CLINICAL AND TRANSLATIONAL MEDICINE Open Access

EASY-APP: An artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis

Journal:	*Clinical and Translational Medicine*
Manuscript ID	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kui, Balázs; Department of Medicine, University of Szeged; Centre for Translational Medicine, Department of Medicine, University of Szeged Pintér, József; Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics Molontay, Roland; Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics; MTA-BME Stochastics Research Group Nagy, Marcell; Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics Farkas, Nelli; Institute for Translational Medicine, Medical School, University of Pécs; Institute of Bioanalysis, Medical School, University of Pécs Gede, Noémi; Institute for Translational Medicine, Medical School, University of Pécs Vincze, Áron; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Bajor, Judit; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Gódi, Szilárd; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Czimmer, József; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Szabó, Imre; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Illés, Anita; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Sarlós, Patrícia; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Hágendorn, Roland; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Pár, Gabriella; Division of Gastroenterology, First Department of Medicine, Medical School, University of Pécs Papp, Mária; Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen Vitális, Zsuzsanna; Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen Kovács, György; Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen

Fehér, Krisztina; Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen
Földi, Ildikó; Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen
Izbéki, Ferenc; Szent György Teaching Hospital of Fejér County
Gajdán, László; Szent György Teaching Hospital of Fejér County
Fejes, Roland; Szent György Teaching Hospital of Fejér County
Németh, Balázs; Department of Medicine, University of Szeged
Török, Imola; County Emergency Clinical Hospital of Târgu Mures - Gastroenterology Clinic and University of Medicine, Pharmacy, Sciences and Technology ”George Emil Palade”
Farkas, Hunor; County Emergency Clinical Hospital of Târgu Mures - Gastroenterology Clinic and University of Medicine, Pharmacy, Sciences and Technology ”George Emil Palade”
Mickevicius, Artautas; Vilnius University Hospital Santaros Clinics
Sallinen, Ville; Department of Transplantation and Liver Surgery, Helsinki University Hospital and University of Helsinki
Galeev, Shamil; Saint Luke Clinical Hospital
Ramírez-Maldonado, Elena; General Surgery, Consorci Sanitari del Garraf
Párniczky, Andrea; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs; Heim Pál National Pediatric Institute
Erőss, Bálint; Division of Pancreatic Diseases, Heart and Vascular Center, Semmelweis University; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs; Centre for Translational Medicine, Semmelweis University
Hegyi, Péter; Division of Pancreatic Diseases, Heart and Vascular Center, Semmelweis University; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs
Márta, Katalin; Division of Pancreatic Diseases, Heart and Vascular Center, Semmelweis University; Centre for Translational Medicine, Semmelweis University
Váncsa, Szilárd; Centre for Translational Medicine, Semmelweis University; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs; Division of Pancreatic Diseases, Heart and Vascular Center, Semmelweis University
Sutton, Robert; Institute of Systems, Molecular and Integrative Biology, University of Liverpool and Liverpool University Hospitals NHS Foundation Trust
Halloran, Chris; Institute of Systems, Molecular and Integrative Biology, University of Liverpool and Liverpool University Hospitals NHS Foundation Trust
Latawiec, Diane; Institute of Systems, Molecular and Integrative Biology, University of Liverpool and Liverpool University Hospitals NHS Foundation Trust
Szatmary, Peter; Institute of Systems, Molecular and Integrative Biology, University of Liverpool and Liverpool University Hospitals NHS Foundation Trust
de-Madaria, Enrique; Gastroenterology Department, Alicante University General Hospital, ISABIAL
Pando, Elizabeth; Department of Hepato-Pancreato-Biliary and Transplant Surgery, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona
Alberti, Piero; Department of Hepato-Pancreato-Biliary and Transplant Surgery, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona
Gómez-Jurado, Maria; Department of Hepato-Pancreato-Biliary and Transplant Surgery, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona
Tantau, Alina; The 4th Medical Clinic, "Iuliu Hatieganu" University of Medicine and Pharmacy; Department of Gastroenterology and

Hepatology Medical Center
Szentesi, Andrea; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs; Centre for Translational Medicine, Department of Medicine, University of Szeged
Hegyi, Peter; Centre for Translational Medicine, Semmelweis University; Division of Pancreatic Diseases, Heart and Vascular Center, Semmelweis University; Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs

Keywords: Pancreatitis, Severity, Prediction, Web application

Themed Topics: Others

Abstract:

Background: Acute pancreatitis (AP) is a potentially severe or even fatal inflammation of the pancreas. Early identification of patients at high risk for developing a severe course of the disease is crucial for preventing organ failure and death. Most of the former predictive scores require many parameters or at least 24 hours to predict the severity, therefore the early therapeutic window is often missed.
Methods: The early achievable severity index (EASY) is a multicentre, multinational, prospective, observational study (ISRCTN10525246). The predictions were made using machine learning models. We used the scikit-learn, xgboost, and catboost Python packages for the modelling. We have evaluated our models using 4-fold cross-validation and the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and accuracy metrics have been calculated on the union of the test sets of the cross-validation. The most critical factors and their contribution to the prediction were identified using a modern tool of explainable artificial intelligence, called SHapley Additive exPlanations (SHAP).
Results: The prediction model is based on the international cohort of 1,184 and a validation cohort of 3,543 patients. The best performing model was an XGBoost classifier with an average AUC score of 0.81 and accuracy of 89.1% and, the model is improving with experience. The six most influential features are the respiratory rate, body temperature, abdominal muscular reflex, gender, age, and glucose level. Using the XGBoost machine learning algorithm for prediction, the SHAP values for the explanation, and the bootstrapping method to estimate confidence we have developed a free and easy-to-use web application in the Streamlit Python-based framework (http://easy-app.org/).
Conclusions: The EASY prediction score is a practical tool for identifying patients at high risk for severe AP within hours of hospital admission. The web application is available for clinicians and contributes to the improvement of the model.

EASY-APP: An artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis

Running title: Early prediction of severity in acute pancreatitis

Balázs Kui[1,2], József Pintér[3], Roland Molontay[3,4], Marcell Nagy[3], Nelli Farkas[5,6], Noémi Gede[5], Áron Vincze[7], Judit Bajor[7], Szilárd Gódi[7], József Czimmer[7], Imre Szabó[7], Anita Illés[7], Patrícia Sarlós[7], Roland Hágendorn[7], Gabriella Pár[7], Mária Papp[8], Zsuzsanna Vitális[8], György Kovács[8], Eszter Fehér[8], Ildikó Földi[8], Ferenc Izbéki[9], László Gajdán[9], Roland Fejes[9], Balázs Csaba Németh[1,2], Imola Török[10], Hunor Farkas[10], Artautas Mickevicius[11], Ville Sallinen[12], Shamil Galeev[13], Elena Ramírez-Maldonado[14], Andrea Párniczky[5,15], Bálint Erőss[5,16,17], Péter Jenő Hegyi[5,16], Katalin Márta[16,17], Szilárd Váncsa[5,16,17], Robert Sutton[18], Peter Szatmary[18], Diane Latawiec[18], Chris Halloran[18], Enrique de-Madaria[19], Elizabeth Pando[20], Piero Alberti[20], Maria José Gómez-Jurado[20], Alina Tantau[21,22], Andrea Szentesi[2,5], Péter Hegyi[5,16,17,†], and the Hungarian Pancreatic Study Group

[1]Department of Medicine, University of Szeged, Szeged, Hungary
[2]Centre for Translational Medicine, Department of Medicine, University of Szeged, Szeged, Hungary
[3]Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics, Budapest, Hungary
[4]MTA-BME Stochastics Research Group, Budapest Hungary
[5]Institute for Translational Medicine, Szentágothai Research Centre, Medical School, University of Pécs, Pécs, Hungary
[6]Institute of Bioanalysis, Medical School, University of Pécs, Pécs, Hungary
[7]Division of Gastroenterology, First Department of Medicine, University of Pécs, Medical School, Pécs, Hungary
[8]Department of Gastroenterology, Institute of Internal Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary
[9]Szent György Teaching Hospital of County Fejér, Székesfehérvár, Hungary
[10]County Emergency Clinical Hospital of Târgu Mures - Gastroenterology Clinic and University of Medicine, Pharmacy, Sciences and Technology "George Emil Palade", Targu Mures, Romania
[11]Vilnius University Hospital Santaros Clinics, Vilnius, Lithuania
[12]Department of Transplantation and Liver Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
[13]Saint Luke Clinical Hospital, St. Petersburg, Russia
[14]General Surgery, Consorci Sanitari del Garraf, Sant Pere de Ribes, Spain
[15]Heim Pál National Pediatric Institute, Budapest, Hungary
[16]Division of Pancreatic Diseases, Heart and Vascular Centre, Semmelweis University, Budapest, Hungary
[17]Centre for Translational Medicine, Semmelweis University, Budapest, Hungary
[18]Institute of Systems, Molecular and Integrative Biology, University of Liverpool and Liverpool University Hospitals NHS Foundation Trust, Liverpool, England, United Kingdom
[19]Gastroenterology Department, Alicante University General Hospital, ISABIAL, Alicante, Spain
[20]Department of Hepato-Pancreato-Biliary and Transplant Surgery, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain
[21]The 4th Medical Clinic, "Iuliu Hatieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania
[22]Department of Gastroenterology and Hepatology Medical Center, Cluj-Napoca, Romania

[†]Correspondence: Péter Hegyi, Centre for Translational Medicine, Semmelweis University, H-1085 Budapest, Üllői út 26., Hungary, hegyi2009@gmail.com

ABSTRACT

Background: Acute pancreatitis (AP) is a potentially severe or even fatal inflammation of the pancreas. Early identification of patients at high risk for developing a severe course of the disease is crucial for preventing organ failure and death. Most of the former predictive scores require many parameters or at least 24 hours to predict the severity, therefore the early therapeutic window is often missed.

Methods: The early achievable severity index **(**EASY) is a multicentre, multinational, prospective, observational study (ISRCTN10525246). The predictions were made using machine learning models. We used the scikit-learn, xgboost, and catboost Python packages for the modelling. We have evaluated our models using 4-fold cross-validation and the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and accuracy metrics have been calculated on the union of the test sets of the cross-validation. The most critical factors and their contribution to the prediction were identified using a modern tool of explainable artificial intelligence, called SHapley Additive exPlanations (SHAP).

Results: The prediction model is based on the international cohort of 1,184 and a validation cohort of 3,543 patients. The best performing model was an XGBoost classifier with an average AUC score of 0.81 and accuracy of 89.1% and, the model is improving with experience. The six most influential features are the respiratory rate, body temperature, abdominal muscular reflex, gender, age, and glucose level. Using the XGBoost machine learning algorithm for prediction, the SHAP values for the explanation, and the bootstrapping method to estimate confidence we have developed a free and easy-to-use web application in the Streamlit Python-based framework (http://easy-app.org/).

Conclusions: The EASY prediction score is a practical tool for identifying patients at high risk for severe AP within hours of hospital admission. The web application is available for clinicians and contributes to the improvement of the model.

Keywords: severity prediction, acute pancreatitis, artificial intelligence

INTRODUCTION

Acute pancreatitis (AP) is one of the most challenging and common gastroenterological diseases which requires hospitalization. The importance of the investigation of AP is uncontroversial: more than 2.6 billion dollars is the annual cost of the treatment of AP in the USA where it causes approximately 300,000 emergency department visits every year [1, 2].

According to the revised Atlanta classification, the severity of AP can be determined as mild, moderately severe and severe disease course [3]. In general, 70-75 % of patients have mild disease with a very low mortality rate; however, the remaining 20-25% of patients have moderately severe, and 5-10% severe disease with high mortality [4, 5]. Moderately severe AP is associated with transient organ failure (less than 48 hours) and/or local complications. In the case of severe AP, organ failure is persistent (more than 48 hours), with a mortality rate up to 50% [3, 6]. Mortality in AP is spread over two periods: during the early phase (first two weeks), indicative of rampant disease, or during the late phase (third week and later) following progressive deterioration [7-9].

Early identification of those patients who are at a greater risk for developing complications is necessary to reduce the risk of adverse disease outcomes and death. Several prediction scores and biochemical markers have been evaluated and compared in the past [10-13]. No laboratory test is consistently accurate for the prediction of AP severity. For example, C-reactive protein (CRP) levels at 48 hours are significantly higher in the severe pancreatitis group than in the others, but cannot be used on admission because of the low accuracy [14]. Concerning multifactorial scoring systems, all have limitations: typically, these require many and/or not easily accessible variables, or 48-72 hours for evaluation. As a result none have been adopted for widespread use in daily clinical practice. The Acute Physiology and Chronic Health Examination (APACHE) II score was developed for the assessment of critically ill patients, not specific to AP. The calculation of APACHE II is complicated requiring invasive measurements, including blood gases [15]. Ranson and Glasgow-Imrie scores contain parameters that are not routinely measured, and completion of these scores requires 48 hours from hospital admission, losing critical time for more intensive resuscitation, analgesia and nutritional support [16, 17]. More recently developed scores for assessing the severity of AP are the Bedside Index of Severity in Acute Pancreatitis (BISAP) and the Harmless Acute Pancreatitis Score (HAPS). While the BISAP score was developed to predict severe AP and mortality, HAPS can predict non-severe AP with high (96-97%) sensitivity and positive predictive value (98%), [18, 19]. The Balthazar score is useful for characterisation of local injury, but is largely useful only several days after admission, and ignores clinical symptoms and signs [20].

Early prediction of AP severity is still awaiting a solution [21, 22].

Many attempts have been made to use artificial intelligence for prediction, as it can detect complex nonlinear relationships between various biochemical parameters and disease outcomes [11]. As a type of artificial intelligence, a machine learning algorithm builds a model based on a training dataset and can improve its performance with experience. Several AP severity prediction models used artificial intelligence and machine learning, but they were based on datasets with low patient numbers and used only internal validation methods [23-26].

Our principal aim was to develop and validate a clinical prediction model of severity in AP that requires parameters easily accessible on admission. Our further aim was to design and develop a practical application for clinicians for easy prediction of severe AP.

