Covid19Risk.ai: An open source repository and online calculator of prediction models for early diagnosis and prognosis of Covid-19

Iva Halilaj[1,2], Avishek Chatterjee[1], Yvonka van Wijk[1], Guangyao Wu[1], Brice van Eeckhout[3], Cary Oberije[1], Philippe Lambin[1]

[1]The D-Lab, Department of Precision Medicine, GROW- School for Oncology, Maastricht University, Maastricht, The Netherlands

[2]Health Innovation Ventures, Maastricht, The Netherlands

[3]Medical Cloud Company, Belgium

*Correspondence: philippe.lambin@maastrichtuniversity.nl

Abstract

Objective

The current pandemic has led to a proliferation of predictive models being developed to address various aspects of COVID-19 patient care. We aimed to develop an online platform that would serve as an open source repository for a curated subset of such models, and provide a simple interface for included models to allow for online calculation. This platform would support doctors during decision-making regarding diagnoses, prognoses, and follow-up of COVID-19 patients, expediting the models' transition from research to clinical practice.

Methods

In this proof-of-principle study, we performed a literature search in PubMed and WHO database to find suitable models for implementation on our platform. All selected models were publicly available (peer reviewed publications or open source repository) and had been validated (TRIPOD type 3 or 2b). We created a method for obtaining the regression coefficients if only the nomogram was available in the original publication. All predictive models were transcribed on a practical graphical user interface using PHP 8.0.0, and published online together with supporting documentation and links to the associated articles.

Results

The open source website https://covid19risk.ai/ currently incorporates nine models from six different research groups, evaluated on datasets from different countries. The website will continue to be populated with other models related to COVID-19 prediction as these become available. This dynamic platform allows COVID-19 researchers to contact us to have their model curated and included on our website, thereby increasing the reach and real-world impact of their work.

Conclusion

We have successfully demonstrated in this proof-of-principle study that our website provides an inclusive platform for predictive models related to COVID-19. It enables doctors to supplement their judgment with patient-specific predictions from externally-validated models in a user-friendly format. Additionally, this platform supports researchers in showcasing their work, which will increase the visibility and use of their models.

Keywords: Covid-19, predictive models, diagnosis, prognosis, nomogram, machine learning.

Introduction

The recent COVID-19 pandemic, at its start, emphasized several key unmet needs in terms of patient stratification using quantifiable metrics [1]. These include (a) identifying, in the uninfected population, at-risk persons who should be subjected to stricter restrictions than the general population [2], and (b) in the infected population, improving the detection of high-risk patients by utilizing all available patient data (e.g., clinical, laboratory, genetic, and radiological features) so as to improve quality of care and use of hospital resources [3][4]. Now, with several vaccines emerging, there is another compelling reason for identifying those who are most at risk and should therefore receive the vaccines first [5–7].

Ideally, one should address the above needs using quantitative tools that (a) help people at home decide (in consultation with their doctor) whether their health status warrants being self-quarantined, and whether their symptoms (if present) indicate the need for visiting the hospital, and (b) help doctors during triage decide if a patient should be sent home, hospitalized in a ward, or admitted to intensive care [8]. Quantifying these probabilities can be done by using predictive machine learning models.

Currently, COVID-19 publications regarding such models are booming. There are numerous studies being published, from multiple countries and all using different inclusion criteria and outcome measures [4]. This heavily complicates the selection of the optimal model for a specific patient [9]. In addition, the quality of the research is sometimes suboptimal, as a recent review paper has shown [4].

We, as researchers working on COVID-19 models, saw an urgent need for a web-based platform that would serve as an open source repository for validated models. Such a platform would allow the user to have a quick overview of the strengths and weaknesses of the curated models that passed our quality checks. The platform would also allow the user to calculate the output of such models by simply providing the inputs in a user-friendly format, rather than creating their own implementation or conducting their own search to find a suitable implementation.

Our aim for this platform is to include validated prediction models (TRIPOD type 2b and 3) [10], acquired from institutions all over the world, related to all aspects of the disease, including risk assessment of being infected, triage at hospital admission, prediction of recovery process during follow-up, and patient inclusion and stratification in clinical trials. We aim to be inclusive, and thus models that are outside the scope of risk assessment and patient stratification are still within the purview of the platform, e.g., diagnostic models. We believe it will be of interest to doctors who want to leverage the results of all the great research that is taking place, and it will also benefit researchers in dissemination of their own work and in learning about the findings of other groups.

