Computers in Biology and Medicine 163 (2023) 107146

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Automated identification of patient subgroups: A case-study on mortality of COVID-19 patients admitted to the ICU

I. Vagliano [a,b,*], M.Y. Kingma [a], D.A. Dongelmans [b,c,d], D.W. de Lange [d,e], N.F. de Keizer [a,b,d], M.C. Schut [a,b,f], On behalf of thethe Dutch COVID-19 ICU Research Consortium

[a] Dept. of Medical Informatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 15, 1105 AZ, Amsterdam, the Netherlands
[b] Amsterdam Public Health (APH), Postbus 7057, 1007 MB, Amsterdam, the Netherlands
[c] Dept. of Intensive Care Medicine, Amsterdam UMC, University of Amsterdam, Meibergdreef 15, 1105 AZ, Amsterdam, the Netherlands
[d] National Intensive Care Evaluation (NICE) Foundation, Postbus 23640, 1100 EC, Amsterdam, the Netherlands
[e] Dept. of Intensive Care, University Medical Center Utrecht, University Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands
[f] Dept. of Clinical Chemistry, Amsterdam UMC, Vrije Universiteit Amsterdam, Meibergdreef 15, 1105 AZ, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Subgroup discovery
Machine learning
COVID-19
In-hospital mortality
Intensive care
Data registry

ABSTRACT

Background: – Subgroup discovery (SGD) is the automated splitting of the data into complex subgroups. Various SGD methods have been applied to the medical domain, but none have been extensively evaluated. We assess the numerical and clinical quality of SGD methods.
Method: – We applied the improved Subgroup Set Discovery (SSD++), Patient Rule Induction Method (PRIM) and APRIORI – Subgroup Discovery (APRIORI-SD) algorithms to obtain patient subgroups on observational data of 14,548 COVID-19 patients admitted to 73 Dutch intensive care units. Hospital mortality was the clinical outcome. Numerical significance of the subgroups was assessed with information-theoretic measures. Clinical significance of the subgroups was assessed by comparing variable importance on population and subgroup levels and by expert evaluation.
Results: – The tested algorithms varied widely in the total number of discovered subgroups (5-62), the number of selected variables, and the predictive value of the subgroups. Qualitative assessment showed that the found subgroups make clinical sense. SSD++ found most subgroups (n = 62), which added predictive value and generally showed high potential for clinical use. APRIORI-SD and PRIM found fewer subgroups (n = 5 and 6), which did not add predictive value and were clinically less relevant.
Conclusion: – Automated SGD methods find clinical subgroups that are relevant when assessed quantitatively (yield added predictive value) and qualitatively (intensivists consider the subgroups significant). Different methods yield different subgroups with varying degrees of predictive performance and clinical quality. External validation is needed to generalize the results to other populations and future research should explore which algorithm performs best in other settings.

1. Introduction

In clinical research, subgroup analyses involve splitting all patients into subgroups, often as a means to make heterogeneous populations more homogeneous, or to answer specific questions about particular patient groups, types of intervention or types of study [1]. Such analyses can have drawbacks, namely (1) groups are defined manually by the researcher resulting in potentially suboptimal groups, and (2) groups can be simple, i.e., based on single variable (e.g., sex) and/or single thresholds (e.g., men versus women or under versus above 67 years). These drawbacks can be resolved by subgroup discovery (SGD) methods that aim to discover patterns in the form of rules induced from labelled data [2]. In the context of clinical subgroup analysis, SGD means the *automated* splitting of the data into *complex* subgroups, i.e., based on multiple variables and/or multiple thresholds.

Various SGD methods exist, e.g., APRIORI – Subgroup Discovery (APRIORI-SD), CN2 – Subgroup discovery (CN2-SD), Diverse Subgroup Set Discovery (DSSD) and Patient Rule Induction Method (PRIM) [3], as

* Corresponding author. Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health research institute, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands.
E-mail address: i.vagliano@amsterdamumc.nl (I. Vagliano).