METHODS

Preliminary settings

The study protocol was discussed at the third meeting of the Hungarian Pancreatic Study Group (HPSG) in 2014, and the pre-study protocol was published in 2015 [27]. Ethical permission was given by the Scientific and Research Ethics Committee of the Hungarian Medical Research Council (30595/2014/EKU) and the trial was registered in the international ISRCTN registry (ISRCTN10525246). The electronic clinical data registration (eCRF) system for data collection and management was developed and run by the HPSG.

Study design

EASY is a multinational, multicentre, prospective, observational study. In the first phase of the study, simple attainable parameters (medical history, anamnestic data, physical examination, laboratory parameters, and imaging details) were recorded on hospital admission from AP patients from 15 countries and 28 medical centres. In the second phase of the study validation data of AP patients were collected and analysed from four international centres. Centre distribution and case numbers are shown in Supplementary table 1 and 2.

Population

AP patients 18 years of age or older assessed within 12 hours from admission were enrolled after giving their informed consent. Both the definition of acute pancreatitis and severity was based on the revised Atlanta classification [3].

Data collection and management

According to literature data of predictive scores (APACHE II, Glasgow-Imrie, HAPS, BISAP) data of potential prognostic parameters (e.g., medical history, laboratory tests, physical examination, and diagnostic imaging details) were collected.

We applied a four-step data quality control system: after local administrative validation and local professional approval, a central administrative and professional check was undertaken by the study coordination team.

Outcome

After classifying the population into severity groups according to the revised Atlanta classification, a composite binary label was constructed based on the dataset to define the severity of pancreatitis used in the model. The new label was 1 if the outcome of the disease was fatal or the AP was classified as severe (severe AP), otherwise, the label was 0 (non-severe AP).

Predictors and machine learning

Our goal was to predict whether a patient will develop severe or non-severe AP (measured by the composite label introduced above), based on the data that are available at the time of hospital admission. In the language of data science, our problem is a binary classification problem, where the target variable (class label) is the binary degree of severity of AP. The explanatory variables are the parameters measured at the time of admission.

There were two main challenges during data preparation: missing data and imbalanced class distribution. We used a k-nearest-neighbour-based data imputer algorithm, called KNNImputer [28] to handle missing data. Since the dataset is highly imbalanced (only 6% of the patients were labelled as severe), we applied the so-called SMOTE algorithm [29] on the training set to over-sample the severe cases.

The predictions were made using several machine learning algorithms, including Decision Tree, Random Forest, Logistic Regression, SVM, CatBoost, and XGBoost. For the modelling, we used the scikit-learn [30], xgboost [31], and catboost [32] Python packages.

For the evaluation of the model, we used 4-fold cross-validation, which means that data were divided into 4 equally sized subsets, and one of these subsets is selected as a test dataset and the remaining data is used to train the machine learning model. The performance of the model is calculated on the selected test subset, then this procedure is repeated for the other subsets as well, i.e., in each round one of the subsets is the test set and the rest are the training data. Cross-validation aggregates the performance metric, namely, it returns the average performance on the test sets.

We have evaluated our models also using the ROC curve, the AUC, and accuracy metrics calculated from the union of the test sets of the cross-validation.

To measure the confidence of the model, many copies of the machine learning model were trained using a bootstrapping method, i.e., we re-sampled the training dataset 100 times and trained the copies of the model independently on these sets and calculated predicted scores. The 10th and 90th percentiles were used to construct a confidence interval over the score of the prediction. The workflow of developing the prediction model is shown in Figure 1.

Explaining the predictions

Besides predicting the severity of AP, another important goal is to identify the most important factors and their contribution to the prediction using a modern tool of explainable artificial intelligence, called SHapley Additive exPlanations (SHAP) [33]. This so-called SHAP value is able to explain the outcome of a machine learning model, using a game-theoretical concept: the Shapley value. The SHAP value quantifies the (marginal) contribution of each feature to the final prediction, which in our case is the severity score of AP. In other words, for a given feature i, the contribution of i, i.e., its SHAP value is the difference between the prediction using the value of i and the mean prediction.

Formally, let f denote the machine learning that for given input parameters $x = (x_1,...,x_d)$ of a patient returns the predicted severity score, moreover, let $D = \{1, 2, ..., 2\}$ denote the set of features. Then let $f_S(x)$ be the conditional expectation of $f(x)$ given the values of the features of the set S, i.e., the values of x_i, where $\forall i \in S$. If S is an empty set, then $f_S(x)$ is the expectation of $f(x)$, formally, $f_{\{\}Other}(x) = E(f(x))$. Using these notations, let us define a function v which calculates the contribution of a feature set S: $\mathrm{v(S)} = \mathrm{f_S(x)} - \mathrm{f_{\{\}}(x)}$, which is the difference between the prediction where we observe the values of the S subset of features and the mean prediction. The contribution $\varphi_i(x)$ of feature i for the prediction of x is defined using v as follows:

$$\varphi_i(x) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!}(v(S \cup \{i\}) - v(S))$$

The SHAP feature importance I_j of feature j is simply the mean absolute contribution of the feature where the average is taken on the whole dataset, i.e.:

$$I_j = \frac{1}{n}\sum_{i=1}^{n}|\varphi_j(x^{(i)})|$$

Using the XGBoost machine learning algorithm for prediction, the SHAP values for the explanation, and the bootstrapping method for the estimation of confidence we have developed a web application in the Streamlit Python-based framework.

Other statistical analyses

Case numbers and percentages were calculated for categorical variables, mean with standard deviation and medians with ranges were calculated for numerical variables in descriptive analysis. A two-sided p-value of <0.05 was considered statistically significant.

RESULTS

Characteristics of the original cohort

1,184 patients diagnosed with AP were included in the analysis. 878 patients (74%) had mild, 243 (21%) moderately severe, and 63 patients (5%) had a severe disease course according to the revised Atlanta classification. There were 26 deaths. With the constructed binary class label, 1,114 patients (94%) were classified as non-severe, while 70 patients (6%) were labelled as having severe disease.

Hence, the data had a highly imbalanced class distribution. The general characteristics of the cohort are detailed in Table 1.

Machine learning models

We trained and evaluated the following binary classifiers: Decision Tree, Random Forest, Logistic Regression, SVM, CatBoost, and XGBoost. The best performing model was an XGBoost classifier with an average AUC score of 0.81 and an accuracy of 89.1%. The receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) are depicted in Figure 2.

Although the size of our dataset is larger than that of previously published studies, we investigated how the performance of the model increases as we increase the size of the training set. We supposed that the model had not reached its maximal performance and could be further improved with more data (Figure 3).

As not all parameters were measured or available on admission, we examined how the performance of the model decreases if it is built from fewer variables. The AUC values for the models built only on the most important *k* attributes (according to their SHAP values) are shown in Figure 4. It is clear that performance increased with the number of variables used for prediction, but reasonable performance is obtained with fewer parameters.

For binary classification, machine learning models usually only predict a score that can be interpreted as the likelihood of the positive class, here the likelihood of severe AP. On the other hand, the confidence of the given prediction usually remains unclear. To assist the physicians in assessing to what extent they can rely on the model's output in decision-making, we also estimated the confidence of the prediction using a bootstrapping method. The confidence intervals for a selected test dataset of 356 records (30% of the dataset) can be seen in Figure 5. The confidence of the model is greater near the endpoints of the spectrum, i.e., when the degree of severity is clearly mild or severe. On the other hand, the width of the confidence intervals in the mid-range is slightly larger.

Explaining the predictions

With the help of the SHAP values, the individual predictions of the machine learning model can be explained, and it is possible to measure the global importance of individual features. The effect of the individual features on the model output and their ranked importance are shown in Figure 6. The top 6 most influential features are the respiratory rate, body temperature, abdominal muscular reflex, gender, age, and glucose level.

Using the SHAP values, explanations of three different predictions are depicted in Figure 7. The features pushing higher the predicted probability of severe AP (compared to the mean prediction, called base value) are shown in orange and those pushing the prediction lower are in green. Moreover, the length of the bars is proportional to the extent the corresponding factor contributes to the prediction.

If most parameters of the patient are normal, the risk of developing severe AP is very low (Figure 7A). The fact that the BMI and glucose level are high, pushes the predicted severity score higher (Figure 7B). In the case of most parameters being outside the normal range (the patient was older, and had a high glucose level, urea nitrogen, BMI, CRP, and respiratory rate), the probability of severe disease increased (Figure 7C). More examples can be found in Supplementary figure 1.

Validation of the results

Our results were validated on external data of four international centres: Alicante, Barcelona, Cluj-Napoca, and Liverpool. Altogether, 3,164 cases were included in the analysis. First, we validated the model's performance by training it on the EASY dataset and then we measured its performance on the four international centres. The AUC score of the model on the Alicante, Barcelona, Cluj-Napoca, and Liverpool data are 0.72, 0.79, 0.74, and 0.77 respectively. We found that the performance of the model improves significantly if we supplement the training data with the international data sets, in this case the cross-validated AUC score is 0.82 on the EASY, 0.79 on the Alicante, 0.82 on the Barcelona, 0.79 on the Cluj-Napoca, and 0.78 on the Liverpool data set. Finally, we measured the model's performance on the union of all the data sets, in this case, the cross-validated AUC score is 0.803. Further details of the validation cohort and the results of the analysis are available in the Supplementary material.

Web application

Using the XGBoost machine learning algorithm for prediction, the SHAP values for the explanation, and the bootstrapping method for the estimation of confidence we have developed a web application (http://easy-app.org/) in the Streamlit Python-based framework. The application is able to operate if not all the input variables are given, however, at least 5 input parameters have to be provided. Although, XGBoost can handle missing data, to be able to interpret the SHAP values we solved this challenge by retraining the model using only the given parameters.

The application returns three different plots that show the probability of having severe pancreatitis according to the model (the predicted severity score), the confidence interval of the prediction severity score, the explanation of the decision of the model, and the distribution of the predicted scores made by the XGBoost models. A prediction in the application is shown in Figure 8.

DISCUSSION

We have applied machine learning to the development and testing of a simple risk score for severe AP of between 0 and 1 that can be calculated on admission from simple bedside parameters. This score has been derived and validated by study of almost 5,000 patients from multiple countries, confirming its applicability at the bedside, now easily used in our web-based application. Furthermore, it is expected that the score will improve with use, as more data are entered. While machine learning models usually lack an explanation behind the output, operating as a "black box" [11], we solved the problem of machine learning model interpretation by applying a novel explainable artificial intelligence (XAI) tool, called SHAP value [33]. This state-of-the-art technique enabled us to identify the variables that affect the prediction, determining the most important factors and their contribution to the prediction. The power of SHAP values has also been illustrated by Lundberg et al. [34] and Haimovich et al. [35] who developed an early prognostic tool for the severity of COVID-19 and used SHAP values to investigate the effects of the individual variables. To the best of our knowledge this is the first work using SHAP values in the prediction of AP severity. In the EASY population, the most relevant factors causing more severe disease were respiratory rate, abdominal guarding, axillary body temperature, serum amylase, gender, and serum glucose level, all routinely and easily obtained. From this, we have developed an easy-to-use web application that gives a prediction for the likelihood of severe AP using a given input of available parameters while explaining the prediction of the machine learning model making it useful not only for prediction but also for education.

Hand-crafted AP severity prediction scores are readily interpretable and easy to understand but have three principal limitations. Firstly, they are unlikely to achieve as high a level of performance as a machine learning model derived from a set of features. Secondly, most of the hand-crafted scores use parameters that are only available at least 24, if not 48 or 72 hours after hospitalization. Thirdly, these scores were developed during the era of the original Atlanta classification that distinguished mild and severe AP, unlike the revised Atlanta classification that distinguishes mild, moderately severe and severe AP. As the predictive capabilities of these scores, comprehensively reviewed by Gurusamy et al. [36], have reached their limit, alternative approaches are required.

In one of the earliest works using machine learning, Pearce et al. [26] applied kernel logistic regression to predict the severity of AP using 8 variables (age, arterial pH, serum C-reactive protein, GCS, pO2onair, respiratory rate, serum creatinine, and white cell count) obtained from 265 patients. Their model achieved a 0.82 AUC score, while the AUC of the APACHE II score was only 0.74.

Qiu et al. [37] used three machine learning models (SVM, logistic regression, neural network) to predict the risk of multiple organ failure in severe AP. The models were built on a relatively small dataset of 263 patients, and the three models' AUC score was between 0.832 and 0.840, while the AUC of the APACHE score was 0.814. They found hematocrit, kinetic time (thromboelastogram), interleukin-6, and creatinine to have the greatest predictive power. Ding et al. [38] used artificial neural networks and logistic regression for an early prediction of in-hospital mortality in AP. The authors used 12 variables that were collected within 24 hours of admission from 337 patients. The performance of the model was relatively low compared to the other works, with an AUC of the neural network at 0.769 and logistic regression at 0.607. Akshintala & Khashab [39] have recently described the application of artificial intelligence to AP prediction in pancreaticobiliary endoscopy, presenting a simple AI-based AP risk prediction calculator and decision making tool. All these previous results derived from relatively small cohorts [11] suggest the potential of machine learning models to improve upon hand-crafted scores, an approach that we have exploited in our work. Our 0.81 AUC value achieved in far larger populations matches than those of the former works, and our model is improving further with use, as it is applied even more widely.

Strength and limitations

There are several strengths of our model. 1) We have used a large international cohort for both the model development and for the external validation. 2) The model is continuously improving with experience. 3) We also explain the prediction with the help of SHAP values, which helps physicians understand the decision of the machine learning model. Moreover, it may also educate patients in finding how to change their lifestyle or behaviour to avoid developing AP again. 4) Our model uses only data that are available at the time of patient admission to the hospital; hence provides a very early prediction of the likelihood of severe AP. 5) Finally, we developed a web application, which for a given set of input parameters returns three outputs: the predicted severity score of AP, the confidence of the model, and the explanation of the prediction that highlights the key factors affecting the severity of AP.

