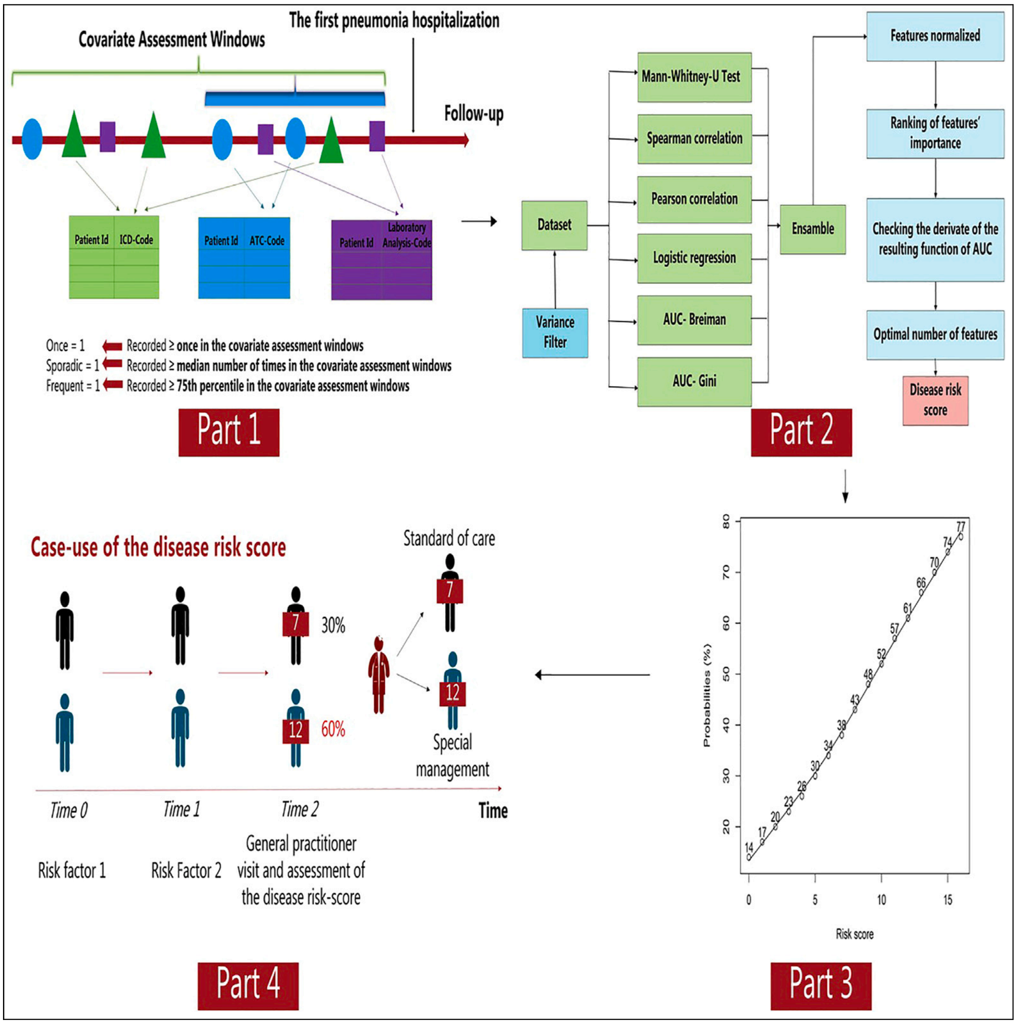
iScience

CellPress
OPEN ACCESS

Article

AI-based disease risk score for community-acquired pneumonia hospitalization


Covariate Assessment Windows
The first pneumonia hospitalization
Follow-up
Patient Id
ICD-Code
Patient Id
ATC-Code
Patient Id
Laboratory Analysis-Code
Once = 1 Recorded ≥ once in the covariate assessment windows
Sporadic = 1 Recorded ≥ median number of times in the covariate assessment windows
Frequent = 1 Recorded ≥ 75th percentile in the covariate assessment windows
Part 1
Dataset
Variance Filter
Mann-Whitney-U Test
Spearman correlation
Pearson correlation
Logistic regression
AUC- Breiman
AUC- Gini
Ensamble
Features normalized
Ranking of features' importance
Checking the derivate of the resulting function of AUC
Optimal number of features
Disease risk score
Part 2
Probabilities (%)
20
30
40
50
60
70
80
14
17
20
23
26
30
34
38
43
48
52
57
61
66
70
74
77
0
5
10
15
Risk score
Part 3
Case-use of the disease risk score
Standard of care
7
30%
12
60%
7
12
Special management
Time 0
Time 1
Time 2
Time
Risk factor 1
Risk Factor 2
General practitioner visit and assessment of the disease risk-score
Part 4

Saeed Shakibfar, Morten Andersen, Maurizio Sessa

maurizio.sessa@sund.ku.dk

Highlights

We developed a disease risk score for CAP hospitalization based on real-world data

This tool can help to identify individuals requiring specific clinical management

Further studies are needed to investigate the optimal use of our disease risk

Shakibfar et al., iScience *26*, 107027
July 21, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.107027

iScience

CellPress OPEN ACCESS

Article

AI-based disease risk score for community-acquired pneumonia hospitalization

Saeed Shakibfar,[1] Morten Andersen,[1] and Maurizio Sessa[1,2,*]

SUMMARY

Community-acquired pneumonia (CAP) is an acute infection involving the parenchyma of the lungs, which is acquired outside of the hospital. Population-wide real-world data and artificial intelligence (AI) were used to develop a disease risk score for CAP hospitalization among older individuals. The source population included residents in Denmark aged 65 years or older in the period January 1, 1996, to July 30, 2018. 137344 individuals were hospitalized for pneumonia during the study period for which, 5 controls were matched leading to a study population of 620908 individuals. The disease risk had an average accuracy of 0.79 based on 5-fold cross-validation in predicting CAP hospitalization. The disease risk score can be useful in clinical practice to identify individuals at higher risk of CAP hospitalization and intervene to minimize their risk of being hospitalized for CAP.

INTRODUCTION

Community-acquired pneumonia (CAP) is an acute infection involving the parenchyma of the lungs, which is acquired outside of the hospital.[1] CAP is a leading cause of hospitalizations, death, and poor quality of life. Its clinical management is associated with a high economic burden for healthcare systems, and long-term sequelae in all age groups worldwide.[1–3]

Recent incidence estimates of CAP were 1.2 and 2.4 cases per 1,000 people in Europe and the USA, respectively, with the highest incidence in individuals aged 65 years or older.[4,5] A recent study has estimated 489 million cases of CAP with more than 2.5 million deaths globally.[1] In the healthcare management of CAP in older individuals, it is crucial to act immediately and prevent the progression of CAP to a more severe disease considering the high incidence of hospitalization and/or mortality of severe CAP in older individuals.[1–5]

Several studies have investigated risk factors associated with hospitalization for CAP in older individuals. However, these studies have produced inconclusive results because of small sample sizes and lack of representativeness that preclude the generalizability of results to other settings.[6–10] Other limitations identified in previously conducted studies include inconsistencies in the definition of CAP if compared to those provided by clinical guidelines,[11] and the use of selected populations with the consequent risk of selection bias.[12] Additionally, the majority of studies have a focus on the use of a few potential predictors, mostly selected in a clinical setting excluding other possible candidate predictors from biomarkers, treatments, comorbidities, and sociodemographic characteristics.[6–12]