The proof-of-concept of such a platform forms the basis of this paper. We intend to maintain this platform as a public service, and increase the number of curated models by encouraging other researchers to share their work through our platform. The benefits to them include (a) helping the researchers to generalize their models by allowing the models to be tested by research groups that are different from the ones that created the model (TRIPOD 4), and (b) improved visibility of their model, which should stimulate usage and citations[11].

Methods

We reviewed the PubMed database of the National Center for Biotechnology Information (NCBI) and the World Health Organization (WHO) database for COVID-19 publications from December 2019 to June 2020. To find relevant publications to our focus we used the terms in the search field: "COVID 2019 prognostic models", "novel coronavirus 2019 diagnostic tools", "COVID-19 predictive models", and "machine-learning COVID 19 models".

The steps that we followed from the literature search until the final stage of publishing online are shown in Figure 1.

Figure 1-The workflow from defining the convenient models until the end phase.

In order to assess the reporting quality of the models from the studies, we tested each paper for its compliance to the TRIPOD (Transparent Reporting of studies on prediction models for Individual Prognosis Or Diagnosis) reporting guideline as shown in Figure 2 [10,12].

Figure 2-TRIPOD types classifications [10].

Getting model coefficients from a Nomogram

In order to improve readability and interpretability by medical specialists, regression models are often published as nomograms, without the model coefficients. To publish the models in a consistent manner on our platform, we used a simple method to extract the coefficients from nomograms. This method is explained using an example taken from one of the implemented models [3], and shown below in Figure 3.

The first step was to determine the relationship between the parameter and the nomogram score, which was done by reading the nomogram, as shown in Table 1.

Figure 3-Nomogram Published in https://doi.org/10.1101/2020.04.03.20052068

Table 1-Point reading for the nomogram

Parameter (Unit)	Equation	Value	Points
Epidemiological history *(yes/no, Boolean)*	x_1	1	9.32
		0	0.00
Wedge/fan-shaped lesion *(yes/no, Boolean)*	x_2	1	10.00
		0	0
Bilateral lower lobes *(yes/no, Boolean)*	x_3	1	8.82

		0	0.00
Ground glass opacities *(yes/no, Boolean)*	x_4	1	3.04
		0	0.00
Crazy paving pattern *(yes/no, Boolean)*	x_5	1	2.10
		0	0.00
WBC *(<4*10^9/L, Boolean)*	x_6	1	0.63
		0	0.00

The relationship between the parameters and the nomogram score P_{total} is described by the following equation:

$$P_{total} = x_1 \cdot 9.32 + x_2 \cdot 10 + x_3 \cdot 8.82 + x_4 \cdot 3.04 + x_5 \cdot 2.10 + x_6 \cdot 0.63$$

The next step is to determine the relationship between the nomogram score and the probability through the regression equation. A logistic regression model follows the following equation:

$$logit(p) = \beta_0 + \sum_n \beta_n \cdot x_n$$

The Logit of the probability and the nomogram score should have a linear relationship, from which the slope was used to determine the value of the coefficients, and the intercept of the model was extracted (Figure 4).

Figure 4 – Logit(P) plotted against the nomogram score

For this example the regression coefficients are shown in Table 2.

Table 2 – Coefficients and intercept extracted from nomogram.

Parameter (Unit)	Coefficient
Epidemiological history	0,93

Wedge/fan-shaped lesion	0,64
Bilateral lower lobes	0,19
Ground glass opacities	0,93
Crazy paving pattern	0,64
WBC	0,19
Intercept	*-4,23*

All the models are written in PHP 8.0.0, where for regression models we set the coefficients and variables in the PHP syntax, thereby making the models operate identically. For the frontend side we used languages such as HTML, CSS and JavaScript for some specific functionalities. The backend of this platform is PHP based and the database is MySQL.

Results

We have created an open source website (https://covid19risk.ai/) to serve as an archive for published AI prediction models related to all aspects of COVID-19, including diagnosis, theragnosis (how to treat the patient, risk stratification), and follow-up (treatment response and complication).

Currently there are nine models implemented and published as illustrated in Table 3. Every showcased model includes a description of the methodology and clinical datasets used for model development and validation, and limitations of each model are explicit.