The principal limitation of this study was imposed by its design, namely the use of binary classification for non-severe and severe AP to derive the model. Binary classification has enabled derivation of the likelihood of the development of severe AP but may not be able to accurately distinguish likelihoods of mild from moderately severe AP as these were entered as one class. This results in a score that is more akin to the original rather than revised Atlanta classification, although there may be limitations in the scores obtained for patients with local complications but without persistent organ failure. While the score calculated for any patient varies between 0 and 1, it may be easier for clinicians to understand percentage likelihoods instead; this feature can be altered in the future. More subtly, the confidence limits for the prediction of severity are wider moving away from the prediction spectrum endpoints, i.e. with scores nearer to the middle of the range. Nevertheless, our model is improving with experience, thus, these limitations might decrease with using the web application and feeding the model with further data.

Implications for practice

Based on the predictions we can identify patients at increased risk for severe AP, thereby the model can assist in early triage to intensive care units and selection of patients for specific interventions. The easy-to-use web application (http://easy-app.org/) is a useful tool for clinicians for early prediction. The more they use this application the better the model becomes.

CONCLUSION

The EASY prediction score is a practical tool for identifying patients at a greater risk for severe AP within 24 hours of hospital admission. The easy-to-use web application is available for clinicians and contributes to the improvement of the model.

Authors' contributions: PH conceptualized the study. BK, ÁV, JB, SG, JC, IS, AI, PS, RH, GP, MP, ZV, GK, EF, IF, FI, LG, RF, BCN, IT, HF, AM, VS, SG, ERM, AS, AP, BE, PJH, KM, SV, RS, PS, DL, CH, EM, EP, PA, MJG, AT contributed to the data collection and quality assurance. BK, JP, RM, MN, NF, NG, and AS extracted and analysed the data. BK, PH, RM, and NF interpreted the data. BK, AS and PH wrote the manuscript. All the authors critically reviewed the manuscript before finalization and submission.

Hungarian Pancreatic Study Group (full names are listed in the Contributiors section and affiliations are detailed in the Supplementary material): AH, VDV, TT, LC, ZS, MM, PP, GP, DS, IOZ, AL, ÁP, KZ, ÁVP, IH, ES, JN, MH, GR, LL, ATI, GC, and DL contributed to the data collection. DI, DP, PV, KO, MFJ, MF, AM, and ZS contributed to the assurance of data quality.

Contributors: Adrienn Halász, Veronika Dunás-Varga, Tamás Takács, László Czakó, Zoltán Szepes, Melania Macarie, Petr Pencik, Goran Poropat, Davor Stimac, Imanta Ozola-Zalite, Andrey Litvin, Árpád Patai, Kristina Zadorozhna, Árpád V. Patai, István Hritz, Elena Stilidi, János Novák, Masayasu Horibe, Georgiana Robu, László Lakatos, Ali Tüzün Ince, Gabriele Capurso, and Dušan Leško contributed to the data collection. Dóra Illés, Dániel Pécsi, Péter Varjú, Klementina Ocskay, Márk Félix Juhász, Mária Földi, Alexandra Mikó, and Zsolt Szakács.

Funding: The research was supported by project grants K131996 to PH, FK131864 to AM, K128222 to LC, FK124632 to BCN and by funding from the University of Pécs Medical School Research Fund (300909) to AS. The funders had no effect on the concept, data collection, analysis, and writing of the manuscript.

Ethics approval and consent to participate: The study was approved by the Scientific and Research Ethics Committee of the Hungarian Medical Research Council (30595/2014/EKU). Written informed consent was obtained from all participants before enrolment.

Consent for Publication: Not applicable.

Conflicts of Interest Statement: Authors do not have any conflicts of interest to declare.

Data availability: The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Figure legends

Figure 1. The workflow of the development of the prediction model

Figure 2. The cross-validated (fold=4) ROC curve of the XGBoost model. The corresponding mean area under the curve (AUC) is 0.809. The standard deviation of the AUC scores is 0.017 and the 90% confidence interval of the mean is [0.775, 0.835].

Figure 1. The performance of the XGBoost model trained on different sized sets. The points show the AUC scores, and the bars are the corresponding confidence intervals.

Figure 2. The performance of the model using varying numbers of attributes with the top k most important features. The importance is calculated using the SHAP importance. The points show the AUC scores and the bars are the corresponding confidence intervals.

Figure 3. The predicted severity score on a selected subset of the dataset and the confidence intervals for the 10th and 90th percentiles and the 25th and 75th percentiles. The records are sorted by the severity score.

Figure 4. A summary plot of the impact of the features on the prediction (severity score) of the model. Each patient is represented by a point in each row. The colour of the points represents the relative value of the feature, and the x-position of the points is the SHAP value, i.e., the impact on the model's prediction.

Figure 5. Three examples of the local explanation of the predictions using the SHAP values. **A)** Predicted mild AP. **B)** Predicted AP with borderline severity. **C)** Predicted severe AP. Factors that push the predicted score higher compared to the base value (mean prediction) are coloured orange, and those pushing lower the prediction are shown in green.

Figure 8. An example output of the web application for the following input parameters: age: 55 years, gender: 0 (woman), body mass index: 22, alcohol consumption: 1 (true), blood pressure/pulse: 140/75/60, body temperature: 37.0 °Celsius, respiratory rate: 25. A) Predicted severity score. B) Explanation of the prediction. C) The kernel density estimate plot of the distribution of the predictions.

Supplementary material

Further contributors of the Hungarian Pancreatic Study Group, centre distribution of the original cohort, characteristics of the validation cohort, validation analysis, availability, and details of the web application, TRIPOD checklist.

References

1. Garber, A., et al., *Mechanisms and management of acute pancreatitis.* Gastroenterology research and practice, 2018. 2018.
2. Peery, A.F., et al., *Burden of gastrointestinal disease in the United States: 2012 update.* Gastroenterology, 2012. 143(5): p. 1179-1187. e3.

3. Banks, P.A., et al., *Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus.* Gut, 2013. 62(1): p. 102-111.
4. Hegyi, P., et al., *Accelerating the translational medicine cycle: the Academia Europaea pilot.* Nature Medicine, 2021. 27(8): p. 1317-1319.
5. Párniczky, A., et al., *Prospective, multicentre, nationwide clinical data from 600 cases of acute pancreatitis.* PLoS One, 2016. 11(10): p. e0165309.
6. Sternby, H., et al., *Determinants of severity in acute pancreatitis: a nation-wide multicenter prospective cohort study.* Annals of surgery, 2019. 270(2): p. 348-355.
7. Johnson, C. and M. Abu-Hilal, *Persistent organ failure during the first week as a marker of fatal outcome in acute pancreatitis.* Gut, 2004. 53(9): p. 1340-1344.
8. Moran, R.A., et al., *Early infection is an independent risk factor for increased mortality in patients with culture-confirmed infected pancreatic necrosis.* Pancreatology, 2021.
9. Párniczky, A., et al., *Antibiotic therapy in acute pancreatitis: From global overuse to evidence based recommendations.* Pancreatology, 2019. 19(4): p. 488-499.
10. Gao, W., H.-X. Yang, and C.-E. Ma, *The value of BISAP score for predicting mortality and severity in acute pancreatitis: a systematic review and meta-analysis.* PloS one, 2015. 10(6): p. e0130412.
11. Gorris, M., et al., *Artificial intelligence for the management of pancreatic diseases.* Digestive Endoscopy, 2021. 33(2): p. 231-241.
12. Mikó, A., et al., *Computed tomography severity index vs. other indices in the prediction of severity and mortality in acute pancreatitis: A predictive accuracy meta-analysis.* Frontiers in physiology, 2019. 10: p. 1002.
13. Yang, Y.-X. and L. Li, *Evaluating the ability of the bedside index for severity of acute pancreatitis score to predict severe acute pancreatitis: a meta-analysis.* Medical Principles and Practice, 2016. 25(2): p. 137-142.
14. Farkas, N., et al., *A multicenter, international cohort analysis of 1435 cases to support clinical trial design in acute pancreatitis.* Frontiers in physiology, 2019. 10: p. 1092.
15. Larvin, M. and M. Mcmahon, *APACHE-II score for assessment and monitoring of acute pancreatitis.* The Lancet, 1989. 334(8656): p. 201-205.
16. Ranson, J.H., et al., *Objective early identification of severe acute pancreatitis.* American Journal of Gastroenterology (Springer Nature), 1974. 61(6).
17. Wilson, C., D. Heath, and C. Imrie, *Prediction of outcome in acute pancreatitis: a comparative study of APACHE II, clinical assessment and multiple factor scoring systems.* Journal of British Surgery, 1990. 77(11): p. 1260-1264.
18. Lankisch, P.G., et al., *The harmless acute pancreatitis score: a clinical algorithm for rapid initial stratification of nonsevere disease.* Clinical gastroenterology and hepatology, 2009. 7(6): p. 702-705.
19. Wu, B.U., et al., *The early prediction of mortality in acute pancreatitis: a large population-based study.* Gut, 2008. 57(12): p. 1698-1703.
20. Choi, H.W., et al., *Early Prediction of the Severity of Acute Pancreatitis Using Radiologic and Clinical Scoring Systems With Classification Tree Analysis.* American Journal of Roentgenology, 2018: p. 1035-1043.
21. Mederos, M.A., H.A. Reber, and M.D. Girgis, *Acute pancreatitis: a review.* JAMA, 2021. 325(4): p. 382-390.
22. Silva-Vaz, P., et al., *Multifactorial scores and biomarkers of prognosis of acute pancreatitis: applications to research and practice.* International journal of molecular sciences, 2020. 21(1): p. 338.
23. Andersson, B., et al., *Prediction of severe acute pancreatitis at admission to hospital using artificial neural networks.* Pancreatology, 2011. 11(3): p. 328-335.
24. Keogan, M.T., et al., *Outcome analysis of patients with acute pancreatitis by using an artificial neural network.* Academic radiology, 2002. 9(4): p. 410-419.
25. Mofidi, R., et al., *Identification of severe acute pancreatitis using an artificial neural network.* Surgery, 2007. 141(1): p. 59-66.

26. Pearce, C.B., et al., *Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-reactive protein.* Pancreatology, 2006. 6(1-2): p. 123-131.
27. Hritz, I. and P. Hegyi, *Early Achievable Severity (EASY) index for simple and accurate expedite risk stratification in acute pancreatitis.* Journal of Gastrointestinal & Liver Diseases, 2015. 24(2).
28. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. 17(6): p. 520-525.
29. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. 16: p. 321-357.
30. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. 12: p. 2825-2830.
31. Chen, T. and C. Guestrin, *XGBoost: A scalable tree boosting system In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(pp. 785–794).* New York, NY, USA: ACM, 2016. 10(2939672.2939785).
32. Dorogush, A.V., V. Ershov, and A. Gulin, *CatBoost: gradient boosting with categorical features support.* arXiv preprint arXiv:1810.11363, 2018.
33. Lundberg, S.M. and S.-I. Lee. *A unified approach to interpreting model predictions*. in *Proceedings of the 31st international conference on neural information processing systems*. 2017.
34. Lundberg, S.M., et al., *From local explanations to global understanding with explainable AI for trees.* Nature machine intelligence, 2020. 2(1): p. 56-67.
35. Haimovich, A.D., et al., *Development and validation of the quick COVID-19 severity index: a prognostic tool for early clinical decompensation.* Annals of emergency medicine, 2020. 76(4): p. 442-453.
36. Gurusamy, K.S., T.P. Debray, and G. Rompianesi, *Prognostic models for predicting the severity and mortality in people with acute pancreatitis.* The Cochrane Database of Systematic Reviews, 2018. 2018(5).
37. Qiu, Q., et al., *Development and validation of three machine-learning models for predicting multiple organ failure in moderately severe and severe acute pancreatitis.* BMC gastroenterology, 2019. 19(1): p. 1-9.
38. Ding, N., et al., *An Artificial Neural Networks Model for Early Predicting In-Hospital Mortality in Acute Pancreatitis in MIMIC-III.* BioMed Research International, 2021. 2021.
39. Akshintala, V.S. and M.A. Khashab, *Artificial intelligence in pancreaticobiliary endoscopy.* Journal of Gastroenterology and Hepatology, 2021. 36(1): p. 25-30.

Table 1. Characteristics of the original cohort

Demographic data			**Data quality***
Gender, male %	58.1%	female/male	100%
Age, mean (SD); min, max	55.7 (16.6)	[17, 95]	100%
BMI, mean (SD); min, max	27.98 (5.86)	[14.8, 50.4]	99%
Anamnestic data			
Alcohol consumption, yes %	54.0%	yes/no	100%
Smoking, yes %	34.4%	yes/no	100%
Length of abdominal pain, mean (SD) in hours; min, max	36.8 (40.4)	[1, 168]	98%
Admission data			
Abdominal guarding, yes %	22.7%	yes/no	99%
Abdominal tenderness, yes %	89.6%	yes/no	99%
Body temperature (axillary), °C mean (SD); min, max	36.7 (0.46)	[34.8, 39.0]	98%
Systolic blood pressure (Hgmm), mean (SD); min, max	141.9 (22.5)	[75, 220]	100%
Diastolic blood pressure (Hgmm), mean (SD); min, max	85.2 (14.3)	[40, 191]	100%
Heart rate, mean (SD); min, max	83.9 (16.5)	[41, 153]	100%
Respiratory rate, mean (SD); min, max	17.7 (3.7)	[10, 45]	99%
Laboratory parameters			
Amylase, mean (SD); min, max	1077 (1117)	[16, 8544]	100%
ASAT/GOT, mean (SD); min, max	147.9 (186)	[4, 1251]	99%
Serum ionized Calcium, mean (SD); min, max	2.31 (0.22)	[1.5, 4.5]	98%
C-reactive protein (mg/l), mean (SD); min, max	49.76 (74.5)	[0.07, 515]	100%
Creatinine, mean (SD); min, max	85.8 (46.7)	[36, 706]	100%
Glucose, mean (SD); min, max	8.23 (3.48)	[2.53, 43.29]	100%
Potassium, mean (SD); min, max	4.12 (0.55)	[2.5, 7]	97%
Sodium, mean (SD); min, max	137.8 (4.1)	[116, 155]	97%
Urea (carbamide), mean (SD); min, max	6.32 (3.85)	[0.98, 40.09]	100%
White blood cell count, mean (SD); min, max	12.78 (5.05)	[1.32, 52.70]	100%
Imaging examinations			
Pleural fluid	12.0%	yes/no	88%
Acute peripancreatic fluid collection	22.2%	yes/no	93%
Abdominal fluid	23.0%	yes/no	96%
Outcome			
The severity of acute pancreatitis, severe %	5.9%	non-severe/severe	100%