In this regard, we conducted a systematic screening of the scientific literature highlighting that, to the best of our knowledge, there are no studies that have evaluated risk factors for CAP by applying AI techniques in population-wide real-world data (Appendices 1 and 2). From the systematic reviews, we identified only one study that has investigated predictors of CAP from sociodemographic and healthcare real-world data; however, this study has been performed using primary care data and did not focus on hospitalization as an outcome.[13]

Therefore, in this study, we took the opportunity of applying AI-based feature selection techniques to population-wide real-world data from Danish healthcare and administrative databases to identify predictors of CAP hospitalization among older individuals and use this information to develop a disease risk score. We aim at developing a disease risk score because it may represent a precious tool for general practice to identify among older individuals those considering the high risk of CAP hospitalization may require specific

[1]Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark

[2]Lead contact

*Correspondence: maurizio.sessa@sund.ku.dk

https://doi.org/10.1016/j.isci.2023.107027

clinical management to mitigate such a risk. This is a key priority in public health as CAP is a significant public health burden, accounting for a significant proportion of hospitalizations and healthcare costs worldwide.[1–5] Additionally, a newly developed disease risk score for CAP hospitalization can help to identify the risk factors for severe disease and improve our ability to predict severe cases, improve their management, and, therefore, reduce the morbidity and mortality associated with CAP.

RESULTS

Sociodemographic characteristics

In total, 137344 individuals were hospitalized for CAP during the study period for which, 5 individuals were matched leading to a study population of 620908 individuals. An overview of sociodemographic characteristics of the study population is provided in Table S1. The mean age (SD) of individuals that were hospitalized for CAP was 77.2 (9.4) years while for individuals not hospitalized for CAP it was 77.0 (9.7) years with a proportion of women of 48.4% and 50.9%, respectively. Sociodemographic characteristics of individuals hospitalized for CAP were comparable with those observed in other Danish studies investigating trends of pneumonia hospitalization in the Danish population.[14]

Performance of the classification models

In total, 735588 features were generated in the data management processes of which 321 were filtered based on variance. The average AUC of the benchmarked classification models built using the iterative inclusion of variables sorted by EFS estimated variable importance in Figure 1. The derivate of the resulting function of average accuracy based on a 5-fold cross-validation approach identified 10 as the optimal number of features to be used in the classification models (Figure 1). The 10 most important predictors for CAP are shown in Figure 2. After clinical validation, phenoxymethylpenicillin, roxithromycin, terbutaline, furosemide, paracetamol, and potassium chloride were considered non-biologically plausible predictors of CAP. Analogously, ICD_{10} codes for hospital admission ''Suspected disease or condition'' was considered not sufficiently detailed predictor and, therefore, considered non-biologically plausible.

The top 10 most important predictors for CAP were used to compare the performance of the classification models of which the average AUC, sensitivity, specificity, and accuracy based on a 5-fold cross-validation approach and a balanced confusion matrix are shown in Table 1 for resampling data with 1:1 matching. Logistic regression was chosen as the final model based on its best performance when considering all together AUC, sensitivity, specificity, and accuracy.

Disease risk score

The disease risk score included weights in a range between 1 and 4 (Table 2), which resulted in a disease risk score ranging between 0 and 16. The average accuracy for predicting CAP based on a 5-fold cross-validation using only the disease risk score and a logistic regression model was 0.79. The predicted probability of CAP for each disease risk score is presented in Figure 3. Observed versus predicted probabilities estimated using the disease risk score with (1) all the predictors, (2) only the disease risk score, and (3) a calibrated model with only the disease risk score are presented in Figures S1, S2, andS3. Models' performance in terms of observed versus predicted probabilities was assessed using the mean absolute error and the model using only the disease risk score was preferred. The predicted probability of CAP for each disease risk score with and without calibration is presented in Figure S4. The median time to the event and the interquartile range was 49 days (7–168), 44 days (16–115), 41 days (15–87), 46 days (17–101), 40 days (15–105), 52 days (7–179), 55 days (18–150), 1.4 years (1.4–3), 0.8 years (0.1.-2.6), 2.2 years (0.3–5.9) for phenoxymethylpenicillin, prednisolone, furosemide, potassium chloride, paracetamol, roxithromycin, terbutaline, chronic obstructive pulmonary disease, chronic obstructive pulmonary disease—acute exacerbation, and ''suspected disease or condition,'' respectively (Figure 4). The network analysis of predictors is provided in Figure 5. For non-biologically plausible predictors, the network analysis and their inter-relationship are provided in Figures S5 and S6. Based on the predictors used to develop the disease risk score, we developed a questionnaire that can be used in general practice to compute the disease risk score for each patient (Table 3).

DISCUSSION

This study has investigated the use of AI-based feature selection techniques to population-wide real-world data from Danish healthcare and administrative databases to identify predictors of CAP hospitalization among older individuals and use this information to develop a disease risk score.

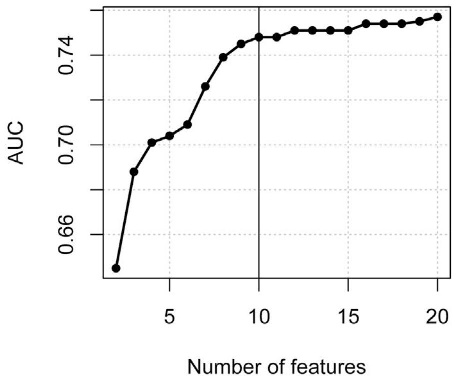


Figure 1. Optimal number of predictors for CAP based on EFS estimated variable importance

This is the first study proposing a disease risk score for CAP hospitalization, which is based on preemptive predictors identified from population-based real-world data. The disease risk had an average accuracy of 0.79 based on 5-fold cross-validation in predicting CAP hospitalization.

If we look at our identified predictors used to develop the disease risk score, they capture most of the aforementioned risk factors directly or indirectly and, additionally, they provide weight to their importance in the context of the risk of CAP hospitalization and they provide relatively easy-to-detect proxies of these conditions with a high predictive power for CAP hospitalization.

Our newly developed disease risk score can have significant implications from both practical and academic perspectives. From a real-world standpoint, the disease risk score may have a big impact on patient care and outcomes. The disease risk score may enable clinicians to start early interventions or preventative measures to improve outcomes by precisely identifying patients who are at risk of severe CAP. This can involve making lifestyle changes, taking medications, or getting checked on or monitored more frequently leading to a more personalized approach to healthcare. From an academic perspective, the disease risk score may also help advance medical research and lead to new standards for preventative measures of severe CAP and it can help to prioritize risk factors for severe CAP. Additionally, our research can eventually stimulate research in related fields, including epidemiology, leading to new insights and interventions for disease prevention and management. In total, 321 predictors were identified, however, 10 contributed more extensively to CAP hospitalization prediction. Redeemed prescription of phenoxymethylpenicillin, roxithromycin, terbutaline, furosemide, paracetamol, and potassium chloride were considered non-biologically plausible predictors of CAP but rather proxies of the events that were more likely related to the occurrence of CAP. In particular, we believe these medicines are detected as predictors for CAP hospitalization due to their approved therapeutic indications for clinical conditions which are known risk factors for CAP hospitalization.