Model Nr	**Input features-Output**	**Cohort type**	**Tripod** **Type**
Model 1	**Input features**: Age, hospital staff (Yes/No) **Output:** Probability of severe illness [1].	Asymptomatic COVID positive patients	Type 2b

Model 2	**Input features**: Age, hospital staff (Yes/No), body temperature, days since onset of symptoms. **Output**: Probability of severe illness [1].	Symptomatic COVID positive patients	Type 2b
Model 3	**Input features:** Age, CT lesion score (0 = no lung parenchyma involved, 1 = up to 5% of lung parenchyma involved, 2 = 5-25%, 3 = 26-50%, 4 = 51-75%, 5 = 76-100% of lung parenchyma involved; final CT score is a total score from five lobes). **Output**: Probability of severe illness [1].	COVID positive patients with semantic CT features	Type 2b
Model 4	**Input features:** Age, Lymphocyte, C-reactive protein, Lactate dehydrogenase, Creatine kinase, Urea, Calcium. **Output**: Probability of severe illness [1].	COVID-19 positive patients with blood test results	Type 3
Model 5	**Input features:** Signs of pneumonia on CT, History of close contact with Covid-19 confirmed, Patient(yes/No), Fever, Age, Gender, Max temperature, Respiratory symptoms, Neutrophil-to-lymphocyte ratio. **Output**: Probability of severe illness [2].	COVID-19 positive patients	Type 2b
Model 6	**Input features:** Age, Direct bilirubin, Red blood cell distribution width, Blood urea nitrogen, C-reactive protein, Lactate dehydrogenase, Albumin **Output**: Probability of severe illness [13].	COVID-19 positive patients	Type 2b
Model 7	**Input features:** Age, Sex, Diabetes, COPD or emphysema, or chronic Bronchitis, Asthma, Cystic fibrosis, Hypertension, Had a heart attack, had a stroke, Coronary atherosclerosis or other heart disease, Congestive heart failure, Rheumatic kidney disease, Chronic kidney disease, Liver disease, Cancer, Neurocognitive conditions, Sickle cell	COVID-19 positive patients (vulnerable to	Type 3

	anemia, HIV infection (Yes/No) Health history: Organ transplant, Hemodialysis treatment, pneumonia, acute bronchitis, influenza or other acute respiratory infection, pregnant last 2 weeks, hospital admission/emergency last year, height, weigh. Symptoms: fever, shortness in breath, cough, fatigue, body aches, headache, diarrhea, sore throat, decrease smell and taste(Yes/No) **Output**: Predicts vulnerability score to serious illness from COVID-19 [14].	develop serious complications)	
Model 8	**Input features:** Age, cardiovascular disease, diabetes, chronic respiratory disease, hypertension, cancer, prior stroke, heart disease, chronic kidney disease. **Output**: Estimate mortality rates in patients with COVID-19 [15].	COVID-19 positive patients	Type 3
Model 9	**Input features**: Epidemiological history, wedge-shaped or fan-shaped lesion parallel or near to the pleura, bilateral lower lobes, ground glass opacities, crazy paving pattern, WBC **Output**: Probability of severe illness [3].	Suspected COVID-19 pneumonia patients	Type 2b

Table 3-For every model: input features, output, cohort type and TRIPOD type.

For each online model given, doctors can find: a) the intended use (predicted outcome) of the model, b) to which patients does this tool apply (particularly among individuals with preexisting medical conditions), c) the information and the parameters that need to be entered by the doctor, and d) how the tool was developed. The doctors can visit the website, choose an applicable model, and fill in the variables asked in order to generate a probability. The COVID-19 predictive models on the website use the same calculations as the models described in the scientific publications on which they are based.

The main result of our work is a broadly applicable platform, which includes validated models regarding different stages, symptoms and outcomes of COVID-19. This repository of COVID-19 predictive models will serve as a decision aid for doctors.

Discussion

This platform can be viewed as a "model zoo" aimed at researchers and clinicians and with adequate grasp of the medical complexities associated with COVID-19. The aim for all showcased models is to stimulate research and supplement clinical judgment, not substitute it. The open source website is not intended for unaided use by laypeople (e.g., patients).

We re-emphasize that this manuscript and the website in its current form are only a proof-of-principle. We do not claim that all models that would pass our selection criteria have been included. Similarly, any model not currently included on the platform should not be seen as problematic. Our inclusion period ranged from December 2019 till June 2020. As many models were published since then, an update of the search and the website needs to be and will be done in the near future.

