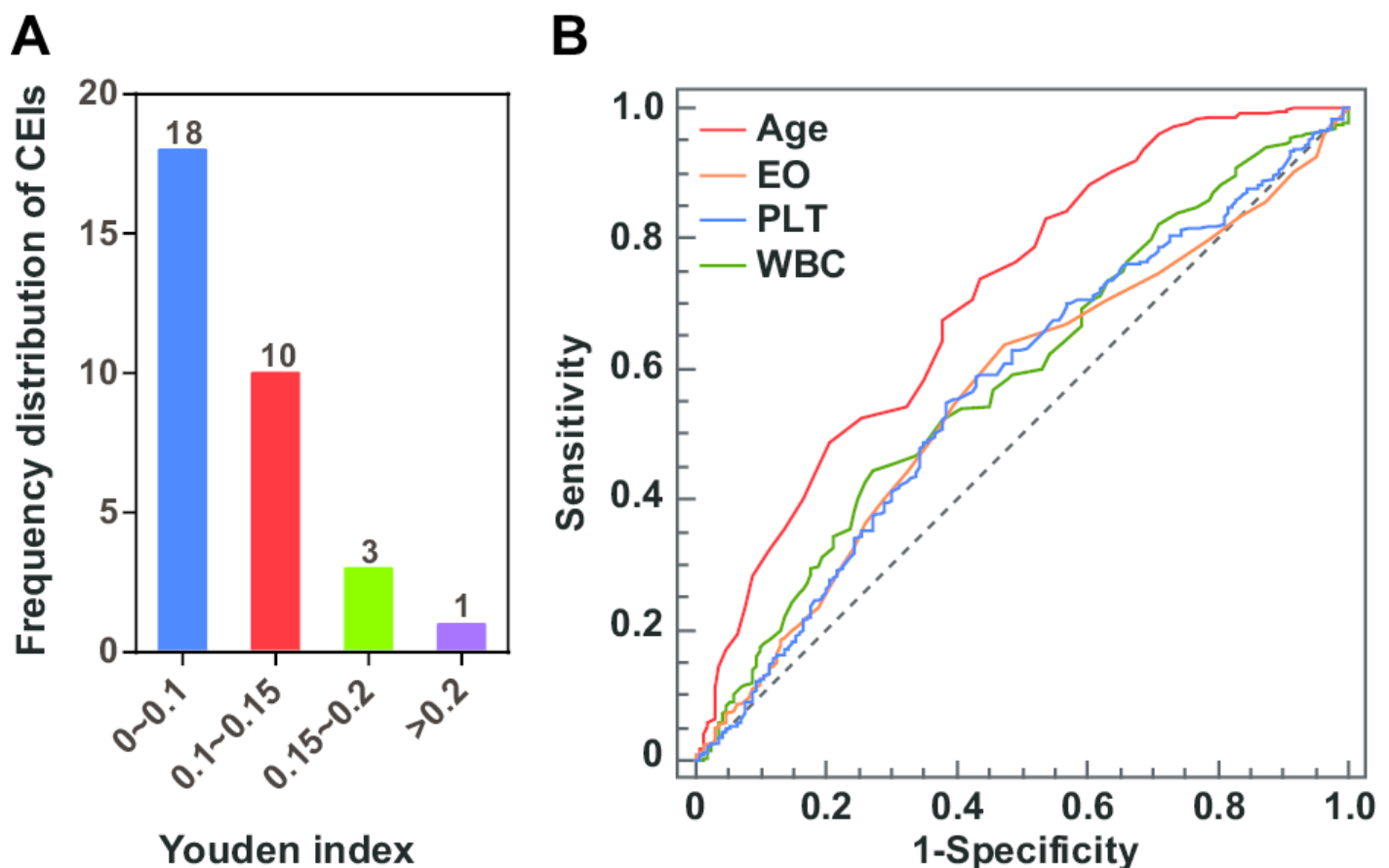
Figure 2

Figure 2. PPSNs of different edges.