In Danish clinical guidelines, phenoxymethylpenicillin is an antibiotic recommended for the treatment of pneumonia in primary care with a posological scheme of 660 mg (1 million IU) x 4 for 5 days.[15] Therefore, it is not surprising to identify this antibiotic as a potential predictor of pneumonia hospitalization. Our epidemiological and clinical reasoning for this predictor is that individuals may have received a diagnosis of pneumonia in primary care due to pulmonary infection, started the treatment with phenoxymethylpenicillin and, as the treatment was not successful, the individuals were furtherly referred to the hospital for secondary care and undergoing hospitalization. A similar consideration applies to roxithromycin, as this antibiotic is recommended 150 mg x 2–3 days for the treatment of pneumonia in primary care in individuals with an allergy to penicillin.[15] Of note, it cannot be excluded that roxitromycin was used as first line treatment for mycoplasma pneumonia.

Terbutaline is prescribed as a rescue treatment for sudden breathlessness or wheezing in people with asthma or chronic obstructive pulmonary disease, which are known risk factors for pneumonia and that were identified as predictors in this study.[16] It is also plausible to believe that terbutaline has been used as a bronchodilator for alleviating acute bronchoconstriction associated with pneumonia.

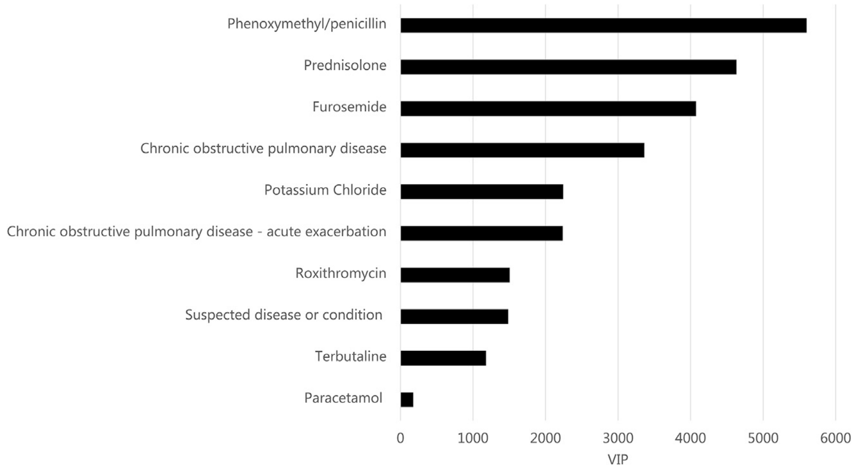


Figure 2. Top ten predictors for CAP

VIP = variable importance.

Furosemide is a potent loop diuretic that is used for edema secondary to various clinical conditions, such as congestive heart failure exacerbation, liver failure, renal failure, and high blood pressure. It is not surprising to find this drug as a potential predictor of CAP hospitalization as cardiovascular disorders are known risk factors for CAP.[17] Another plausible explanation for finding furosemide as a predictor of CAP hospitalization is that the exacerbation of the underlying cardiovascular condition of individuals that developed pneumonia leads to the medical need of prescribing furosemide.[18]

Potassium chloride is a potassium salt used to treat hypokalemia. Electrolyte abnormalities in pneumonia, including hypokalemia (serum potassium less than or equal to 3.5 mEq/L), have been extensively described in the scientific literature among individuals hospitalized for CAP.[19]

Paracetamol is sold as over-the-counter in Denmark. However, it is possible to obtain a medical prescription for reimbursement in case of chronic pain. In Danish administrative registers, over-the-counter medications are not traceable, therefore, these individuals that redeemed prescriptions of paracetamol were mostly individuals exposed to chronic pain. The underlying causes and etiology of chronic pain and its pharmacological treatment are associated with immunosuppression, which is a known risk factor for CAP.[20]

Regarding prednisolone as a predictor for CAP hospitalization, we believe that prednisolone may have been used before CAP hospitalization because several studies suggested that its use reduces mortality and morbidity in adults with severe CAP.[21] Prednisolone is also used in the pharmacological treatment of known risk factors of CAP, such as chronic obstructive pulmonary disease and asthma.[22]

Several predictors did not show biological plausibility and, therefore, we have performed a network analysis. Among the reasons for not observing biological plausibility, there are reverse causality and/or confounding. As shown in our network analysis, the aforementioned predictors are strongly inter-related and such inter-relationships have been addressed with a biological rationale in a recent narrative review by Aliberti and colleagues.[2] Of note, our analytical approach was able to identify in a data-driven way all known risk factors for CAP hospitalization mentioned in the aforementioned review.[23] Therefore, it is not surprising to observe that when predictors of CAP hospitalization were combined in a disease risk score

Table 1. Models performance

Models	AUC	Sensitivity	Specificity	Accuracy
Logistic Regression	0.75	0.68	0.70	0.69
Random Forest	0.74	0.66	0.72	0.69
Random Forest Ranger	0.75	0.65	0.72	0.69
Random Partitioning	0.69	0.69	0.66	0.67

Table 2. Weights applied to the predictors selected based on the optimal number of predictors

Predictors	Clinical condition to evaluate	Weights
Phenoxymethyl/penicillin	Pulmonary infection requiring phenoxymethyl/ penicillin	4
Prednisolone	Recent exposure to prednisolone	3
Furosemide	Recent exposure to furosemide	3
Chronic obstructive pulmonary disease	Chronic obstructive pulmonary disease	2
Chronic obstructive pulmonary disease - acute exacerbation	Chronic obstructive pulmonary disease	2
Potassium Chloride	Hypokalemia	2
Paracetamol	Chronic pain requiring paracetamol as pain killer	1
Suspected disease or condition	Suspicious of severe pneumonia based on sign and symptoms	1
Roxithromycin	Pulmonary infection requiring roxithromycin	1
Terbutaline	Need for bronchodilation	1

which, accounts for their inter-relationship, the prediction accuracy resulted of our classification model outperformed the classification model with the individual predictors reaching an average accuracy for predicting CAP hospitalization of 0.79 or rather that on average 79% (8 out of 10) of patients were correctly predicted to be/not be hospitalized for CAP based on their disease risk score. These results suggest that independently of the etiology for CAP hospitalization, the identified predictors were found to have a strong predictive power for this clinical event. In this regard, it is interesting to observe linearity between the predicted probability of CAP hospitalization and the disease risk score which may facilitate its future use in clinical practice. Finally, it has to be emphasized that the identified predictors of CAP hospitalization were aligned with those previously mentioned in the scientific literature.