This paper should be seen by researchers from outside our collaboration as an invitation to participate on this platform, with the option of keeping the code hidden from the end user while still offering full functionality. We will assist external researchers for the successful incorporation of their models on our platform. This will create synergies that are bound to accelerate AI research on COVID-19. It will also ensure that models get the recognition they deserve and are used widely, instead of gathering dust as often happens when there are many publications on the same broad theme during a short period (a certainty in the context of COVID-19, given its world-changing nature).

The method we used for retrieving coefficients of a regression model from a nomogram has certain limitations. For one, the accuracy is highly dependent on the resolution of the published model. Another limitation is that though the coefficients of the model are retrieved, the standard error for the coefficients of the parameters cannot be obtained from a nomogram alone. However, the method can be applied to any nomogram, making it a tool that can be broadly used, not restricted to COVID-19.

Conclusions

Our platform (https://covid19risk.ai/), at the current proof-of-principle stage, includes nine validated machine-learning models to serve as decision aids to doctors for various aspects of COVID-19 patient care. Our method for obtaining regression coefficients from a nomogram can be used by other researchers, including in non-COVID contexts. Our platform will be maintained and regularly updated for at least three years, since we have secured funding for this period (DRAGON grant). Therefore, we are encouraging research groups to collaborate with us to share their models with the world.

Acknowledgments

Authors acknowledge financial support from the European Commission's Horizon 2020 research and innovation programme under grant agreement MSCA-ITN-PREDICT n° 766276, and the IMI2 Joint undertaking program under grant agreement DRAGON n° 101005122.

Disclosure

Dr Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic/DNAmito, Health

Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Oncoradiomics, BHV, Merck, Varian, Elekta, ptTheragnostic and Convert pharmaceuticals. Dr Lambin has shares in the company Oncoradiomics SA, Convert pharmaceuticals SA and The Medical Cloud Company SPRL and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Oncoradiomics and Health Innovation Ventures and three non-issues, non licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper.

References

1. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a Clinical Decision Support System for Severity Risk Prediction and Triage of COVID-19 Patients at Hospital Admission: an International Multicenter Study. European Respiratory Journal. 2020. p. 2001104. doi:10.1183/13993003.01104-2020

2. Song C-Y, Xu J, He J-Q, Lu Y-Q. COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. doi:10.1101/2020.03.05.20031906

3. Wang Z, Weng J, Li Z, Hou R, Zhou L, Ye H, et al. Development and Validation of a Diagnostic Nomogram to Predict COVID-19 Pneumonia. doi:10.1101/2020.04.03.20052068

4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020. p. m1328. doi:10.1136/bmj.m1328

5. Reiter PL, Pennell ML, Katz ML. Acceptability of a COVID-19 vaccine among adults in the United States: How many people would get vaccinated? Vaccine. 2020;38: 6500–6507.

6. Williams L, Gallant AJ, Rasmussen S, Brown Nicholls LA, Cogan N, Deakin K, et al. Towards intervention development to increase the uptake of COVID-19 vaccination among those at high risk: Outlining evidence-based and theoretically informed future intervention content. British Journal of Health Psychology. 2020. pp. 1039–1054.

doi:10.1111/bjhp.12468

7. Persad G, Peek ME, Emanuel EJ. Fairly Prioritizing Groups for Access to COVID-19 Vaccines. JAMA. 2020. doi:10.1001/jama.2020.18513

8. Helms J, CRICS TRIGGERSEP Group (Clinical Research in Intensive Care and Sepsis Trial Group for Global Evaluation and Research in Sepsis), Tacquard C, Severac F, Leonard-Lorant I, Ohana M, et al. High risk of thrombosis in patients with severe SARS-CoV-2 infection: a multicenter prospective cohort study. Intensive Care Medicine. 2020. pp. 1089–1098. doi:10.1007/s00134-020-06062-x

9. Jehi L, Ji X, Milinovich A, Erzurum S, Merlino A, Gordon S, et al. Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19. PLOS ONE. 2020. p. e0237419. doi:10.1371/journal.pone.0237419

10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. European Urology. 2015. pp. 1142–1151. doi:10.1016/j.eururo.2014.11.025