*Data not missing

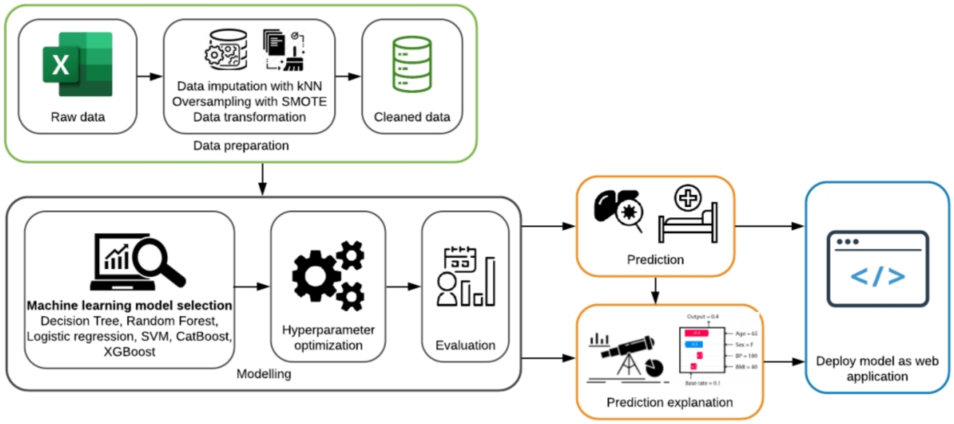


Figure 1. The workflow of the development of the prediction model

754x360mm (47 x 47 DPI)

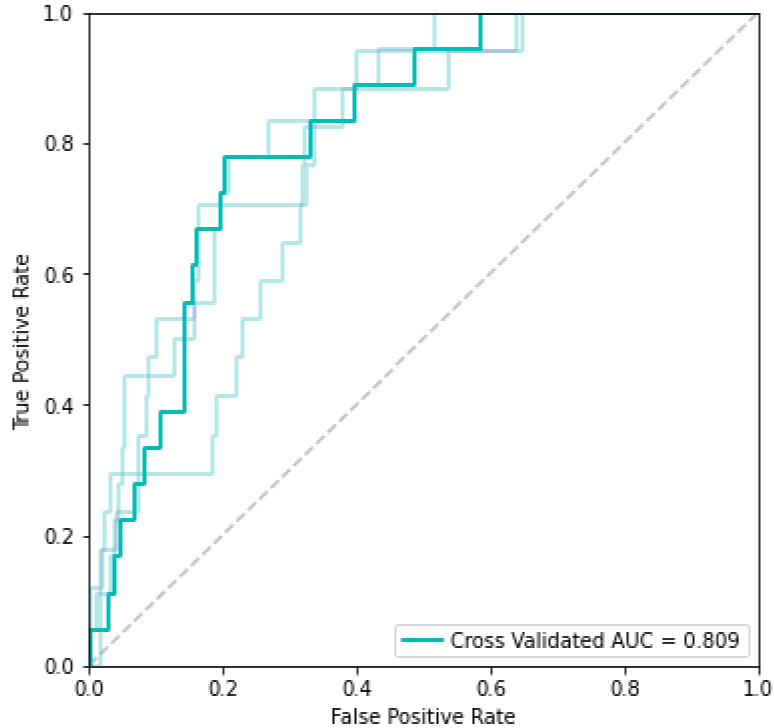


Figure 2. The cross-validated (fold=4) ROC curve of the XGBoost model. The corresponding mean area under the curve (AUC) is 0.809. The standard deviation of the AUC scores is 0.017 and the 90% confidence interval of the mean is [0.775, 0.835].

212x202mm (47 x 47 DPI)

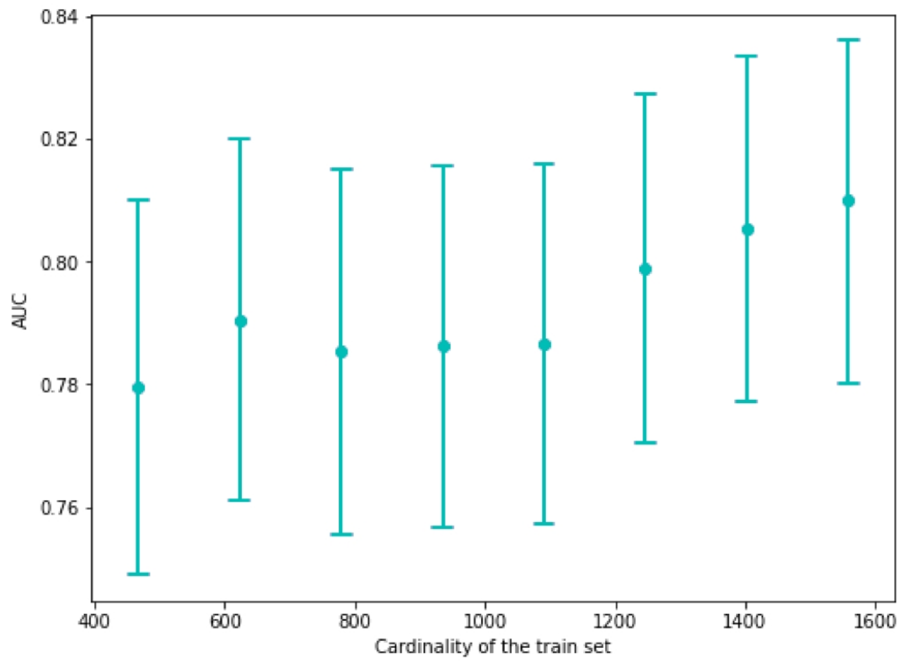


Figure 3. The performance of the XGBoost model trained on different sized sets. The points show the AUC scores, and the bars are the corresponding confidence intervals.

645x484mm (118 x 118 DPI)

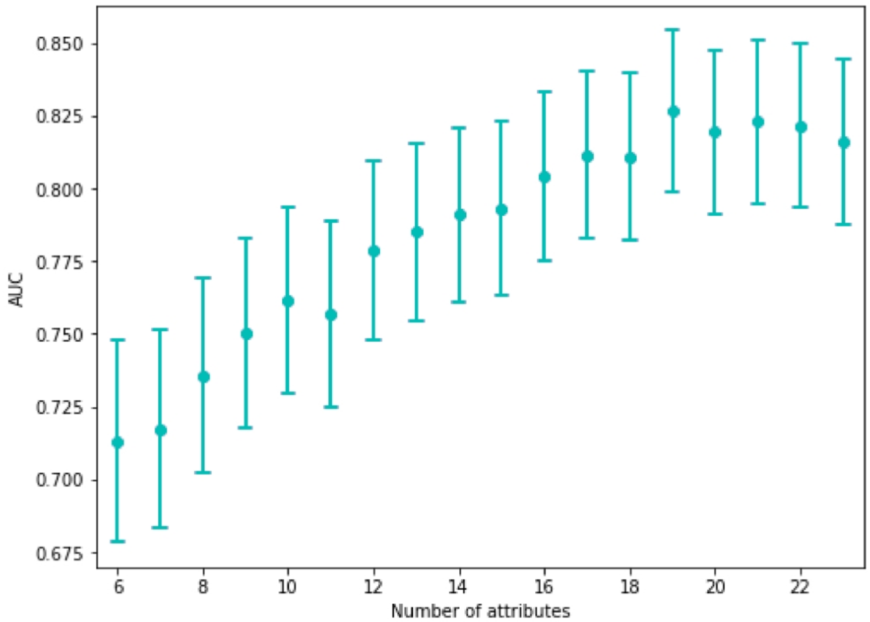


Figure 4. The performance of the model using varying numbers of attributes with the top k most important features. The importance is calculated using the SHAP importance. The points show the AUC scores and the bars are the corresponding confidence intervals.

645x484mm (118 x 118 DPI)

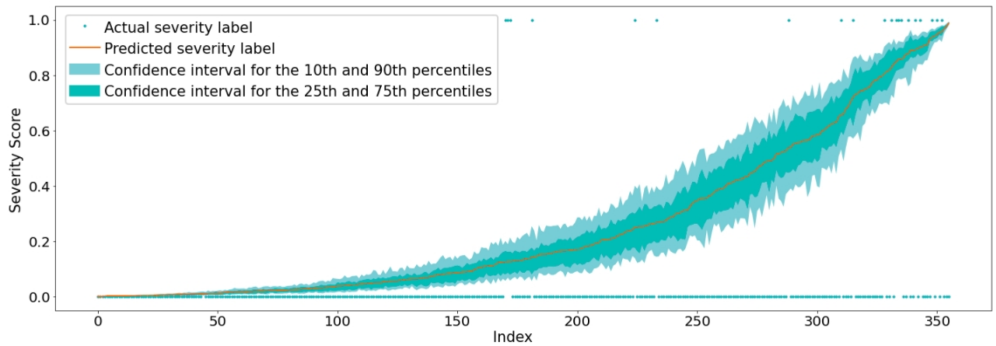


Figure 5. The predicted severity score on a selected subset of the dataset and the confidence intervals for the 10th and 90th percentiles and the 25th and 75th percentiles. The records are sorted by the severity score.

773x269mm (47 x 47 DPI)

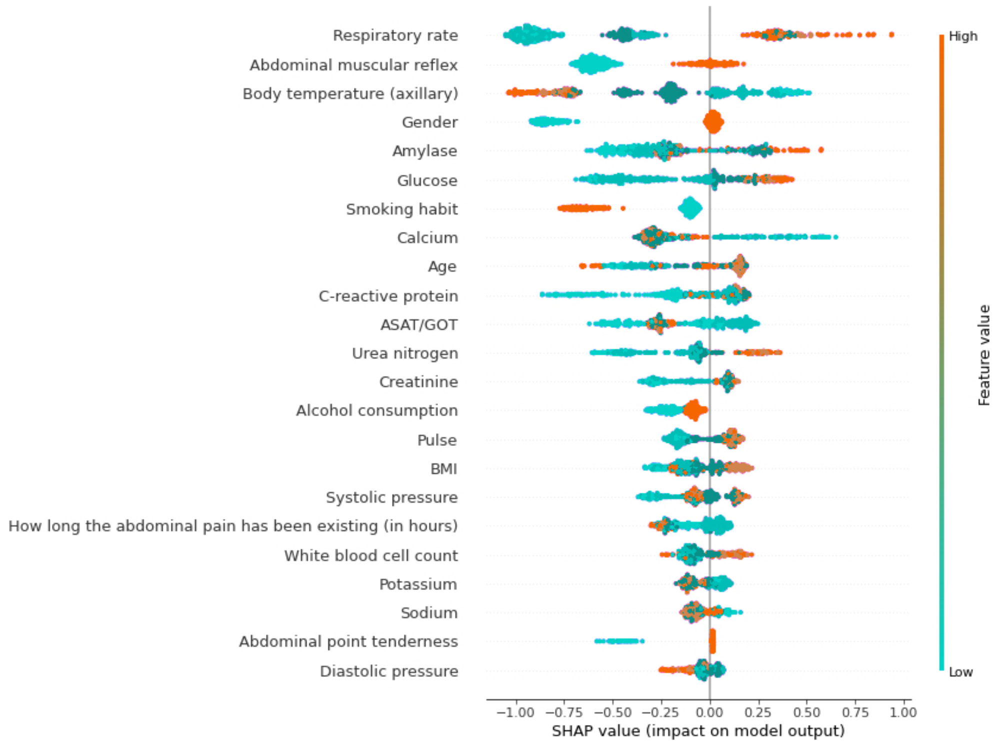


Figure 6. A summary plot of the impact of the features on the prediction (severity score) of the model. Each patient is represented by a point in each row. The colour of the points represents the relative value of the feature, and the x-position of the points is the SHAP value, i.e., the impact on the model's prediction.

454x342mm (47 x 47 DPI)

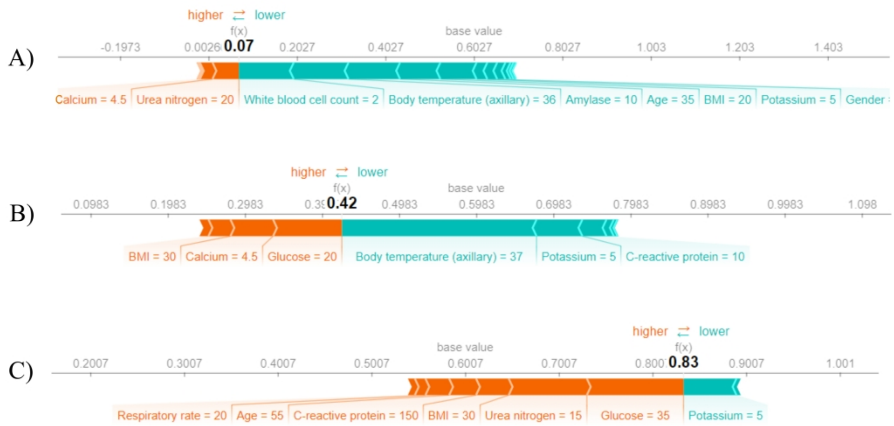


Figure 7. Three examples of the local explanation of the predictions using the SHAP values. A) Predicted mild AP. B) Predicted AP with borderline severity. C) Predicted severe AP. Factors that push the predicted score higher compared to the base value (mean prediction) are coloured orange, and those pushing lower the prediction are shown in green.