https://doi.org/10.1016/j.compbiomed.2023.107146
Received 13 February 2023; Received in revised form 31 May 2023; Accepted 6 June 2023
Available online 15 June 2023

well as the improved Subgroup Set Discovery (SSD++) [4]. These algorithms discover subgroups that are represented as combinations of constraints on the variables (e.g., age >= 25 and BMI <19), which can also be interpreted as clinical rules. Typically, SGD methods differ from each other in the type of subgroup searching and selection (i.e., *exhaustive*: looks at all possible subgroups given the patient population, which requires large amounts of computation; or *heuristic*: finds subgroups faster and more efficient, but sacrifices optimality, accuracy, precision or completeness for speed, i.e., lower run time) and which quality measures are used for searching, e.g., unusualness, coverage, redundancy, and novelty, also known as weighted relative accuracy (WRAcc). These same quality measures are also used to assess the numerical significance of found subgroups. The clinical significance of subgroups can also be assessed by determining whether the subgroups are clinically relevant. To do this, on the one hand, we can consider variable importance (i.e., affect the overall risk prediction. On the other hand, clinicians could assess the subgroups (i.e., the rules describing the subgroups) to determine clinical relevance.

Multiple SGD methods have been applied to the medical domain [5–11], as well as specifically to the intensive care [12,13] but, to the best of our knowledge, not to COVID-19 patients. Furthermore, none of these studies extensively evaluated different SGD methods and the discovered subgroups. These studies provide either quality measures [5,7,11,12], predictive power of the discovered subgroups [6,13], or simply applied SGD methods to their problem and only qualitatively assessed the results of applying the method in terms of meaningful insight on their data and problem [8–10]. In contrast, our study assesses SGD methods and the discovered subgroups in terms of both quality measures and predictive power as well as it provides clinical validation.

This study proposes a new approach to systematically assess the numerical and clinical quality of automated patient subgroup discovery methods. Such an assessment is informative for clinicians that consider using SGD to perform complex subgroup analysis as to which SGD method is best applicable. SGD can pave the way to personalized medicine, and our approach can ease the implementation of SGD in clinical decision support systems. As a second contribution, we provide a case study on the prediction of hospital mortality in a registry cohort of ICU-admitted COVID-19 patients. We identified the subgroups that make clinical sense and there is much potential in using these subgroups in an automated way, for example for flagging, or as clinical decision rules.

The paper is organized as follow: Section 2 introduces the patient population, the SGD methods used, and their evaluation; Section 3 presents the discovered SGD groups and the evaluation results; Section 4 discusses our results; Section 5 concludes the paper.

2. Related work

Various subgroup discovery methods, e.g. Refs. [5–11], have been applied to the medical domain. Recent studies use subgroup discovery to identify subgroups of cancer patients [14,15], identify predictive factors for diabetic ketoacidosis [16], or discover subgroups of patients undergoing transcatheter aortic valve implantation with high model prediction error and their distribution over the centres [17]. Gamberger et al. [11] demonstrated the applicability of SGD analysis for in brain ischaemia. Abu-Hanna et al. [12] compared the established algorithms Classification and Regression Trees and PRIM in an SGD task on a large real-world high-dimensional ICU database. Nannings et al. [13] applied the PRIM to identify very elderly ICU patients at high risk of mortality and compared the results with those of a conventional logistic regression model. SGD has been used to find disease markers from gene expression data [18]. Techniques other than SGD are also used for identifying patient subgroups, including clustering [19–22], latent profile analyses [23], or a combination of clustering and subgroups discovery [24]. Subgroup discovery was also used to assess personalized treatment effects in order to identify patient subgroups that react exceptionally bad or well to treatment [25]. Multi-omics Clustering Variational Autoencoders (MCluster-VAEs) was used to extract representations on multi-omics data to discover cancer subtypes [26]. Risk profiles for negative and positive COVID-19 hospitalized patients were identified through partition around medoids clustering [27]. However, none of these studies did extensively consider the evaluation of different SGD methods and the discovered subgroups. We considered the subgroups defined by the discovered rules and we compared these with individual variables (estimated coefficients of a linear regression) to assess their predictive performance and redundancy.

3. Material and methods

3.1. Data

This study used prospectively collected data on all patients admitted between February 21st, 2020 and May 24th, 2022 with confirmed COVID-19 to a Dutch ICU extracted from the Dutch National Intensive Care Evaluation (NICE) registry. The NICE dataset contains, amongst other items, demographic data, minimum and maximum values of laboratory and monitor data in the first 24 hours of ICU stays, diagnoses (reason for admission as well as comorbidities), information on ICU admissions, i.e., hospital length of stay before ICU admission and referring specialism, ICU as well as hospital length of stay, and ICU as well as in-hospital mortality data [28]. Data is collected in a standardized manner according to strict definitions and stringent checks ensure high data quality [29]. The outcome variable was in-hospital mortality.