We believe that the disease risk score can be useful in clinical practice as it relied on the identification of a few, clinically validated predictors to identify individuals that with a higher probability will be hospitalized for CAP. A great advantage of our disease risk score is that identified predictors had a time-to-event or rather the time from which such predictors were recorded in administrative databases and the time from which the individuals were hospitalized for CAP of at least 41 days. This let us believe that from the

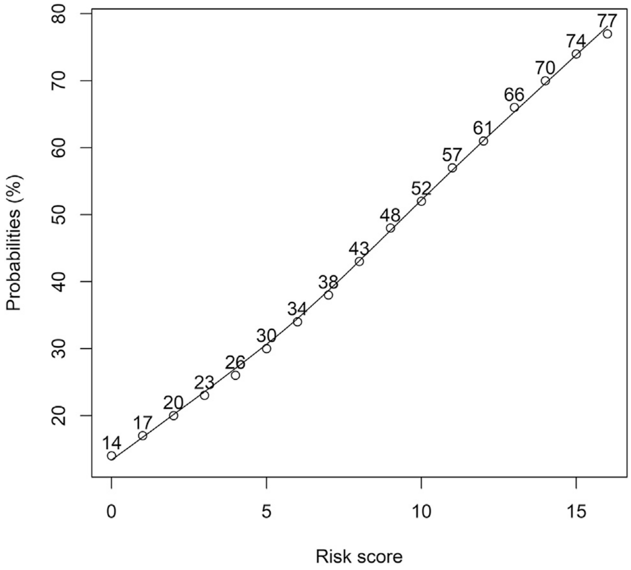


Figure 3. Predicted probabilities of developing the outcome by disease risk score values

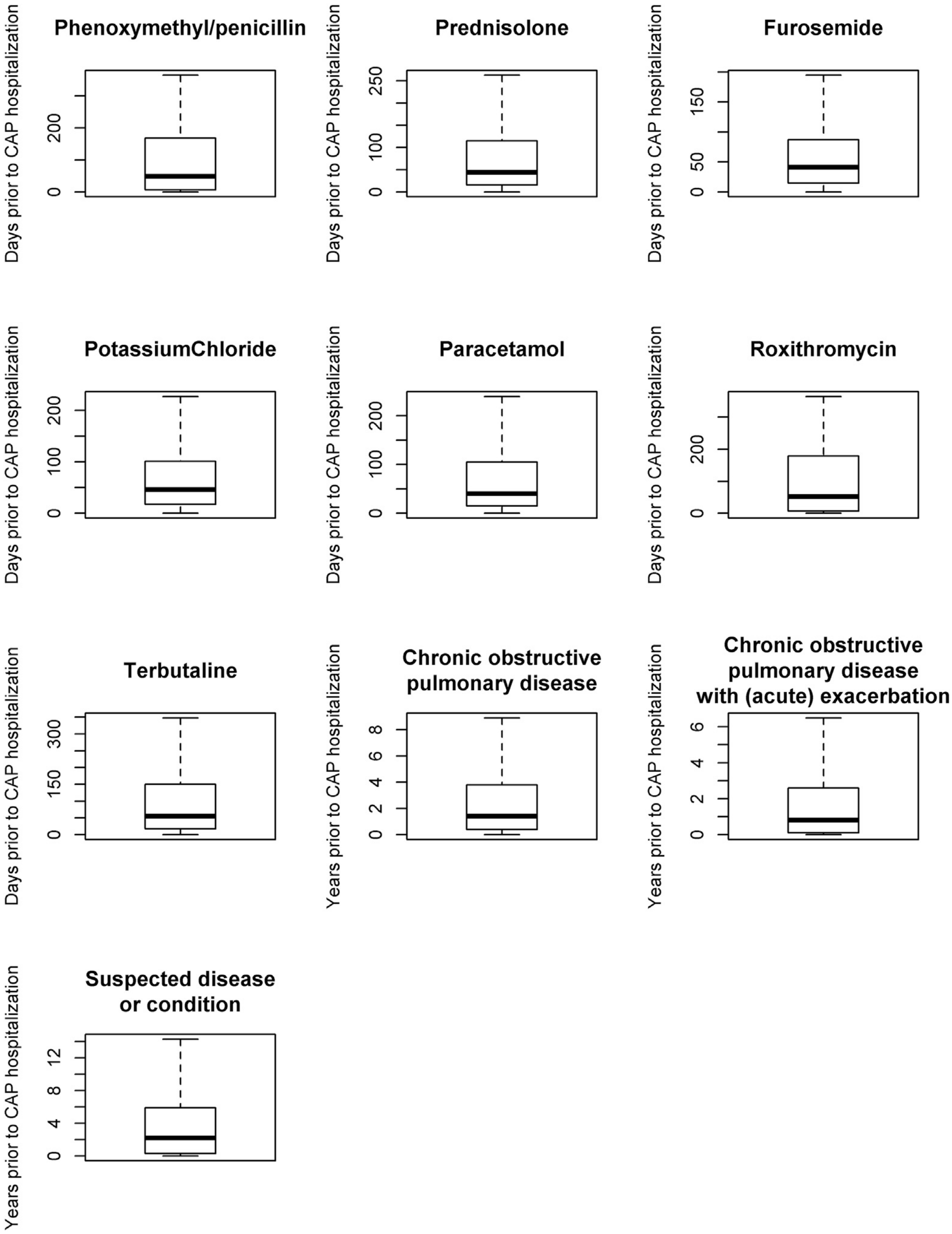


Figure 4. Boxplots of time to event for the top 10 predictors

detection of the predictor to the occurrence of CAP hospitalization there may be sufficient time to perform a clinical intervention to mitigate such risks.

We have envisioned different disease risk scores to be used in various clinical settings, including primary care, emergency rooms, and specialist clinics, and for different purposes. In primary care settings, such as in general practice, the disease risk scores can be used for routine screenings. In emergency room settings, the disease risk score may be less commonly used, as healthcare providers often focused on acute care and triage. However, it cannot be excluded that the disease risk score may be used to aid in the diagnosis and guide management. Regarding the possible applications of the disease risk score, we believe that it can be used for screening to identify patients who are at a higher risk of CAP allowing clinicians to initiate preventative measures, such as