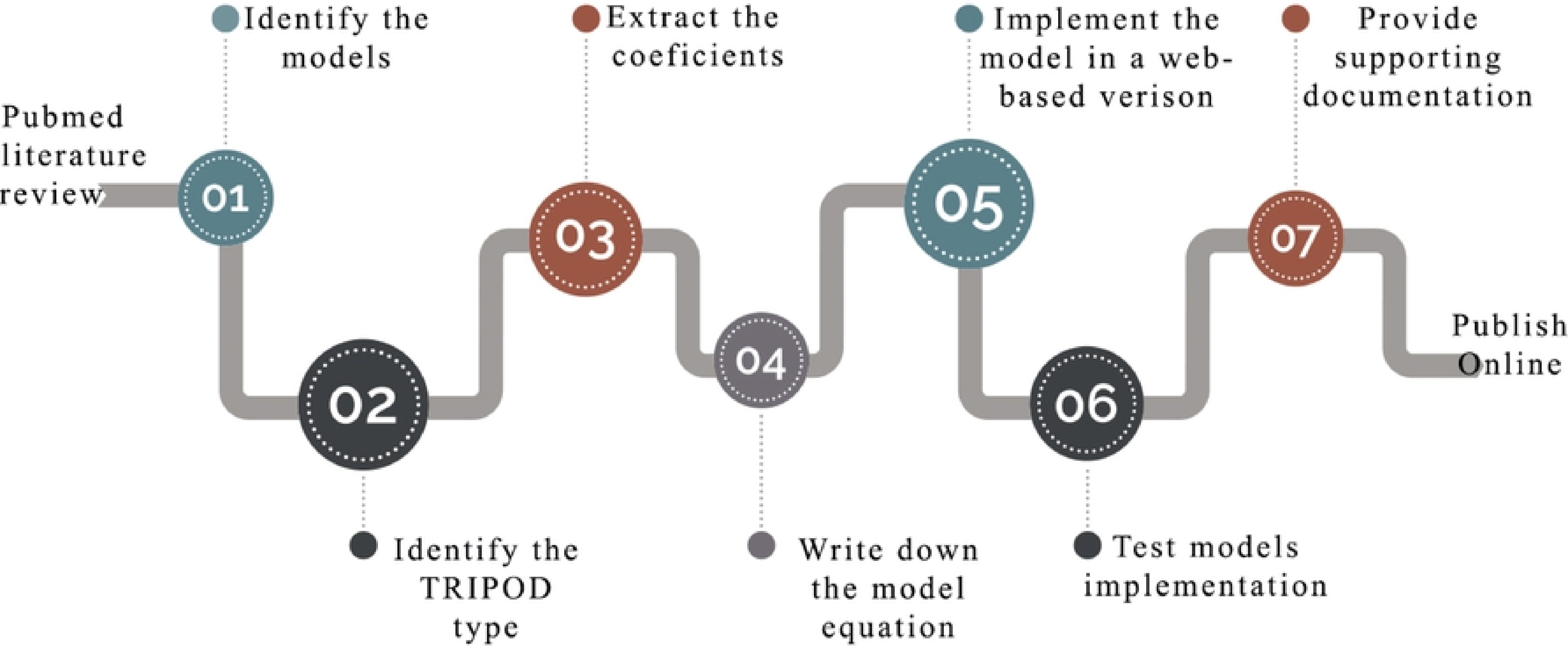
11. Naseem M, Akhund R, Arshad H, Ibrahim MT. Exploring the Potential of Artificial Intelligence and Machine Learning to Combat COVID-19 and Existing Opportunities for LMIC: A Scoping Review. J Prim Care Community Health. 2020;11: 2150132720963634.