After variable selection (see Section 3.3), the used data consisted of about 60 variables. A total of 14,548 confirmed COVID-19 patients were included, of which 4000 patients (27.5%) died during their hospital stay. Survivors were significantly younger (59.5 vs 68.4 years old, $p < 0.001$), more often females (33.3% vs 27.3%, $p < 0.001$) and with slightly higher body mass index (29.8 vs 28.9, $p < 0.001$) than non-survivors. Table 1 and Table S3 show the descriptive summary statistics of the patient population.

3.2. Patient inclusion

Patients were considered to have COVID-19 when the RT-PCR of their respiratory secretions was positive for SARS-CoV-2. Surgery patients were excluded as they are typically admitted patients with COVID-19 rather than patients admitted because of COVID-19.

3.3. Analyses

Preprocessing – included the handling of missing data and variable selection. Missing values were imputed by using the multiple imputation by chained equations (MICE) [30]. Variables with only one unique value (or almost, >= 99% frequency) were excluded.

Patient subgroups – were obtained by application of selected heuristic (SSD++ [4], PRIM [31]) and exhaustive (APRIORI-SD [32]) algorithms[1]. The algorithms were selected to form a diverse mix of algorithms based on association rules (APRIORI-SD), decision trees (PRIM) and inductive inference (SSD++). Per algorithm, we interpreted the subgroups independently of each other: a patient can belong to one or more subgroups, by definition of adherence to the subgroup's conditions: if the conditions fit the patient, it belongs to the subgroup.

Model optimization – the three subgroup algorithms use parameters to control the learning process, called *hyperparameters*. These parameters need to be set such that the algorithm performs optimal. We did this optimization as follows. For APRIORI-SD, we performed a grid search for the *number of subgroups* (5, 10, 25, 50, 75, 100) and the *maximum selector*

[1] APRIORI-SD can also be considered heuristic depending on whether the characterization of its search is based on the rule generation or on the rule post-processing.

Table 1
Descriptive statistics of the population, stratified by hospital mortality.

		Overall	Survivors	Non-surviros	Missing
n		14548	10548	4000	
Age, mean (SD)		62.0 (12.3)	59.5 (12.5)	68.4 (9.1)	0
Body mass index, mean (SD)		29.5 (5.7)	29.8 (5.7)	28.9 (5.5)	303
Gender, n (%)	Female	4603 (31.6)	3513 (33.3)	1090 (27.3)	0
Origin of admission, n (%)	General ward of the same hospital	9707 (67.0)	7089 (67.5)	2618 (65.7)	57
	Emergency room of the same hospital	4088 (28.2)	2897 (27.6)	1191 (29.9)	
	CCU/IC of another hospital	359 (2.5)	262 (2.5)	97 (2.4)	
	CCU/IC of the same hospital	82 (0.6)	48 (0.5)	34 (0.9)	
	Emergency room of another hospital	80 (0.6)	65 (0.6)	15 (0.4)	
	Home	75 (0.5)	62 (0.6)	13 (0.3)	
	General ward of another hospital	59 (0.4)	48 (0.5)	11 (0.3)	
	Special/ Medium care of the same hospital	27 (0.2)	21 (0.2)	6 (0.2)	
	Other	8 (0.1)	7 (0.1)	1 (0.0)	
	CCU/IC other location of same hospital, transport by ambulance	4 (0.0)	3 (0.0)	1 (0.0)	
	Recovery of the same hospital	1 (0.0)	1 (0.0)		
	Special/ Medium care of another hospital	1 (0.0)	1 (0.0)		
Hospital length of stay before the ICU admission, mean (SD)		2.6 (3.4)	2.5 (3.2)	2.9 (4.0)	6
Wave of infection, n (%)	1 (from 2020-02-01 to 2020-05-15)	2258 (15.5)	1560 (14.8)	698 (17.4)	0
	Patients in between waves 1 and 2	389 (2.7)	288 (2.7)	101 (2.5)	
	2 (from 2020-10-01 to 2020-11-30)	1754 (12.1)	1183 (11.2)	571 (14.3)	
	3 (from 2020-12-01 to 2021-01-31)	1932 (13.3)	1329 (12.6)	603 (15.1)	
	4 (2021-02-01 to 2021-06-30)	4121 (28.3)	3152 (29.9)	969 (24.2)	
	5 (from 2021-07-01 to 2021-09-30)	867 (6.0)	698 (6.6)	169 (4.2)	
	6 (from 2021-10-01 to 2022-05-24)	3227 (22.2)	2338 (22.2)	889 (22.2)	
Comorbidities					
Acute renal failure, n (%)		1051 (7.2)	489 (4.6)	562 (14.1)	0
Chronic cardiovascular insufficiency, n (%)		229 (1.6)	112 (1.1)	117 (2.9)	0
Chronic renal insufficiency, n (%)		654 (4.5)	304 (2.9)	350 (8.8)	0
Chronic Obstructive Pulmonary Disease, n (%)		1364 (9.4)	840 (8.0)	524 (13.1)	0
Chronic respiratory insufficiency, n (%)		649 (4.5)	399 (3.8)	250 (6.2)	0
Diabetes, n (%)		3242 (22.3)	2166 (20.5)	1076 (26.9)	0
Haematological malignancy, n (%)		308 (2.1)	151 (1.4)	157 (3.9)	0
Immunological insufficiency, n (%)		1504 (10.3)	872 (8.3)	632 (15.8)	0
Number of chronic comorbidities, mean (SD)	0	10,781 (74.1)	8308 (78.8)	2473 (61.8)	0
	1	2885 (19.8)	1825 (17.3)	1060 (26.5)	
	2	805 (5.5)	388 (3.7)	417 (10.4)	
	3	73 (0.5)	26 (0.2)	47 (1.2)	
	4	4 (0.0)	1 (0.0)	3 (0.1)	