861x484mm (118 x 118 DPI)

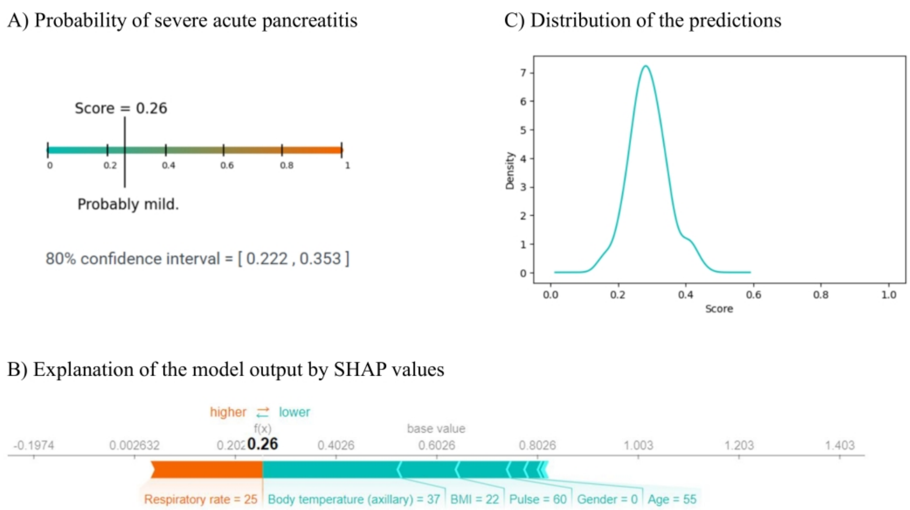


Figure 8. An example output of the web application for the following input parameters: age: 55 years, gender: 0 (woman), body mass index: 22, alcohol consumption: 1 (true), blood pressure/pulse: 140/75/60, body temperature: 37.0 °Celsius, respiratory rate: 25. A) Predicted severity score. B) Explanation of the prediction. C) The kernel density estimate plot of the distribution of the predictions.

861x484mm (118 x 118 DPI)

Supplementary material

EASY-APP: An artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis

HUNGARIAN PANCREATIC STUDY GROUP CONTRIBUTORS

Adrienn Halász[1,2], Veronika Dunás-Varga[1], Tamás Takács[3], László Czakó[3], Zoltán Szepes[3], Melania Macarie[4], Petr Pencik[5], Goran Poropat[6], Davor Stimac[6], Imanta Ozola-Zalite[7], Andrey Litvin[8], Árpád Patai[9], Kristina Zadorozhna[10], Árpád V. Patai[11], István Hritz[11], Elena Stilidi[12], János Novák[13], Masayasu Horibe[14], Georgiana Robu[15], László Lakatos[16], Ali Tüzün Ince[17], Gabriele Capurso[18], Dušan Leško[19], Dóra Illés[3], Dániel Pécsi[20], Péter Varjú[20], Klementina Ocskay[20,21], Márk Félix Juhász[20,22], Mária Földi[22,23], Alexandra Mikó[20,24], and Zsolt Szakács[20,25]

[1]Szent György Teaching Hospital of County Fejér, 1st Department of Internal Medicine, Székesfehérvár, Hungary
[2]Doctoral School of Clinical Medicine, University of Szeged, Szeged
[3]Department of Medicine, University of Szeged, Szeged, Hungary
[4]County Emergency Clinical Hospital of Târgu Mures - Gastroenterology Clinic and University of Medicine, Pharmacy, Sciences and Technology ”George Emil Palade”, Targu Mures, Romania
[5]Centrum péče o zažívací trakt, Vítkovická nemocnice a.s., Ostrava, Czech Republic
[6]Clinical Hospital Center Rijeka, Rijeka, Croatia
[7]Gastroenterology, Hepatology and Nutritional Centre, Pauls Stradins Clinical University Hospital, Riga, Latvia
[8]Immanuel Kant Baltic Federal University, Kaliningrad, Russia, Gomel Regional Clinical Hospital, Gomel, Belarus
[9]Markusovszky University Teaching Hospital, Szombathely, Hungary
[10]Bogomolets National Medical University, Kiev, Ukraine
[11]Department of Surgery, Transplantation and Gastroenterology, Semmelweis University, Budapest, Hungary
[12]Hospital of Medical Academy named after SI Georgievsky, Simferopol, Russia
[13]Pándy Kálmán Hospital of Békés County, Gyula, Hungary
[14]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan
[15]Central Military Emergency Hospital "Dr Carol Davila", Bucharest, Romania
[16]Centre of Internal Medicine, Csolnoky Ferenc Hospital, Veszprém, Hungary
[17]Hospital of Bezmialem Vakif University, School of Medicine, Istanbul, Turkey
[18]Digestive and Liver Disease Unit, S. Andrea Hospital University "Sapienza", Rome, Italy
[19]1st Department of Surgery, Pavol Jozef Šafárik University, Košice, Slovakia
[20]Institute for Translational Medicine, Medical School, University of Pécs, Pécs, Hungary
[21]Centre for Translational Medicine, Semmelweis University, Budapest, Hungary
[22]Heim Pál National Pediatric Institute
[23]Centre for Translational Medicine, Department of Medicine, University of Szeged, Szeged, Hungary
[24]Department of Medical Genetics, Medical School, University of Pécs, Pécs
[25]First Department of Medicine, Medical School, University of Pécs, Pécs, Hungary

Contributions: AH, VDV, TT, LC, ZS, MM, PP, GP, DS, IOZ, AL, ÁP, KZ, ÁVP, IH, ES, JN, MH, GR, LL, ATI, GC, and DL contributed to the data collection. DI, DP, PV, KO, MFJ, MF, AM, and ZS contributed to the assurance of data quality.

Supplementary material

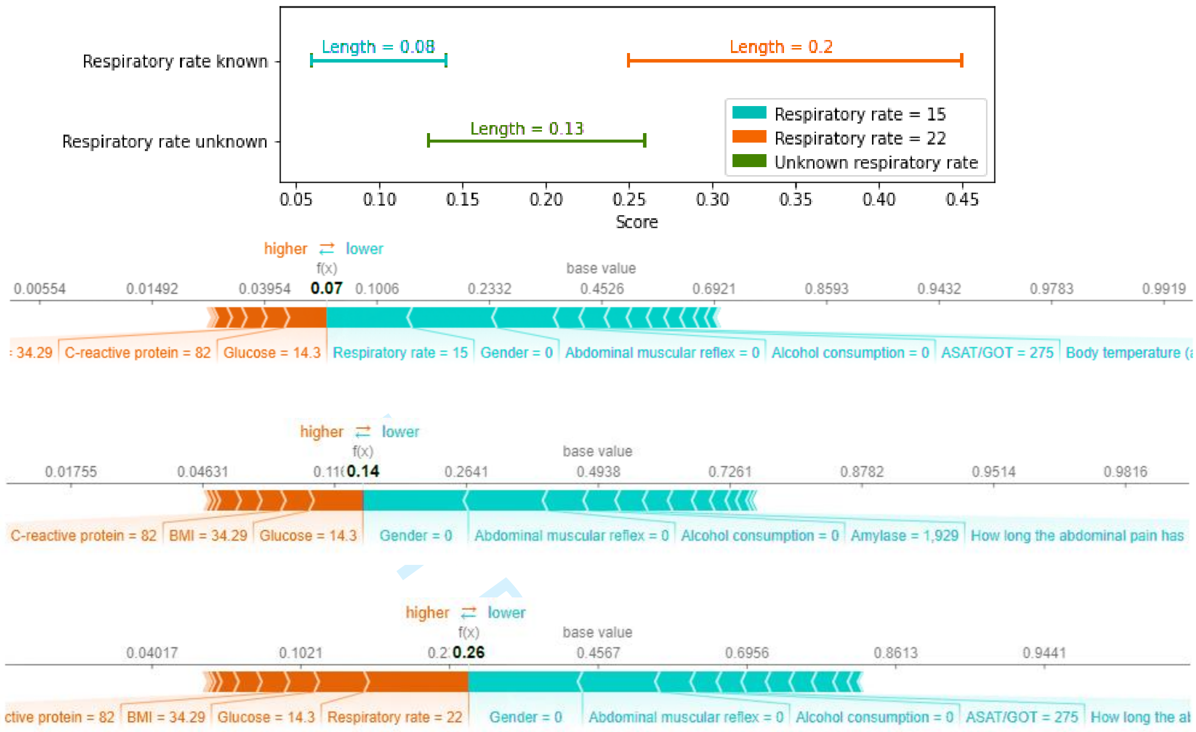
CENTRES OF THE ORIGINAL COHORT OF THE EASY STUDY

EASY study - Institutes			**Case no.**
First Department of Medicine, University of Pécs	Pécs	Hungary	584
Department of Internal Medicine, University of Debrecen	Debrecen	Hungary	149
Szent György University Teaching Hospital of County Fejér	Székesfehérvár	Hungary	93
Department of Medicine, University of Szeged	Szeged	Hungary	84
County Emergency Clinical Hospital of Târgu Mures	Targu Mures	Romania	76
Vilnius University Hospital Santariskiu Klinikos	Vilnius	Lithuania	31
General Surgery, Consorci Sanitari del Garraf	Sant Pere de Ribes	Spain	30
Helsinki University Central Hospital	Helsinki	Finland	30
Saint Luke Clinical Hospital	St. Petersburgh	Russia	30
Centrum Péče o Zažívací trakt, Vítkovická Nemocnice A.S.	Ostrava	Czech Republic	13
Clinical Hospital Centre Rijeka	Rijeka	Croatia	11
Gastroenterology, Hepatology and Nutritional Centre, Pauls Stradins Clinical University Hospital	Riga	Latvia	9
Gomel Regional Clinical Hospital	Gomel	Belarus	9
Markusovszky University Teaching Hospital	Szombathely	Hungary	9
Bogomolets National Medical University	Kiev	Ukraine	6
Second Department of Medicine, Semmelweis University	Budapest	Hungary	6
Medical Academy named after SI Georgievsky	Simferopol	Russia	5
Pándy Kálmán Hospital of County Békés	Gyula	Hungary	4
Keio University	Tokyo	Japan	2
Central Military Emergency Hospital "Dr Carol Davila"	Bucharest	Romania	1
Csolnoky Ferenc Hospital	Veszprém	Hungary	1
Jahn Ferenc South-Pest Hospital	Budapest	Hungary	1
Total			**1,184**

Supplementary table 1. Centres of the original cohort of the EASY study.

Supplementary material

FURTHER EXAMPLES OF PREDICTIONS



Supplementary figure 1. Further examples of predictions: the top plot shows the predicted severity score and its confidence interval with a low, high, and unknown value of respiratory rate, the lower figures show the corresponding local explanations.

Supplementary figure 1. shows the effect on the predicted severity score and its confidence when the value of the most important feature, the respiratory rate is low (respiratory rate = 15), high (respiratory rate = 22), and unknown (the other features are unchanged). The bottom part of the figure shows the three corresponding explanations of the predictions. From the figure, it is apparent, that the low respiratory rate pushes the predicted severity score lower (prediction = 0.07), and in this case, the confidence of the model is also higher (length of the confidence interval is 0.08). On the other hand, when the respiratory rate is relatively high, the predicted severity score also increases from 0.07 to 0.26, and the respiratory rate is 22 is the most influential negative factor that pushes the severity score lower. When the respiratory rate is unknown, both the predicted severity score and its confidence interval lies between the previous values.

Supplementary material

CHARACTERISTICS OF THE VALIDATION COHORT

INSTITUTES			No. of cases
Alicante University General Hospital, Alicante University	Alicante	Spain	1 655
Liverpool University Hospitals, University of Liverpool	Liverpool	UK	647
Hospital Val D'Hebron, University of Barcelona	Barcelona	Spain	454
The 4th Medical Clinic, "Iuliu Hatieganu" University of Medicine and Pharmacy	Cluj Napoca	Romania	408
Total			**3 164**

Supplementary table 2. Institutes of the validation cohort.

ALICANTE			Data Quality*
Demographic data			
Gender, male %	53.8%	female/male	100%
Age, mean (SD); min, max	64.5 (17.3)	[19, 100]	100%
BMI, mean (SD); min, max	27.4 (4.7)	[16.0, 53.9]	94%
Anamnestic data			
Alcoholic AP consumption, yes %	15.2%	yes/no	100%
Smoking, yes %	22.2%	yes/no	100%
Length of abdominal pain, mean (SD) in hours; min, max	no data		
Admission data			
Abdominal guarding, yes %	no data		
Abdominal tenderness, yes %	no data		
Body temperature (axillary), °C mean (SD); min, max	36.3 (0.7)	[34.0, 40.0]	98%
Systolic blood pressure (Hgmm), mean (SD); min, max	no data		
Diastolic blood pressure (Hgmm), mean (SD); min, max	no data		
Heart rate, mean (SD); min, max	82 (18)	[35, 160]	97%
Respiratory rate, mean (SD); min, max	18 (4)	[8, 40]	73%
Laboratory parameters			
Amylase, mean (SD); min, max	no data		
ASAT/GOT, mean (SD); min, max	187.6 (242.0)	[4.0, 2122.0]	82%
Serum ionized Calcium, mean (SD); min, max	2.3 (0.2)	[1.3, 4.0]	60%
C-reactive protein (mg/l), mean (SD); min, max	no data		
Creatinine, mean (SD); min, max	102.3 (59.2)	[53.4, 884.0]	56%
Glucose, mean (SD); min, max	7.7 (3.6)	[1.0, 32.6]	99%
Potassium, mean (SD); min, max	4.1 (0.6)	[2.0, 7.0]	57%
Sodium, mean (SD); min, max	138.2 (3.8)	[115.0, 152.0]	99%
Urea (carbamide), mean (SD); min, max	6.8 (3.8)	[0.5, 42.3]	99%
White blood cell count, mean (SD); min, max	11.5 (6.0)	[1.0, 38.7]	99%
Outcome			
The severity of acute pancreatitis, severe %	7.5%	non-severe/severe	100%

Supplementary table 3. Summary of the dataset from Alicante.