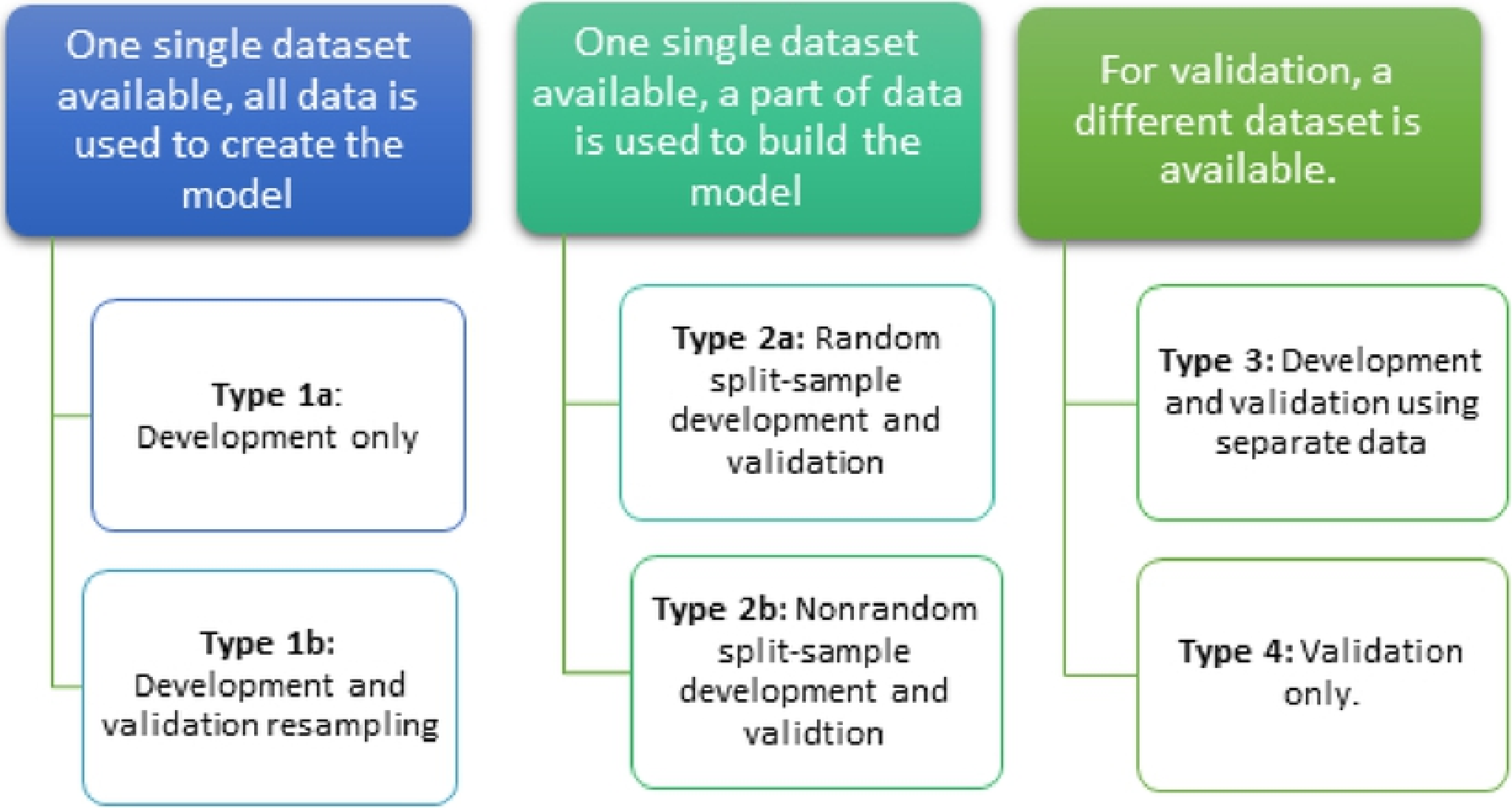
12. van Wijk Y, Halilaj I, van Limbergen E, Walsh S, Lutgens L, Lambin P, et al. Decision Support Systems in Prostate Cancer Treatment: An Overview. Biomed Res Int. 2019;2019: 4961768.

13. Gong J, Ou J, Qiu X, Jie Y, Chen Y, Yuan L, et al. A Tool to Early Predict Severe Corona Virus Disease 2019 (COVID-19) : A Multicenter Study using the Risk Nomogram in Wuhan and Guangdong, China. doi:10.1101/2020.03.17.20037515

14. COVID-19 Vulnerability Index (cv19index) - ClosedLoop.ai. [cited 15 Dec 2020]. Available: https://closedloop.ai/c19index/

15. COVID-19 Prognostic Tool. [cited 15 Dec 2020]. Available: https://qxmd.com/calculate


Pubmed literature review
01
Identify the models
02
Identify the TRIPOD type
03
Extract the coeficients
04
Write down the model equation
05
Implement the model in a web-based verison
06
Test models implementation
07
Provide supporting documentation
Publish Online