A. PPSN of 1500 edges; **B.** PPSN of 2000 edges; **C.** PPSN of 2500 edges; **D.** PPSN of 3000 edges. Green dots represent patients with healthy lungs and red dots indicate patients complicated with ILD. Edge between two dots that represent patients means a short Euclidean distance between the two patients in a 4-dimensional space, in which the four ILD-associated CEIs were re-defined as coordinates. Although 533 patients were included in the calculation of Euclidean distance, only the shortest distances could be visualized as edges in PPSN. For each edge setting, the average number of neighbors was also calculated.

A

B

Average number of neighbors = 5.629
(1500 edges)

Average number of neighbors = 7.505
(2000 edges)

C

D

Average number of neighbors = 9.381
(2500 edges)

Average number of neighbors = 11.257
(3000 edges)
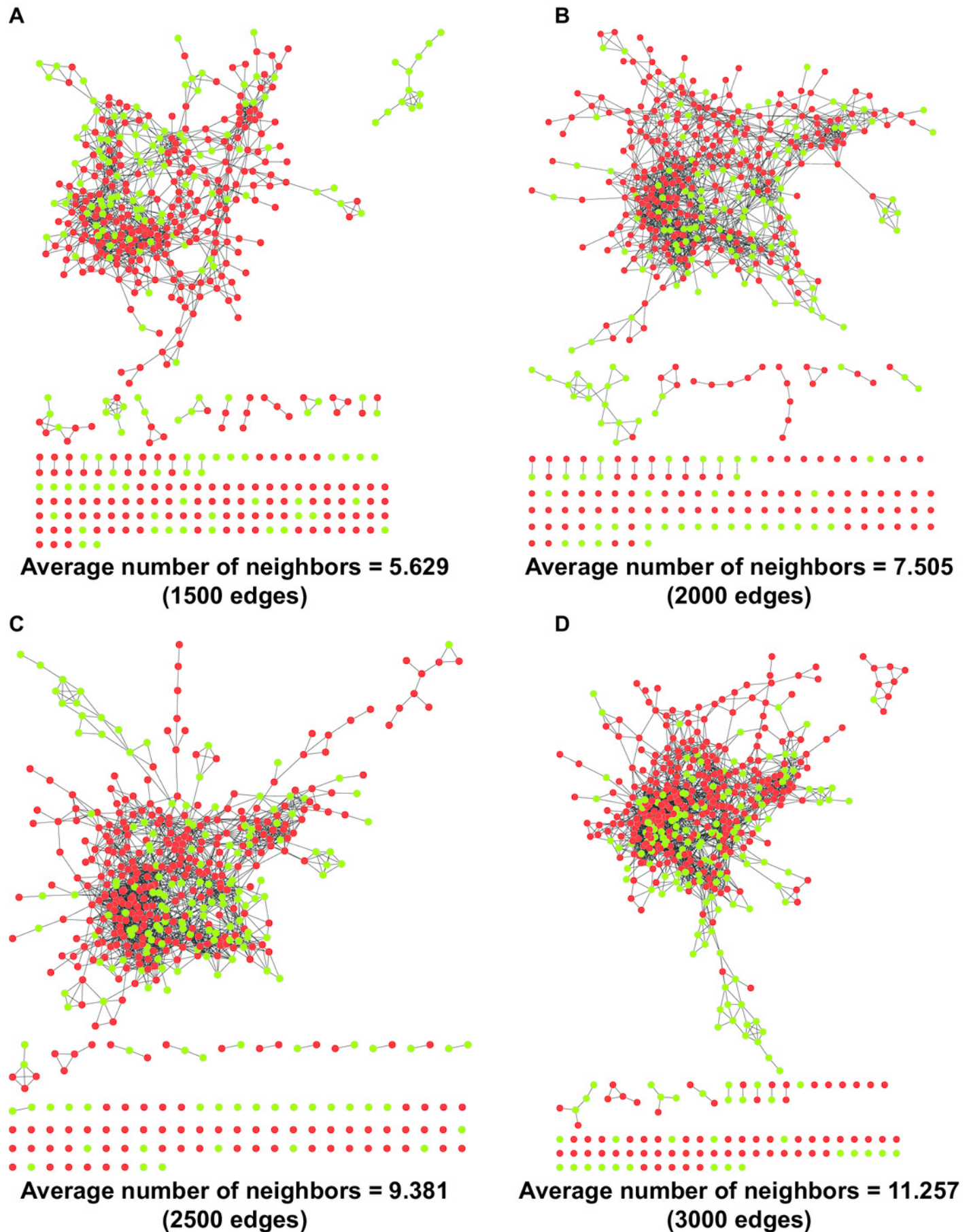
Figure 3

Figure 3. Optimization and evaluation of ANN II.