Supplementary material

LIVERPOOL			**Data quality***
Demographic data			
Gender, male %	48.8%	female/male	100%
Age, mean (SD); min, max	55.4 (18.1)	[18, 96]	99%
BMI, mean (SD); min, max	no data		
Anamnestic data			
Alcoholic AP consumption, yes %	no data		
Smoking, yes %	no data		
Length of abdominal pain, mean (SD) in hours; min, max	22.1 (32.7)	[1, 168]	86%
Admission data			
Abdominal guarding, yes %	no data		
Abdominal tenderness, yes %	no data		
Body temperature (axillary), °C mean (SD); min, max	36.7 (0.5)	[34.0, 41.0]	99%
Systolic blood pressure (Hgmm), mean (SD); min, max	135.1 (26.2)	[59, 224]	100%
Diastolic blood pressure (Hgmm), mean (SD); min, max	79.1 (15.9)	[21, 186]	100%
Heart rate, mean (SD); min, max	83.3 (19-9)	[45, 165]	100%
Respiratory rate, mean (SD); min, max	17.7 (3.3)	[12, 46]	99%
Laboratory parameters			
Amylase, mean (SD); min, max	1471.9 (1057.0)	[33.0, 6303.0]	91%
ASAT/GOT, mean (SD); min, max	no data		
Serum ionized Calcium, mean (SD); min, max	2.3 (0.1)	[1.4, 2.9]	62%
C-reactive protein (mg/l), mean (SD); min, max	33.4 (71.0)	[1.0, 529.0]	71%
Creatinine, mean (SD); min, max	86.0 (55.7)	[8.1, 694.0]	98%
Glucose, mean (SD); min, max	7.6 (4.2)	[2.1, 71.0]	67%
Potassium, mean (SD); min, max	4.2 (0.5)	[2.2, 7.6]	92%
Sodium, mean (SD); min, max	138.2 (3.8)	[120.0, 159.0]	98%
Urea nitrogen (carbamide), mean (SD); min, max	5.8 (4.7)	[1.1, 65.0]	98%
White blood cell count, mean (SD); min, max	13.4 (5.5)	[1.4, 68.0]	97%
Outcome			
The severity of acute pancreatitis, severe %	8.8%	non-severe/severe	100%

Supplementary table 4. Summary of the dataset from Liverpool.

Supplementary material

BARCELONA			**Data quality***
Demographic data			
Gender, male %	51.3%	female/male	100%
Age, mean (SD); min, max	64.9 (18.5)	[17, 98]	100%
BMI, mean (SD); min, max	28.2 (5.3)	[16.4, 55.3]	98%
Anamnestic data			
Alcoholic AP consumption, yes %	16.9%	yes/no	94%
Smoking, yes %	24.9%	yes/no	97%
Length of abdominal pain, mean (SD) in hours; min, max	41.4 (61.8)	[1, 360]	100%
Admission data			
Abdominal guarding, yes %	37.6%	yes/no	97%
Abdominal tenderness, yes %	7.5%	yes/no	97%
Body temperature (axillary), °C mean (SD); min, max	36.4 (0.7)	[34.0, 38.6]	99%
Systolic blood pressure (Hgmm), mean (SD); min, max	132.9 (24.6)	[60, 207]	100%
Diastolic blood pressure (Hgmm), mean (SD); min, max	72.3 (15.1)	[30, 167]	100%
Heart rate, mean (SD); min, max	82.3 (18.0)	[40, 148]	100%
Respiratory rate, mean (SD); min, max	16.5 (3.2)	[12, 40]	95%
Laboratory parameters			
Amylase, mean (SD); min, max	1219.4 (1424.9)	[6.0, 20420.0]	99%
ASAT/GOT, mean (SD); min, max	236.0 (321.9)	[8.0, 3515.0]	100%
Serum ionized Calcium, mean (SD); min, max	2.3 (0.2)	[1.3, 2.9]	98%
C-reactive protein (mg/l), mean (SD); min, max	55.7 (80.3)	[0.3, 437.7]	99%
Creatinine, mean (SD); min, max	88.4 (43.3)	[30.9, 321.8]	100%
Glucose, mean (SD); min, max	8.3 (3.6)	[3.2, 35.4]	100%
Potassium, mean (SD); min, max	3.9 (0.5)	[2.5, 5.7]	99%
Sodium, mean (SD); min, max	137.0 (3.7)	[116.3, 154.8]	100%
Urea nitrogen (carbamide), mean (SD); min, max	16.1 (9.1)	[2.2, 72.1]	100%
White blood cell count, mean (SD); min, max	13.5 (5.6)	[2.3, 45.5]	100%
Outcome			
The severity of acute pancreatitis, severe %	11.7%	non-severe/severe	100%

Supplementary table 5. Summary of the dataset from Barcelona.

Supplementary material

CLUJ NAPOCA			**Data quality***
Demographic data			
Gender, male %	51.0%	female/male	100%
Age, mean (SD); min, max	60.1 (16.5)	[21, 93]	100%
BMI, mean (SD); min, max	no data		
Anamnestic data			
Alcoholic AP consumption, yes %	18.9%	yes/no	100%
Smoking, yes %	no data		
Length of abdominal pain, mean (SD) in hours; min, max	58.0 (63.2)	[2, 504]	77%
Admission data			
Abdominal guarding, yes %	11.5%	yes/no	100%
Abdominal tenderness, yes %	92.6%	yes/no	100%
Body temperature (axillary), °C mean (SD); min, max	no data		
Systolic blood pressure (Hgmm), mean (SD); min, max	138.0 (23.7)	[60, 231]	98%
Diastolic blood pressure (Hgmm), mean (SD); min, max	77.5 (13.7)	[30, 150]	98%
Heart rate, mean (SD); min, max	81.9 (16.9)	[50, 150]	97%
Respiratory rate, mean (SD); min, max	no data		
Laboratory parameters			
Amylase, mean (SD); min, max	804.6 (844.3)	[10.5, 5349.0]	100%
ASAT/GOT, mean (SD); min, max	173.9 (259.8)	[8.0, 3481.0]	100%
Serum ionized Calcium, mean (SD); min, max	2.3 (0.2)	[1.0, 3.5]	88%
C-reactive protein (mg/l), mean (SD); min, max	82.4 (93.8)	[0.6, 523.0]	98%
Creatinine, mean (SD); min, max	96.0 (69.9)	[30.1, 534.1]	100%
Glucose, mean (SD); min, max	7.6 (3.9)	[2.6, 39.8]	99%
Potassium, mean (SD); min, max	3.9 (0.6)	[2.6, 7.2]	100%
Sodium, mean (SD); min, max	135.2 (3.5)	[124.0, 145.0]	100%
Urea nitrogen (carbamide), mean (SD); min, max	16.3 (12.1)	[1.1, 93.9]	100%
White blood cell count, mean (SD); min, max	12.8 (5.7)	[1.3, 39.4]	100%
Outcome			
The severity of acute pancreatitis, severe %	17.2%	non-severe/severe	100%

Supplementary table 6. Summary of the dataset from Cluj Napoca.

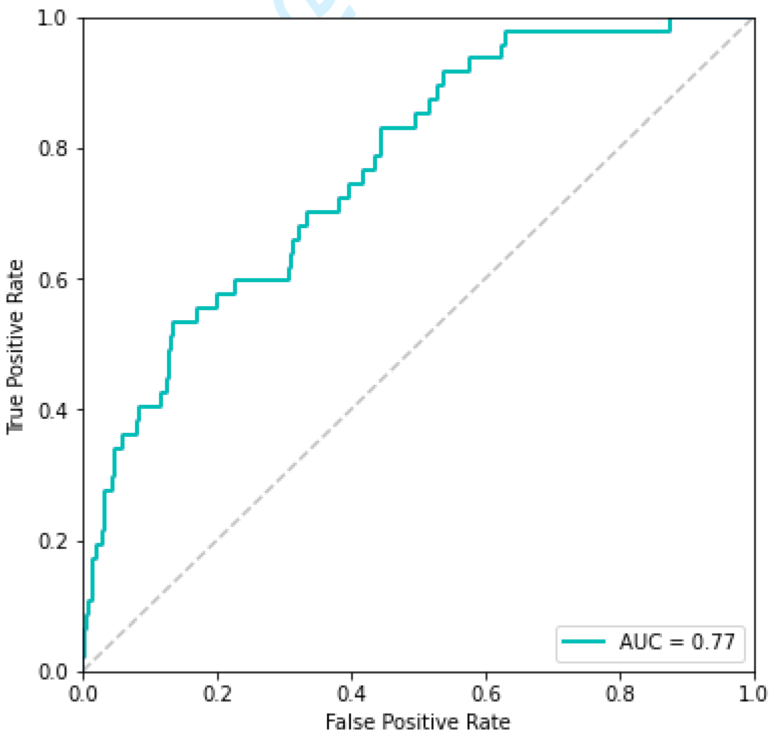
Supplementary material

VALIDATION ANALYSIS

1. Training set: Original EASY cohort, test set: data from Liverpool, Cluj Napoca, Alicante, Barcelona.

In the first part of the validation phase, we used our whole EASY database as a training set, and other different studies were used as test sets, i.e., the machine learning model was trained on the EASY cohort, and then we tested its performance on the other international datasets separately (Liverpool, Barcelona, Cluj Napoca, and Alicante). After cleaning and processing the data, we analysed them. A significant problem was that these studies were created for other purposes instead of the severity prediction of pancreatitis. Thus, many parameters were missing, which made harder the analysis, and decreased the AUC values.

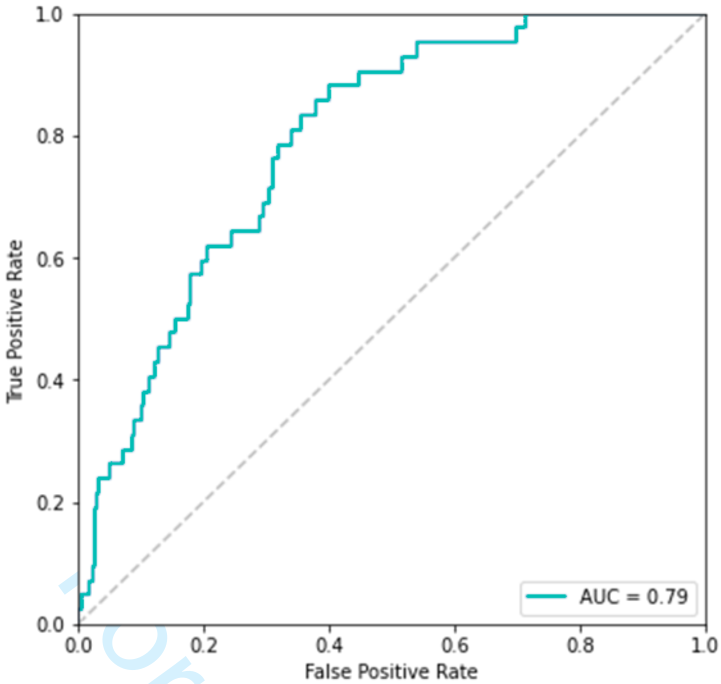
Data from Liverpool study group: In the case of this study the record of BMI, abdominal tenderness, abdominal guarding, GOT values were missing. These parameters were not recorded because of the design of this study protocol. As Supplementary figure 2. shows, the ROC curve has a lower AUC score on the Liverpool data (0.77), compared to the EASY cohort (0.81).



Supplementary figure 2. The ROC curve of the model that was trained on the EASY data and evaluated on the Liverpool data.

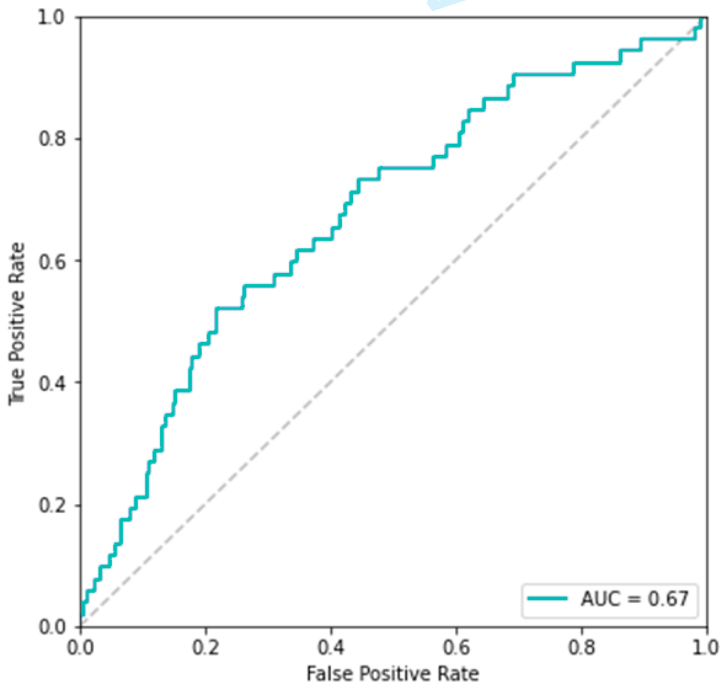
In the case of the data from Barcelona, the ROC curve has a similar AUC score (0.79) to the EASY cohort, as it can be seen in Supplementary figure 3.

Supplementary material



Supplementary figure 3. The ROC curve of the model that was trained on the EASY data and evaluated on the Barcelona data.

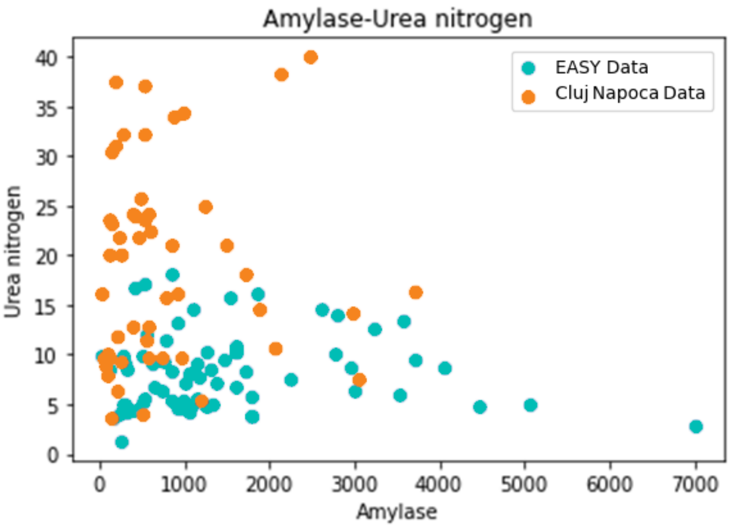
In the case of the Cluj Napoca data, for 90% of the patients, BMI, respiratory rate, body temperature, and smoking habit parameters were missing, which caused a significant problem, because BMI and body temperature are very important predictive factors of severe AP.