One single dataset available, all data is used to create the model

Type 1a: Development only

Type 1b: Development and validation resampling

One single dataset available, a part of data is used to build the model

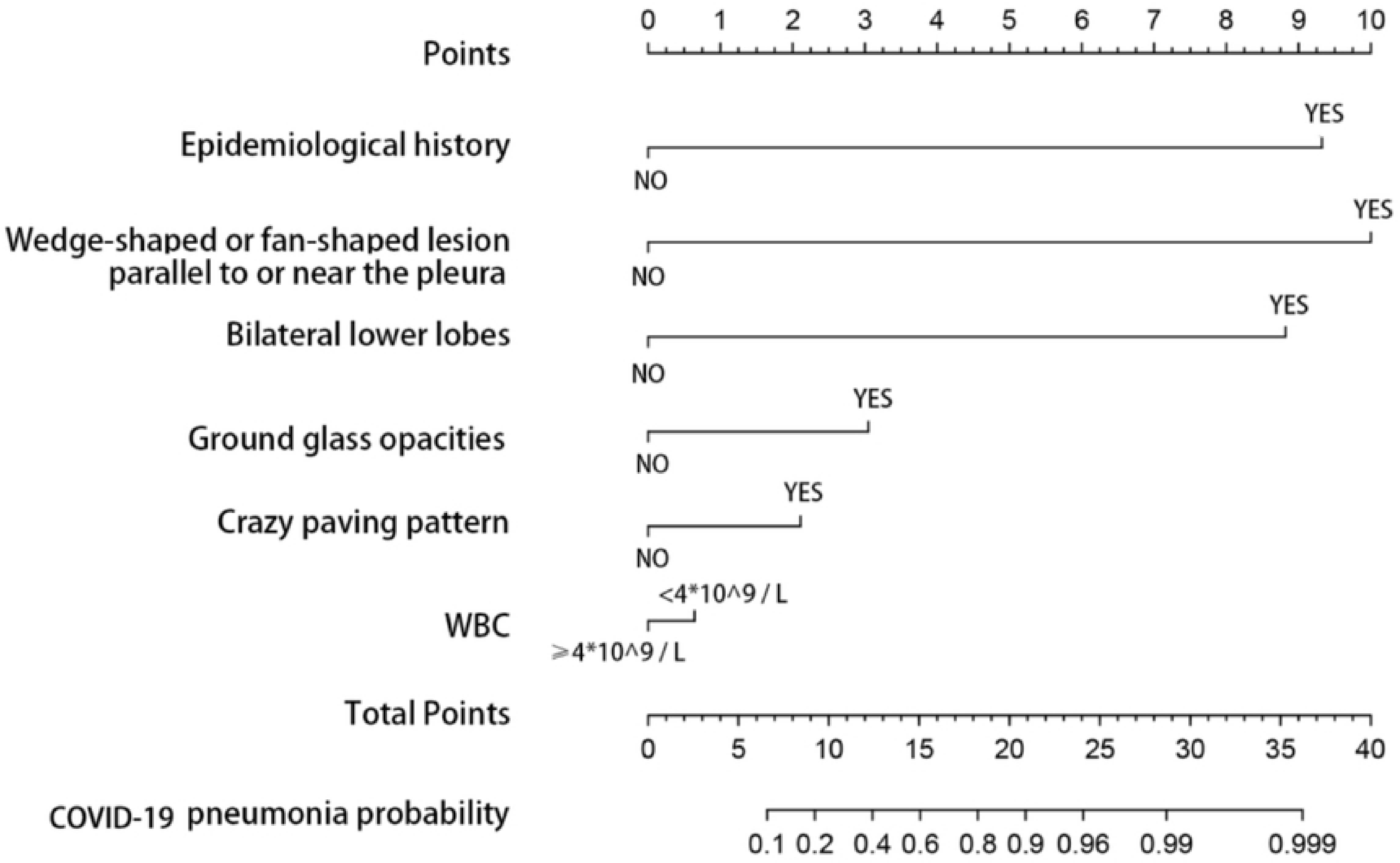
Type 2a: Random split-sample development and validation