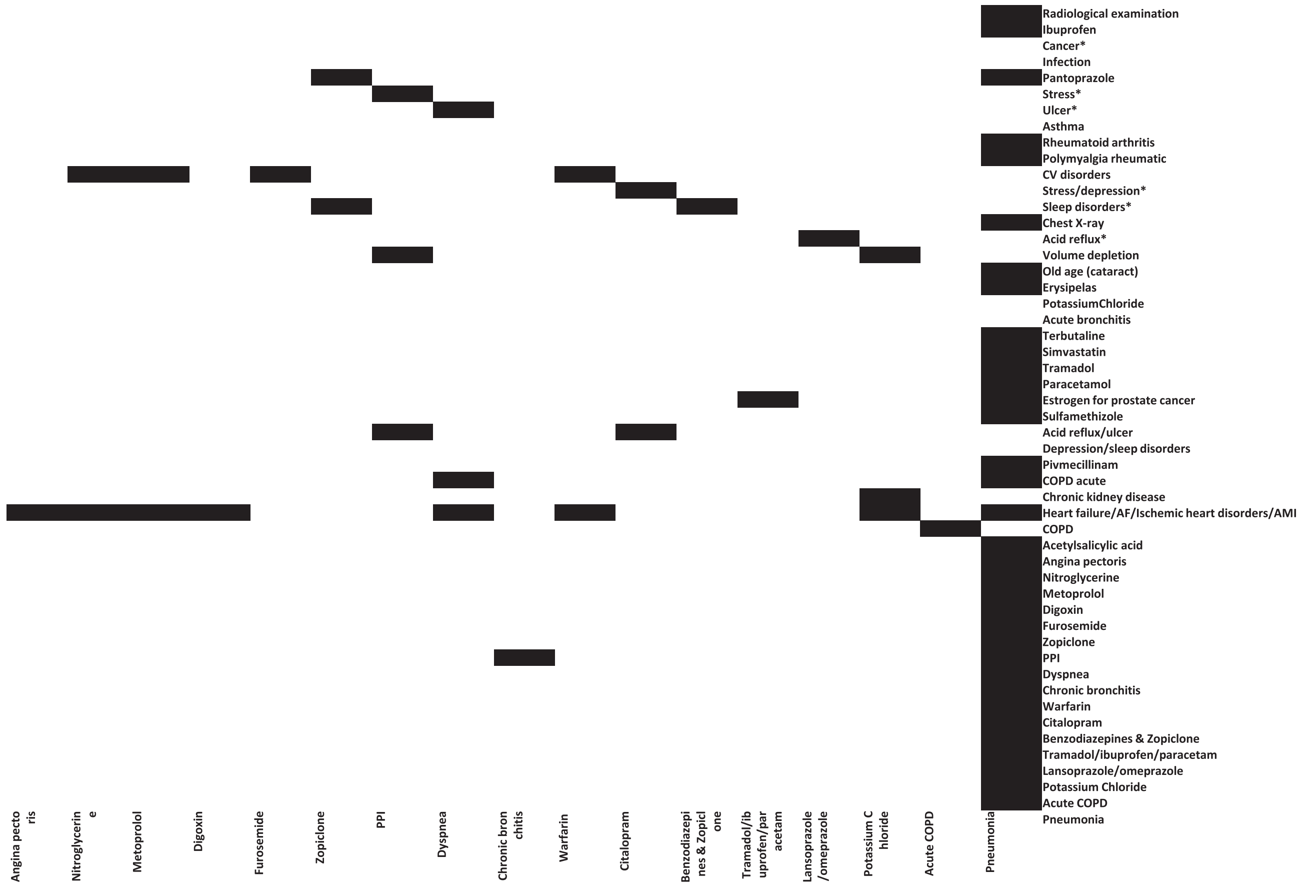


Figure 5. Network analysis of predictors of predictors

Predictors marked with * have been used hypothesized based on literature screening.

lifestyle modifications or medication, to reduce the risk of future complications. Additionally, it can be used for diagnosis as it can also help clinicians confirm or rule out a diagnosis of severe CAP. Our disease risk score can also use to guide treatment decisions as it can help clinicians determine whether to pursue more aggressive treatment options or to just monitor the patient's condition. Finally, it can also be used for deciding follow-up visits and deciding how to monitor patients over time as it can help identify patients who are at a higher risk of developing severe CAP and, therefore, requiring more frequent screenings.

It should be emphasized that our disease risk scores should be used along with other clinical information and it should take into account patient preferences. Clinical judgment and patient input should always be taken into account when making treatment decisions.

In conclusion, this study identified 10 crucial predictors for CAP hospitalization which, when combined into a disease risk score had accuracy in predicting CAP hospitalization of 0.79. These results suggest that independently of the etiology for CAP hospitalization, the disease risk score was found to have a strong predictive power for this clinical event.

Limitations of the study

We missed information on dates of death in 2017. However, considering that a small fraction of cases enrolled in 2017, we believe the potential bias introduced in the selection of controls to be limited. The study is based on population-wide real-world data from Danish healthcare and administrative databases, which may limit the generalizability of the results to other countries or healthcare systems, which differ extensively from the Danish system. Another limitation of this study is that clinical and epidemiological reasoning was conducted only by two individuals or rather a pharmacist (MS) and a medical doctor (MA).

Table 3. Survey to compute the disease risk score in general practice

Rows	Predictors	Clinical condition to evaluate	Yes	No
1	Phenoxymethylpenicillin	Did the patient during the last 2 months require treatment with phenoxymethylpenicillin for pulmonary infection?	+4	+0
2	Prednisolone	Did the patient during the last 2 months require treatment with Prednisolone for pulmonary infection?	+3	+0
3	Furosemide	Did the patient during the last 2 months require treatment with furosemide for cardiovascular disorders?	+3	+0
4	Chronic obstructive pulmonary disease	Is the patient diagnosed with a chronic obstructive pulmonary disease?	+2	+0
5	Chronic obstructive pulmonary disease - acute exacerbation	Did the patient during the last year have a chronic obstructive pulmonary disease exacerbation?	+2	+0
6	Potassium Chloride	Did the patient during the last 2 months require treatment with potassium chloride for hypokalemia?	+2	+0
7	Paracetamol	Did the patient during the last year require recurrent treatment with paracetamol for chronic diseases?	+1	+0
8	Suspected disease or condition	Did the patient have recent hospital access for an unspecified condition of the lungs?	+1	+0
9	Roxithromycin	Did the patient during the last 2 months require treatment with roxithromycin for pulmonary infection?	+1	+0
10	Terbutaline	Did the patient during the last 2 months require treatment with terbutaline for bronchodilation?	+1	+0

Instructions: To obtain the disease risk score you should sum the value in the column “Yes” for each row where you answered yes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Material availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Study design and methods
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data analysis
  - Feature filtering and prioritization
  - Systematic review – Part 1
  - Systematic review – Part 2

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107027.

ACKNOWLEDGMENTS

This work has been performed as part of the Nordic Coherence project, project no. 105670 funded by NordForsk under the Nordic Council of Ministers and the EU-COVID-19 project, project no. 312707 funded by Norwegian Research Council and by a grant from the Novo Nordisk Foundation to the University of Copenhagen (NNF15SA0018404).

AUTHOR CONTRIBUTIONS

S.S. wrote the original draft, conceptualized the study, and performed the data analysis. M.A. supervised the study, acquired the funding, and revised the manuscript. M.S. planned the study, performed the data analysis, wrote the original draft, and supervised the study.

DECLARATION OF INTERESTS

None.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: January 24, 2023
Revised: April 3, 2023
Accepted: May 30, 2023
Published: June 7, 2023

REFERENCES

1. Torres, A., Cilloniz, C., Niederman, M.S., Menéndez, R., Chalmers, J.D., Wunderink, R.G., and van der Poll, T. (2021). Pneumonia. Nat. Rev. Dis. Primers *7*, 25. https://doi.org/10.1038/s41572-021-00259-0.

2. File, T.M., Jr. (2003). Community-acquired pneumonia. Lancet *362*, 1991–2001.

3. Ticona, J.H., Zaccone, V.M., and McFarlane, I.M. (2021). Community-acquired pneumonia: a focused review. Am. J. Med. Case Rep. *9*, 45.

4. Torres, A., Peetermans, W.E., Viegi, G., and Blasi, F. (2013). Risk factors for community-acquired pneumonia in adults in Europe: a literature review. Thorax *68*, 1057–1065.

5. Jain, S., Self, W.H., Wunderink, R.G., Fakhran, S., Balk, R., Bramley, A.M., Reed, C., Grijalva, C.G., Anderson, E.J., Courtney, D.M., et al. (2015). Community-acquired pneumonia requiring hospitalization among US adults. N. Engl. J. Med. *373*, 415–427.