Supplementary figure 4. The ROC curve of the model that was trained on the EASY data and evaluated on the Cluj Napoca data.

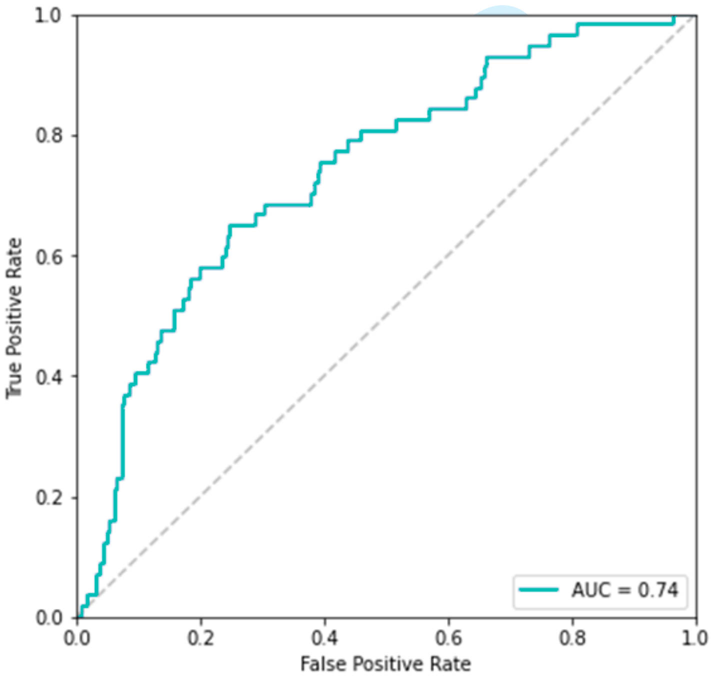
Supplementary material

That is one of the reasons why the ROC curve has a poor 0.67 AUC score (Supplementary figure 4). Another possible reason for the low AUC score, is that we have found that the joint distribution of serum carbamide and amylase is different on the EASY and the Cluj-Napoca data in the case of severe AP patients (Supplementary figure 5.). Maybe this phenomenon was caused by the different populations.



Supplementary figure 5. Joint distribution of serum carbamide and amylase in the severe cases in the EASY and Cluj Napoca dataset.

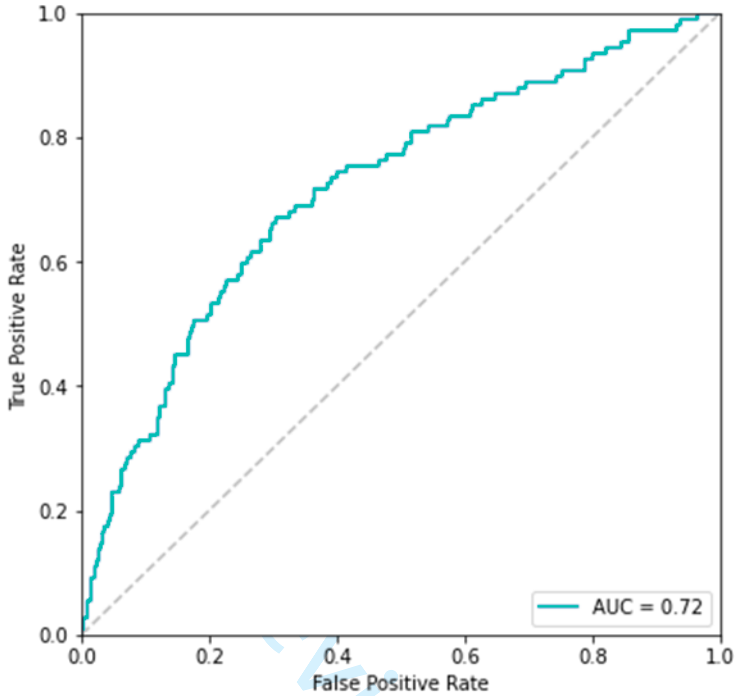
As Supplementary figure 6. illustrates, without serum amylase and carbamide parameters, the ROC curve improved significantly.



Supplementary figure 6. The ROC curve of the model that was trained on the EASY data and evaluated on the Cluj Napoca data without serum amylase and carbamide parameters.

Supplementary material

Finally, we evaluated our model (which was trained on the EASY cohort) on the Atlantis data (Alicante study group), this dataset also showed a lower AUC score (Supplementary figure 7), because of the lack of CRP, abdominal pain duration time, systolic/diastolic blood pressure, abdominal tenderness, abdominal guarding, serum amylase values.



Supplementary figure 7. The ROC curve of the model that was trained on the EASY data and evaluated on the Atlantis data.

2. Training and also test set: data from Liverpool, Cluj Napoca, Alicante, Barcelona.

Centres	Liverpool (n=647)	Barcelona (n=454)	Cluj Napoca (n=408)	Alicante (n=1655)
AUC	0.749	0.782	0.760	0.779

Supplementary table 7. Cross-validated AUC of models trained on the corresponding dataset.

Supplementary table 7. shows the performance (cross-validated AUC) of the model when we train and evaluate it separately on the other international datasets.

In what follows, we study how the performance of the model increases if the training set contains observations from different datasets. The new centres' data were divided into equally sized subsets , and then x% of these subsets were used as a training data set together with the whole EASY data. The rest of the data (the other subsets) were used for validation, i.e., to measure the performance of the model. This process is repeated in a cross-validation manner. The results are detailed in Supplementary table 8.. From the table it is apparent, that the performance of the model increases with more training data. In other words, the more data we add (from other centres) to the EASY cohort for training, the higher the AUC score of the model on the remaining test data set is.

Training set EASY+ x % of the given study	Liverpool AUC (n=647)	Barcelona AUC (n=454)	Cluj Napoca AUC (n=408)	Alicante AUC (n=1655)
0%	0.772	0.790	0.736	0.718
33%	0.773	0.792	0.780	0.764
50%	0.776	0.791	0.785	0.777
67%	0.780	0.803	0.784	0.786
80%	0.781	0.804	0.793	0.791

Supplementary table 8. Cross-validated AUC scores of the model trained on the union of the EASY data complemented with x% of the corresponding international dataset..

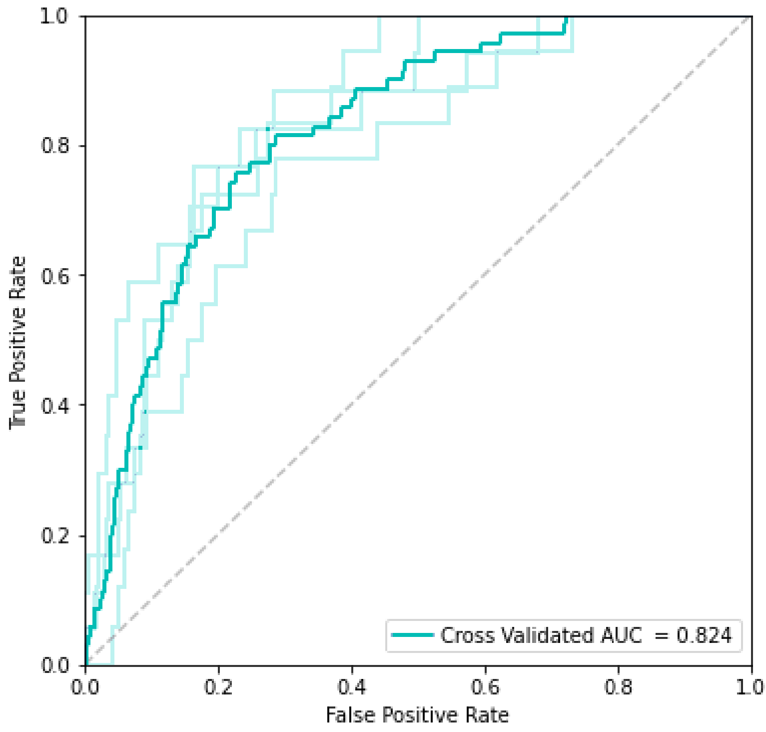
ROC curves with cross-validation.

Finally, we evaluate the model's performance using cross-validation as follows. We select one of the cohorts on which we test the model. Then we divide the data into 4 equally sized subsets. The data of the other cohorts and three subsets of the selected cohort are used for training and the remaining one subset of the selected cohort is used to measure the model's performance (AUC).

Supplementary figure 8., 9., 10., 11., and 12. show the cross-validated ROC curve and the corresponding AUC score of the model on the EASY, Liverpool, Barcelona, Cluj Napoca, and Alicante respectively. The difference between the 'A' and 'B' figures, is that in the case of the 'A' figures, the model was trained and tested separately on the different cohorts, and in the 'B' figures, the training set was supplemented with the other cohorts in the aforementioned way.

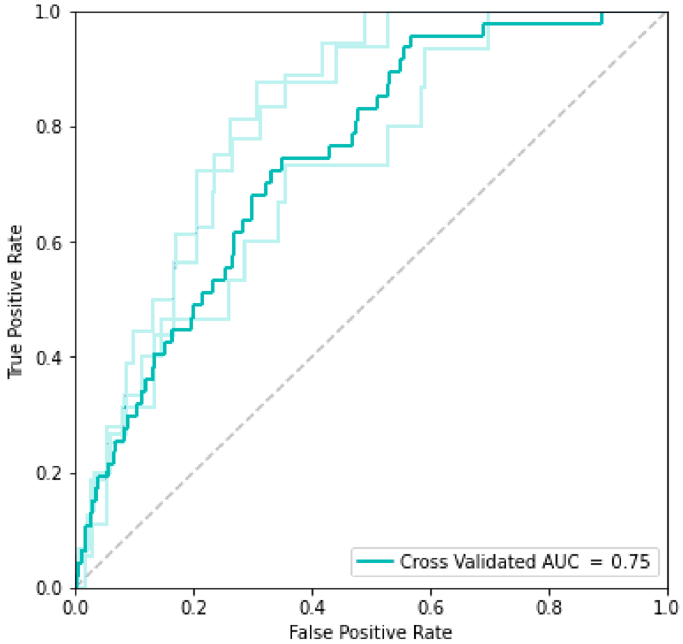
The result of the 4-fold cross-validation on the EASY cohort, where the training datasets (3 folds, i.e., 75% of the EASY data) were supplemented with the Liverpool, Cluj Napoca, Alicante, and Barcelona data.

Supplementary material

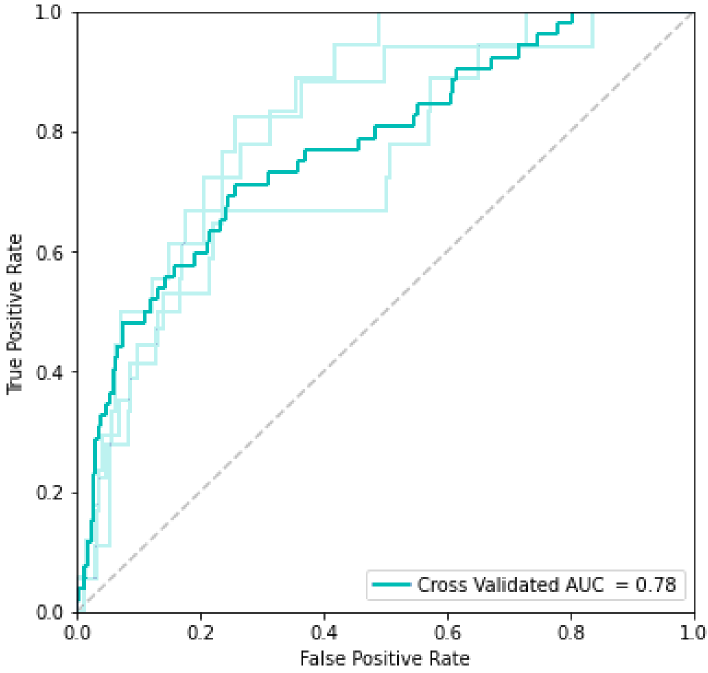


Supplementary figure 8. The result of the 4-fold cross-validation on the EASY cohort, where the training datasets (75% of the EASY data) were supplemented with the Liverpool, Cluj Napoca, Alicante, and Barcelona data.

Supplementary material

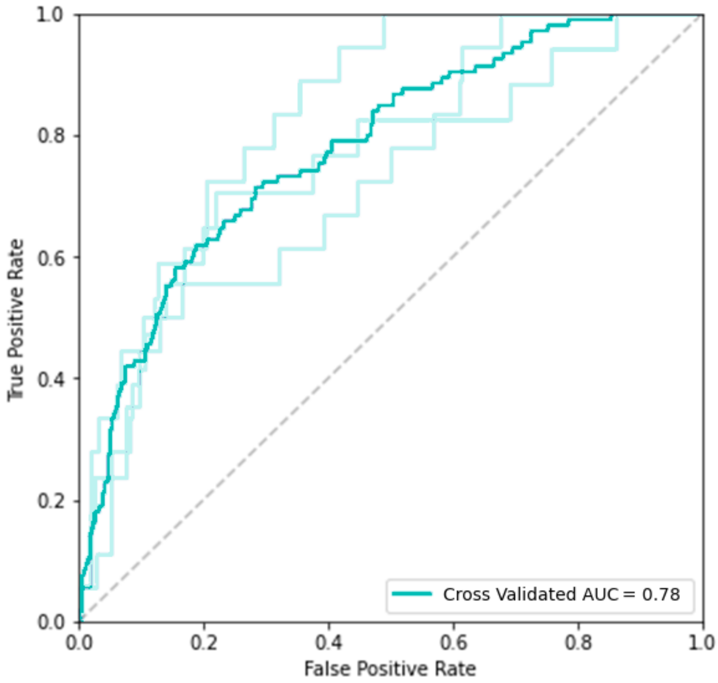


Supplementary figure 9.A. The result of the 4-fold cross-validation on the Liverpool dataset.