Type 2b: Nonrandom split-sample development and validtion

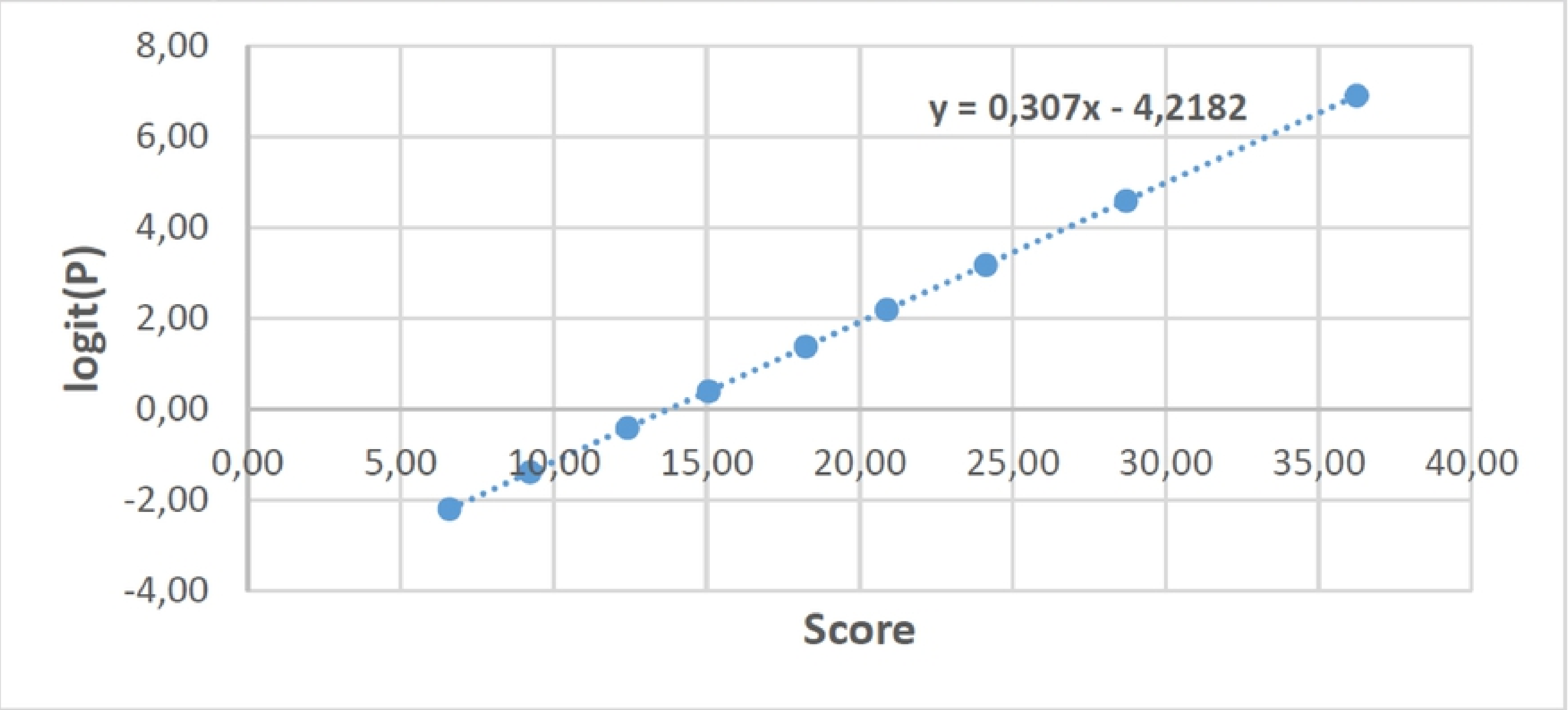
For validation, a different dataset is available.

Type 3: Development and validation using separate data

Type 4: Validation only.


Points
0
1
2
3
4
5
6
7
8
9
10
Epidemiological history
YES
NO
Wedge-shaped or fan-shaped lesion parallel to or near the pleura
YES
NO
Bilateral lower lobes
YES
NO
Ground glass opacities
YES
NO
Crazy paving pattern
YES
NO
WBC
<4*10^9 / L
⩾4*10^9 / L
Total Points
0
5
10
15
20
25
30
35
40
COVID-19 pneumonia probability
0.1
0.2
0.4
0.6
0.8
0.9
0.96
0.99
0.999

Score	6,59	9,23	12,41	15,06	18,24	20,88	24,12	28,71	36,24
logit(P)	-2,20	-1,39	-0,41	0,41	1,39	2,20	3,18	4,60	6,91


y = 0,307x - 4,2182
logit(P)
Score