6. Capelastegui, A., España Yandiola, P.P., Quintana, J.M., Bilbao, A., Diez, R., Pascual, S., Pulido, E., and Egurrola, M. (2009). Predictors of short-term rehospitalization following discharge of patients hospitalized with community-acquired pneumonia. Chest *136*, 1079–1085.

7. Jasti, H., Mortensen, E.M., Obrosky, D.S., Kapoor, W.N., and Fine, M.J. (2008). Causes and risk factors for rehospitalization of patients hospitalized with community-acquired pneumonia. Clin. Infect. Dis. *46*, 550–556.

8. Shorr, A.F., Zilberberg, M.D., Reichley, R., Kan, J., Hoban, A., Hoffman, J., Micek, S.T., and Kollef, M.H. (2013). Readmission following hospitalization for pneumonia: the impact of pneumonia type and its implication for hospitals. Clin. Infect. Dis. *57*, 362–367.

9. Mather, J.F., Fortunato, G.J., Ash, J.L., Davis, M.J., and Kumar, A. (2014). Prediction of pneumonia 30-day readmissions: a single-center attempt to increase model performance. Respir. Care *59*, 199–208.

10. Adamuz, J., Viasus, D., Campreciós-Rodríguez, P., Cañavate-Jurado, O., Jiménez-Martínez, E., Isla, P., García-Vidal, C., and Carratalà, J. (2011). A prospective cohort study of healthcare visits and rehospitalizations after discharge of patients with community-acquired pneumonia. Respirology *16*, 1119–1126.

11. Klausen, H.H., Petersen, J., Lindhardt, T., Bandholm, T., Hendriksen, C., Kehlet, H., Vestbo, J., and Andersen, O. (2012). Outcomes in elderly Danish citizens admitted with community-acquired pneumonia. Regional differencties, in a public healthcare system. Respir. Med. *106*, 1778–1787.

12. Tang, V.L., Halm, E.A., Fine, M.J., Johnson, C.S., Anzueto, A., and Mortensen, E.M. (2014). Predictors of rehospitalization after admission for pneumonia in the veterans affairs healthcare system. J. Hosp. Med. *9*, 379–383.

13. Sun, X., Douiri, A., and Gulliford, M. (2022). Applying machine learning algorithms to electronic health records predicted pneumonia after respiratory tract infection. J. Clin. Epidemiol. *145*, 154–163.

14. Søgaard, M., Nielsen, R.B., Schønheyder, H.C., Nørgaard, M., and Thomsen, R.W. (2014). Nationwide trends in pneumonia hospitalization rates and mortality, Denmark 1997–2011. Respir. Med. *108*, 1214–1222.

15. Fagudvalget for hensigtsmassig anvendelse af antibiotika i primar- og sekundarsektoren under Rådet for Anvendelse af Dyr Sygehusmedicin. Baggrundsnotat for hensigtsmæssig anvendelse af antibiotika ved nedre luftvejsinfektioner i almen praksis og på hospital. (2016). RADS 1.1

16. Cavallazzi, R., and Ramirez, J. (2020). Community-acquired pneumonia in chronic obstructive pulmonary disease. Curr. Opin. Infect. Dis. *33*, 173–181.

17. Restrepo, M.I., and Reyes, L.F. (2018). Pneumonia as a cardiovascular disease. Respirology *23*, 250–259.

18. Corrales-Medina, V.F., Alvarez, K.N., Weissfeld, L.A., Angus, D.C., Chirinos, J.A., Chang, C.-C.H., Newman, A., Loehr, L., Folsom, A.R., Elkind, M.S., et al. (2015). Association between hospitalization for pneumonia and subsequent risk of cardiovascular disease. JAMA *313*, 264–274.

19. Sankaran, R.T., Mattana, J., Pollack, S., Bhat, P., Ahuja, T., Patel, A., and Singhal, P.C. (1997). Laboratory abnormalities in patients with bacterial pneumonia. Chest *111*, 595–600.

20. Letourneau, A.R., Issa, N.C., and Baden, L.R. (2014). Pneumonia in the immunocompromised host. Curr. Opin. Pulm. Med. *20*, 272–279.

21. Stern, A., Skalsky, K., Avni, T., Carrara, E., Leibovici, L., and Paul, M. (2017). Corticosteroids for pneumonia. Cochrane Database Syst. Rev. *12*, CD007720.

22. Larj, M.J., and Bleecker, E.R. (2004). Therapeutic responses in asthma and COPD: corticosteroids. Chest *126*, 138S–149S. discussion 159S-161S.

23. Aliberti, S., Dela Cruz, C.S., Amati, F., Sotgiu, G., and Restrepo, M.I. (2021). Community-acquired pneumonia. Lancet *398*, 906–919.

24. Sessa, M., Mascolo, A., Mortensen, R.N., Andersen, M.P., Rosano, G.M.C., Capuano, A., Rossi, F., Gislason, G., Enghusen-Poulsen,

H., and Torp-Pedersen, C. (2018). Relationship between heart failure, concurrent chronic obstructive pulmonary disease and beta-blocker use: a Danish nationwide cohort study. Eur. J. Heart Fail. *20*, 548–556.

25. Sessa, M., Mascolo, A., Andersen, M.P., Rosano, G., Rossi, F., Capuano, A., and Torp-Pedersen, C. (2016). Effect of chronic kidney diseases on mortality among digoxin users treated for non-valvular atrial fibrillation: a nationwide register-based retrospective cohort study. PLoS One *11*, e0160337.

26. Hagengaard, L., Søgaard, P., Espersen, M., Sessa, M., Lund, P.E., Krogager, M.L., Torp-Pedersen, C., Kragholm, K.H., and Polcwiartek, C. (2020). Association between serum potassium levels and short-term mortality in patients with atrial fibrillation or flutter co-treated with diuretics and rate-or rhythm-controlling drugs. Eur. Heart J. Cardiovasc. Pharmacother. *6*, 137–144.

27. Sessa, M., Mascolo, A., Scavone, C., Perone, I., di Giorgio, A., Tari, M., Fucile, A., de Angelis, A., Rasmussen, D.B., Jensen, M.T., et al. (2018). Comparison of long-term clinical implications of beta-blockade in patients with obstructive airway diseases exposed to beta-blockers with different β1-adrenoreceptor selectivity: an Italian Population-Based Cohort Study. Front. Pharmacol. *9*, 1212.

28. Sessa, M., Rasmussen, D.B., Jensen, M.T., Kragholm, K., Torp-Pedersen, C., and Andersen, M. (2020). Metoprolol versus carvedilol in patients with heart failure, chronic obstructive pulmonary disease, diabetes mellitus, and renal failure. Am. J. Cardiol. *125*, 1069–1076.

29. Sessa, M., Mascolo, A., Dalhoff, K.P., and Andersen, M. (2020). The risk of fractures, acute myocardial infarction, atrial fibrillation and ventricular arrhythmia in geriatric patients exposed to promethazine. Expert Opin. Drug Saf. *19*, 349–357.

30. Pedersen, C.B. (2011). The Danish civil registration system. Scand. J. Public Health *39*, 22–25.

31. Lynge, E., Sandegaard, J.L., and Rebolj, M. (2011). The Danish national patient register. Scand. J. Public Health *39*, 30–33.

32. Andersen, J.S., Olivarius, N.D.F., and Krasnik, A. (2011). The Danish national health service register. Scand. J. Public Health *39*, 34–37.