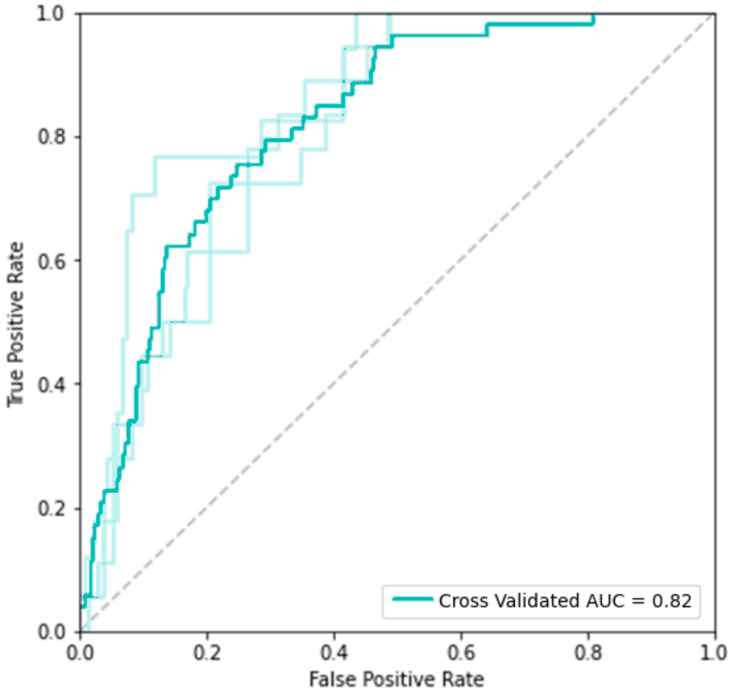


Supplementary figure 9.B. The result of the 4-fold cross-validation on the Liverpool cohort, where the training datasets (75% of the Liverpool data) were supplemented with the EASY, Cluj Napoca, Alicante, and Barcelona data.

Supplementary material

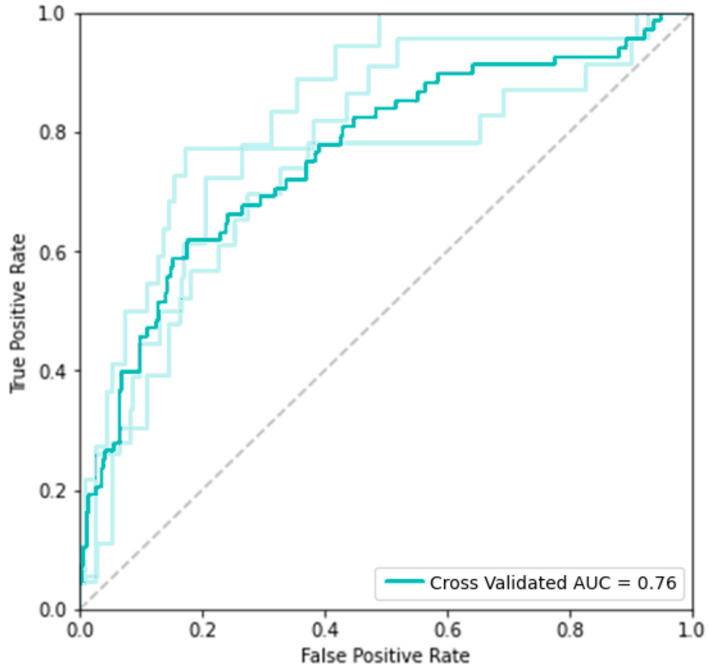


Supplementary figure 10.A. The result of the 4-fold cross-validation on the Barcelona dataset.

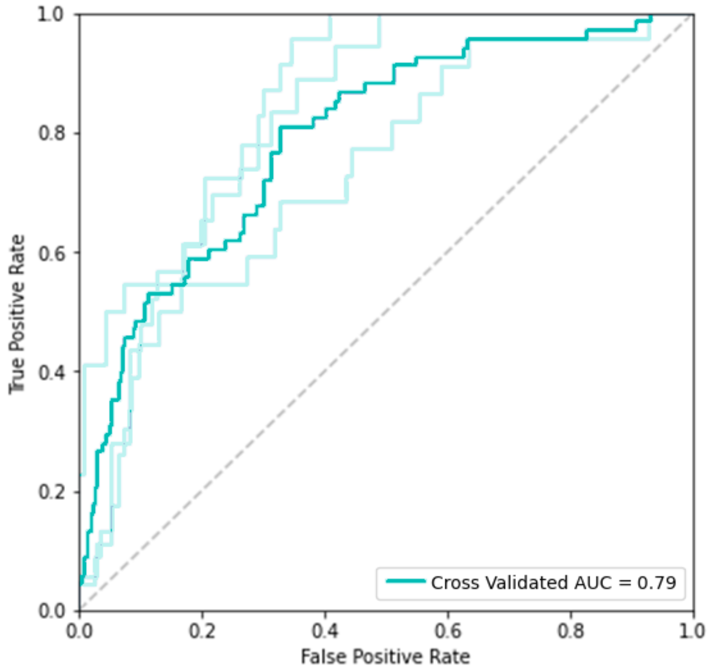


Supplementary figure 10.B. The result of the 4-fold cross-validation on the Barcelona cohort, where the training datasets (75% of the Barcelona data) were supplemented with the Cluj Napoca, Liverpool, Alicante, and EASY data.

Supplementary material

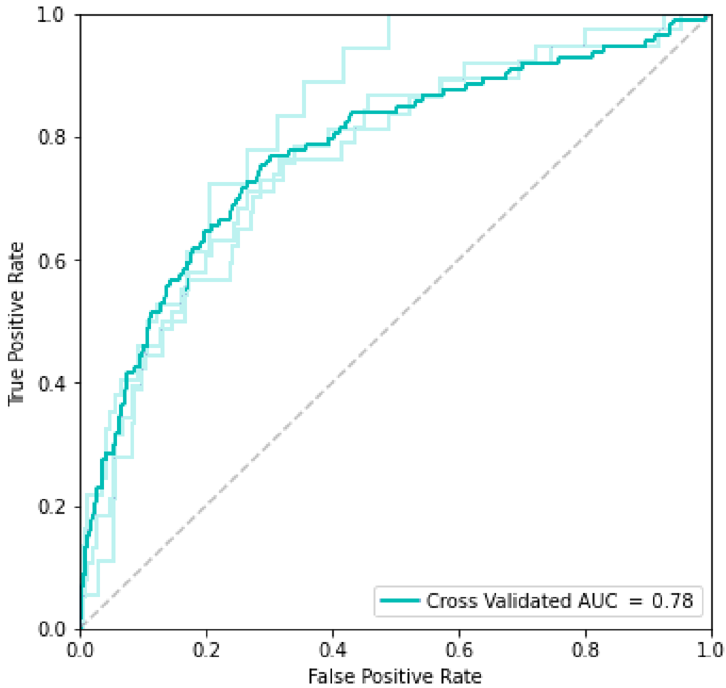


Supplementary figure 11.A. The result of the 4-fold cross-validation on the Cluj Napoca dataset.

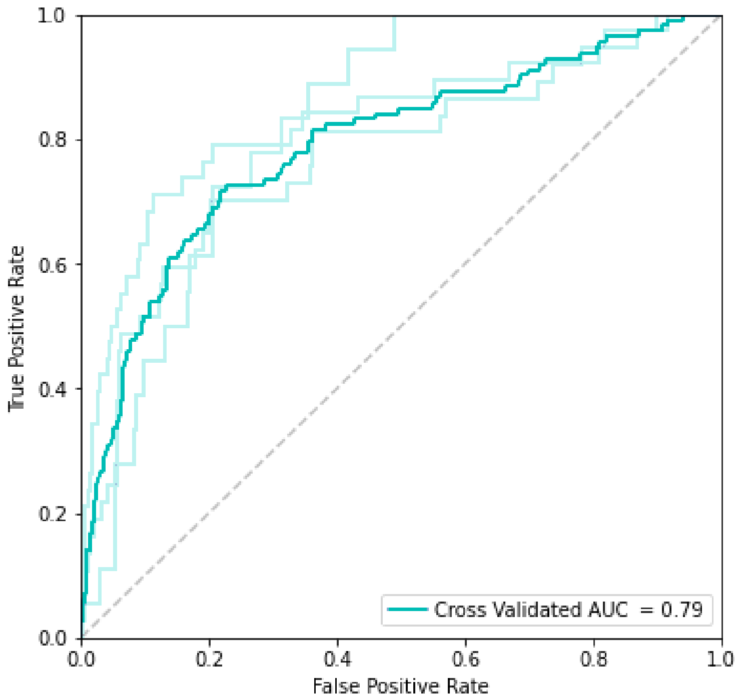


Supplementary figure 11.B. The result of the 4-fold cross-validation on the Cluj Napoca cohort, where the training datasets (75% of the Cluj Napoca data) were supplemented with the Barcelona, Liverpool, Alicante, and EASY data.

Supplementary material



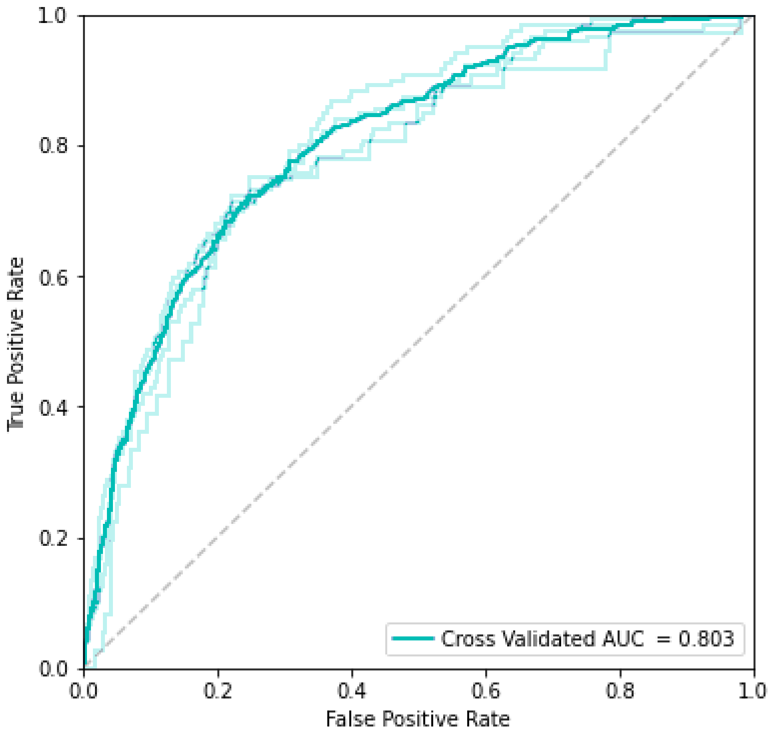
Supplementary figure 12.A. The result of the 4-fold cross-validation on the Alicante dataset.



Supplementary figure 12.B. The result of the 4-fold cross-validation on the Alicante cohort, where the training datasets (75% of the Alicante data) were supplemented with the Barcelona, Liverpool, Cluj Napoca, and EASY data.

3. **The model is cross-validated on the union of the five different cohorts (Alicante, Barcelona, Cluj Napoca, EASY, Liverpool).**

The results of the 4-fold cross-validation on the whole dataset (union of all data sets) can be seen on Supplementary figure 13. The figure suggests that the performance of the model improves when we train the model on the union of all the datasets. In this case the cross-validated AUC score is 0.803, however, when we train the model separately then the average cross-validated AUC score is only 0.776.

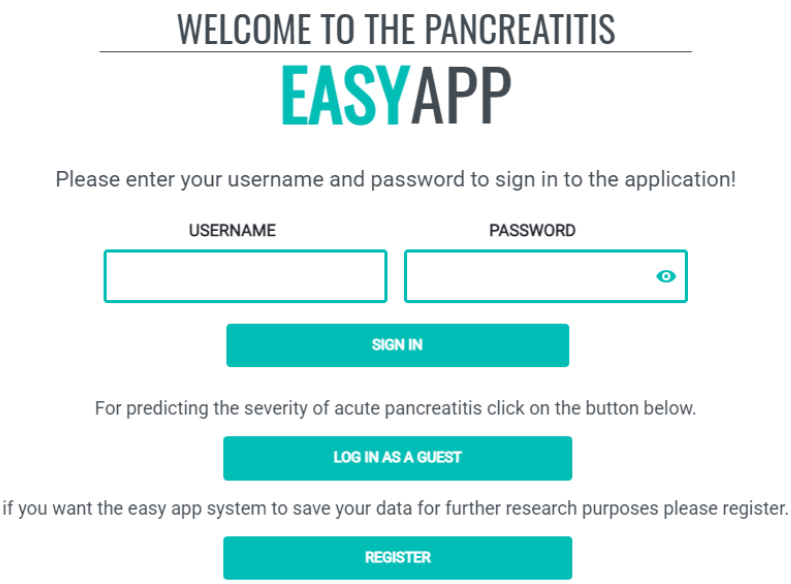


Supplementary figure 13. The result of the 4-fold cross-validation on the whole dataset together.

Supplementary material

EASY-APP AVAILABILITY AND REGISTRATION

The web application is available at: http://easy-app.org/



The model calculates a numerical probability value between 0 and 1. The higher the number, the higher the risk for severe acute pancreatitis. Together with the numerical values, a textual interpretation is given as well. Upon request, the application provides the confidence interval in addition to the numerical value. For educational purposes, with the help of the SHAP values, the explanation of the prediction highlighting the key factors affecting the severity of AP is shown too. Built-in validations filter out invalid values.

The application can be used in two ways. The prediction is possible without registration, however, in the case of registration, the data and the prediction will be stored and the given prediction contributes to the development of the model.

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	4
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	4
	5b	D;V	Describe eligibility criteria for participants.	4
	5c	D;V	Give details of treatments received, if relevant.	4
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	not applicable
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	5
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	not applicable
Sample size	8	D;V	Explain how the study size was arrived at.	8
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	6
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6-7
	10c	V	For validation, describe how the predictions were calculated.	9
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	9-11
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	14+ supplementary
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	not applicable
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	14+ supplementary
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8-9
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8-9
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	8-9
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	8-9
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	not applicable
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	not applicable
	15b	D	Explain how to the use the prediction model.	13-14
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	9-10
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	14+ supplementary
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	15-16
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	14+ supplementary
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	14-17
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	13-16
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	13-16
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	17

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.