A. Distribution of the Youden index values of ANN models built by dividing all the patients with healthy lungs and those complicated with ILD into 5, 10, 15, or 20 groups, respectively. The first ANN models of different division sizes 5D1, 10D1, 15D1, and 20D1 are highlighted as larger dots. The position of the dotted line is 0.387, and values of Youden index were calculated using ANN I. **B.** Comparison of ANNs I and II in ROC curves. A significant difference in AUCs was found (*P* = 0.025). **C.** The 10-fold cross-validation (10FCV) result of ANN II.
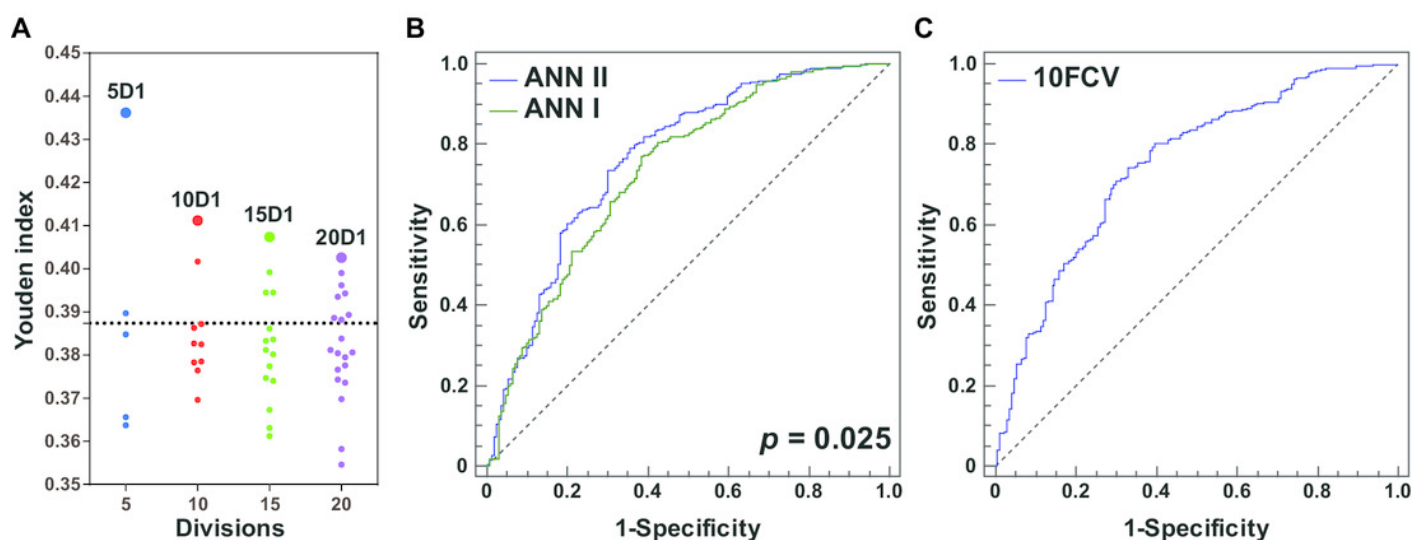
Figure 4

Figure 4. Model evaluation using an independent patient set of 87 RA patients with PF and an external validation set of 218 RA patients.

A. Comparison of single ILD-associated CEIs, ANNs I and II in identifying patients with PF. **B.** Scatter plots of ANN II outputs for patients with healthy lungs and those complicated with ILD or PF. The position of the dotted line is 0.622, the optimal ROC curve cut-off point of ANN II. **C.** The external validation result of ANN II.