33. Kildemoes, H.W., Sørensen, H.T., and Hallas, J. (2011). The Danish national prescription registry. Scand. J. Public Health *39*, 38–41.

34. Helweg-Larsen, K. (2011). The Danish register of causes of death. Scand. J. Public Health *39*, 26–29.

35. Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling (Springer).

36. Neumann, U., Genze, N., and Heider, D. (2017). EFS: an ensemble feature selection tool implemented as R-package and web-application. BioData Min. *10*, 21. https://doi.org/10.1186/s13040-017-0142-8.

37. Yu, L., and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. *5*, 1205–1224.

38. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

39. Nembrini, S., König, I.R., and Wright, M.N. (2018). The revival of the Gini importance? Bioinformatics *34*, 3711–3718.

40. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z.M. (2016). Mlr: machine learning in R. J. Mach. Learn. Res. *17*, 5938–5942.

41. Luke, D.A., and Harris, J.K. (2007). Network analysis in public health: history, methods, and applications. Annu. Rev. Public Health *28*, 69–93.

42. Branco, P., Torgo, L., and Ribeiro, R.P. (2015). A survey of predictive modelling under imbalanced distributions. Preprint at arXiv. https://doi.org/10.48550/arXiv.1505.01658.

43. Graffelman, J., and van Eeuwijk, F. (2005). Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. Biom. J. *47*, 863–879.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Danish Civil Registration System	Statistik Denmark	https://pubmed.ncbi.nlm.nih.gov/21775345/
Danish National Patient Register	Statistik Denmark	https://pubmed.ncbi.nlm.nih.gov/21775347/
Danish register for laboratory results for research	Statistik Denmark	https://pubmed.ncbi.nlm.nih.gov/32547238/
Danish National Prescription Registry	Statistik Denmark	https://pubmed.ncbi.nlm.nih.gov/21775349/
Danish Register of Causes of Death	Statistik Denmark	https://pubmed.ncbi.nlm.nih.gov/21775346/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data sources should be directed to and will be fulfilled by the lead contact, Maurizio Sessa (maurizio.sessa@sund.ku.dk).

Material availability

This study did not generate new unique reagents.

Data and code availability

- All data will be shared upon request to the lead contact. No standardized datatype data were generated in this study.
- This study did not generate new code.
- Any additional analysis information for this work is available by request to the lead contact.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The study does not need ethical approval and patient consent because Danish register-based cohort studies are exempted. Patient records/information before the analysis was pseudonymized. The University of Copenhagen and Statistics Denmark (project number 707278) have appropriate data approval from the Regional Capital Area Data Protection Agency to facilitate the conduct of the present study. In particular, this study is part of a bigger project entitled ''Identification and risk minimization of potentially inappropriate prescriptions in patients aged 65 or older'', for which Dr. Sessa has established a cohort of individuals aged 65 or older in Denmark.

METHOD DETAILS

Study design and methods

Study design and setting

A population-based study including residents in Denmark aged 65 years or older in the period 01/01/1996 to 30/06/2018.

Study population

The study population was composed of individuals hospitalized for pneumonia between January 1996 to June 2018 and up to 5 individuals that were risk-set matched by age and sex at the day of hospitalization (i.e., index date) were at the time of the matching were not dead, migrated or hospitalized for CAP. We defined individuals as being hospitalized for CAP if they have been admitted to the hospital with an admission diagnosis of CAP (International Classification of Diseases, ICD_{10}: J12-J18, A709, or A481) (20).

Data sources

Danish administrative and healthcare databases were used as real-world data sources. These data sources have been extensively used in epidemiological research.[24–29] Sociodemographic characteristics of residents in Denmark including age, sex, immigration, and emigration information were obtained from the Danish Civil Registration System.[30] Data on the information and diagnoses of hospitalizations, surgery and medical procedures, biomarkers, and redeemed medications from territorial pharmacies were obtained from the Danish National Patient Register,[31] the Danish register for laboratory results for research,[32] and the Danish National Prescription Registry,[33] respectively. Finally, the date of death was obtained from the Danish Register of Causes of Death.[34] These registers contained sufficient information to virtually track the entire medical history of individuals included in the study population.

Variables

Two different covariate assessment windows were used to generate the high-dimensional sets of variables furtherly used in AI models to predict CAP hospitalization. Redeemed prescriptions from community pharmacies and results of laboratory analysis were assessed 365 days before the index date. Surgeries/procedures and other hospital inpatient and outpatient admissions were assessed using all the information available in the Danish registers before the index date. Within each of *p* data dimensions (i.e., inpatient/outpatient diagnostic codes, procedures/surgeries, drugs dispensed, and laboratory analysis results) codes were sorted by their prevalence. Prevalence was measured as the proportion of individuals having a specific code at least once during the covariate assessment windows. The top *n* most prevalent codes were identified as candidate empirical covariates. The prevalence of each code (and therefore its empirical ranking) depends on the granularity of the coding of each data dimension. For ICD-10 codes we set the granularity to 3 digits. Granularity decisions have been considered for all data dimensions, including medication coding for which the entire latest version of the ATC code (anatomical therapeutic chemical classification code) of the medicine was used and laboratory analysis codes to generate binary variables for which the full code was used. For the top *n* most prevalent codes in each data dimension, we assessed how frequently that code was recorded for each patient during the covariate assessment windows. We created for each code 3 binary variables: code occurred 1 time (no/yes), code occurred more than median number of times, and code occurred more than 75th percentile number of times. A code that appeared above the 75th percentile number of times would have a true value for all 3 recurrence variables. If any of the values were equal, the variable representing the higher cut-point was dropped. In our data structure it resulted in 3 (ICD-10 codes, ATC codes, and laboratory analysis codes) x n (most prevalent codes in each data dimension*)* x 3 covariates (code occurred 1 time, median number of times, and 75th percentile number of times) = 9n covariates + sociodemographic characteristics (age at the time of matching, sex, equalized family income, and highest achieved education).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis

Descriptive analysis of the study population were presented as mean age, proportion of individuals of female sex, and the year of inclusion in the study population.

Feature filtering and prioritization

Considering the high dimensionality of the analytical dataset we performed features filtering and prioritization. The first filtering approach was performed to reduce dimensionality and it was based on variance. The total list of features generated using the approach described in section 2.4 was screened and we removed variables having ≥99% identical values across individuals in the study population.[35] The idea of this filter is to get rid of features that only consist of noise and therefore have very little variation.[35] Then an ensemble feature selection (EFS) approach was implemented to rank features importance for pneumonia hospitalization with the final goal of prioritizing the most important predictors for this outcome.[36] The approach incorporates 6 feature prioritization methods for binary classifications and, in particular.

1) P-value from the Mann-Whitney-U Test of being classified as being or not being hospitalized for pneumonia.

2 and 3) P-value from the Pearson- and Spearman-correlation analysis based on relevance and redundancy according to Yu and Liu.[37]

4) β-coefficients from a logistic regression of Z-transformed predictors.
5) area under receiver operating characteristic curve (AUC)-based variable importance measure from ensembles of multiple decision trees based on the random forest algorithm according to Breiman et al.[38]
6) AUC-based variable importance measure from ensembles of multiple decision trees based on the Gini impurity index.[39]

The results of feature prioritization methods were normalized to a common scale ranging in an interval from 0 to 1/6 which, represented was defined ensemble score.[36] The normalized results were used for ranking of features importance that was furtherly used to identify the optimal number of features. The optimal number of features were identified by looking at how the AUC of the classification models changed when we included new features for predicting pneumonia hospitalization. In particular, we first ranked features by the ensemble score and then we iterated the analysis for each individual predictor by including it in the classification model starting from the feature with the highest ensemble score. The iteration was performed as many times as the number of identified predictors and to get the optimal number of features we calculated the derivate of the resulting function of average AUC based on a logistic regression model and a 5-fold cross validation approach. Then we benchmarked 4 machine leaning classification models that incorporate the optimal set of features and, in particular, random partitioning, random forest, ranger random forest, and logistic regression using the R package *caret* (Bischl et al., 2016).[40] The benchmarking of different machine learning models was used to identify the model with the best performance to be used for the prediction of CAP hospitalization using the disease risk score (explained in section 2.5.4).

Clinical and epidemiological reasoning

A team composed of a pharmacist (MS) and a medical doctor (MA) with pluriannual clinical, pharmacovigilance, and pharmacoepidemiology experience performed the clinical and epidemiological assessment of identified predictors. Clinical and epidemiological assessment included the evaluation of plausibility which, was assessed by searching in the scientific literature if the identified predictors have been previously investigated in clinical studies assessing their role in the occurrence of CAP.

Predictors were plenary discussed and disagreements were solved by consensus. In particular, for predictors for which the team assumed that there was no clinical/epidemiological plausibility for CAP, we investigated their predictors using the same approach described in section 2.5.1. For example, if an antibiotic that is recommended for the treatment of infections (including those leading to pneumonia) was identified as a predictor for pneumonia hospitalization, we tried to identify predictors for the antibiotic prescription. The final goal was to differentiate between predictors with and without biological plausibility for the outcome. For all identified predictors, we performed a network analysis to describe the inter-relationship among predictors and we assessed the median time and the interquartile range from the last detection of the predictor to the CAP hospitalization.[41] Finally, we plot the median time to event and the interquartile range between for the optimal set of predictors in a boxplot.

Model performance

To estimate the models' performance and to avoid any possible overfitting problem of benchmarked classification models, a 5-fold cross-validation method was applied. Finally, overall model performance was assessed by averaging model performances for each fold. For assessing the model performance the accuracy, area AUC, sensitivity, specificity, and accuracy were measured for all models using a confusion matrix. Accuracy is an evaluation metric that measures the number of correct predictions made by a model to the total number of predictions made. It is calculated by dividing the number of correct predictions by the total number of predictions. However, in this study, the data with the class imbalance (i.e., unequal size of the dependent variable), can substantially affect the performance in the confusion matrix. We, therefore, added one more extra performance assessment for resampling data with 1:1 matching (1 randomly selected control for each case of pneumonia hospitalization based on risk-set matching). It adopted the synthetic minority over-sampling technique to training data by under-sampling the adequate class and over-sampling the inadequate class to improve the model performance.[42] A sensitivity analysis was conducted using 10-fold cross-validation. The model's performance for each fold of the 5- and 10-fold cross-validation is provided in the Tables S2 and S3.

Disease risk score

The disease risk score has been developed using the ensemble score and the optimal set of predictors. After scaling the ensemble score, we have applied the following formula to obtain weights that were >0 and positive (Equation 1).

$$weights = normalized\ value + 2 * |\min\ (normalized\ values)| \qquad \text{Equation 1}$$

Development of the weights using the ensemble score.

By applying the weights to each individual predictor we have calculated the individual risk score of each patient and then, calculated the predicted probability of developing the outcome based on the disease risk score by average AUC based on a 5-fold cross-validation approach using the classification model with the best performance (explained in section 2.5.3) based on resampling data with 1:1 matching. We have plotted the predicted probability of developing the outcome given the disease risk score.

Predicted versus observed probabilities for the outcome were computed using two logistic regression models: 1) one model using all the predictors, 2) one model using only the disease risk score.

Finally, calibration was performed using bootstrapping to get bias-corrected (overfitting-corrected) estimates of predicted vs. observed values based on nonparametric smoothers for the logistic regression models using the disease risk score.[43]

Predicted probabilities of developing the outcome given the disease risk score with and without calibration were plotted.

Systematic review – Part 1

We evaluated reviews and systematic reviews investigating risk factors for community-acquired pneumonia hospitalization. Community-acquired pneumonia was defined as an acute infection involving the parenchyma of the lungs, which is acquired outside of the hospital. Only studies for which the full text was available in the English language were considered eligible. Abstracts sent to international or national conferences, letters to the editor, and case reports/series were considered ineligible.

Outcome

The main outcome was a description of risk factors for community-acquired pneumonia of studies published per year from 01 January 2017 to 23 March 2022.

Search methods

Ovid MEDLINE (from at 12:54 GTM+1) was searched along with the references listed in the reviews identified with our research query Pneumonia AND Artificial intelligence.

Selection of studies

In the first screening procedure, titles and abstracts of retrieved records were screened by two independent researchers (SS and MS) for obvious exclusions. All articles that were considered eligible at the first screening procedure underwent a full-text evaluation. If disagreements arose during the two steps evaluation process, it was resolved by consensus.

The articles were also screened to identify risk factors for CAP hospitalization. The risk factors have been listed in Table S3.

In total, 56,685 records were identified in Ovid MEDLINE and in the reference list of reviews retrieved with the search query. After title/abstract screening, 56.654 records were eliminated because of ineligibility and 2 articles underwent a full-text evaluation. 2 articles were considered eligible to be included in this systematic review and the identified risk factors were listed (Table S4). The PRISMA flowchart of the selection process is shown in Figure S7.

Systematic review – Part 2

We evaluated observational studies, meta-analyses, and clinical trials investigating risk factors for community-acquired pneumonia hospitalization in Denmark. Community-acquired pneumonia was defined as ''an acute infection involving the parenchyma of the lungs, which is acquired outside of the hospital'' (1). Only studies for which the full text was available in the English language were considered eligible. Abstracts sent to international or national conferences, letters to the editor, and case reports/series were considered ineligible. The reference list of narrative and systematic reviews included with our MEDLINE query were further screened for undetected records.

Outcome

The main outcome was the frequency of studies published per year from 01 January 1950 to 15 Match 2022, a narrative overview of their findings, and a lay description of risk factors for Community-acquired pneumonia.

Search methods for the identification of studies

Ovid MEDLINE (from at 12:54 GTM+1) was searched along with the references listed in the reviews identified with our research query ''Pneumonia Denmark''.

Selection of studies

In the first screening procedure, titles and abstracts of retrieved records were screened by two independent researchers (SS and ML) for obvious exclusions. All articles that were considered eligible at the first screening procedure underwent a full-text evaluation. If disagreements arose during the two steps evaluation process, it was resolved by consensus.

In total, 3533 records were identified in Ovid MEDLINE and the reference list of reviews retrieved with the search query. After title/abstract screening, 3454 records were eliminated because of ineligibility and 79 articles underwent a full-text evaluation. 29 articles were considered eligible to be included in this systematic review and the identified risk factors were listed (Table S5). The PRISMA flowchart of the selection process is shown in Figure S8.