SD stands for standard deviation, CCU for Coronary Care Unit, and IC for intensive care.

depth (5-10). For PRIM, we performed a grid search for α (the degree of patience when looking for a sub-optimal solution) and β (the minimum size of the boxes found) with values 0.03, 0.04, 0.05, 0.06, 0.08, 0.1. For SSD++, we performed a grid search for the *maximum selector depth* (5-10) and *beam,* i.e., the pre-defined number of best partial solutions taken as candidates (25, 50, 100).

Numerical significance – of the obtained subgroups was evaluated by means of (1a) information-theoretic quality in terms of coverage, support, rule length, significance, novelty (WRAcc), confidence and redundancy (see Table S2 and [33] for their definition as well as Appendix A for an example of these measures), and (1b) formal evaluation of the benefit of subgroups for prediction. For the latter, we inspected whether it pays to increase the complexity of a prediction model by including subgroup indicator variables in order to improve prediction of the outcome. To this end, a logistic regression model was created with a backward stepwise variable selection model based on the Akaike information criterion (AIC) with the patient variables plus the indicators of the discovered subgroups [34]. The subgroup indicator variables evaluate to TRUE if and only if a patient belongs to the particular subgroup. We then inspected whether subgroup indicators were selected by the selection process. Also, we statistically tested, at the $p = 0.05$ level, whether to reject the hypothesis that the subgroups are redundant with a log likelihood ratio (ANOVA) test. For each individual subgroup indicator, we compared a logistic regression model with only the patient variables with a model that also included the subgroup indicator. Additionally, we did an ANOVA test comparing models with patient variables without subgroups to models with patient variables and all subgroup indicators.

Clinical significance – of the subgroups was evaluated by means of (2a) comparative analysis of the rule descriptions and a regression model, and (2b) expert opinion. For 2a, we informally compared the description of the obtained subgroups with the coefficients of a linear regression (LinR) model fit on hospital mortality (dichotomous outcome

was made continuous to provide more model flexibility). For the LinR model, we did backward stepwise variable selection, which was based on the Akaike information criterion (AIC). For 2b, we put forward the found subgroups (i.e., the rules describing the subgroups) to two intensivists (DD, DdL) with over 20 years of clinical expertise and asked them to evaluate, independent of each other, the rules as fit or unfit for the specific purpose of use by intensivists for triage on ICU admission of COVID-19 patients. If a rule was considered unfit, an explanation was asked for the evaluation. The form used for evaluation is available in Appendix B.

3.4. Statistical analysis

All the analyses were performed using Python v3.6 and R version 3.5.1 x64 with publicly available software packages. Notably, our implementation of APRIORI-SD is based on pysubgroup (https://pysubgroup.readthedocs.io) [35], PRIM is based on a publicly-available python implementation of PRIM (https://github.com/martinsps/PRIM), and SSD++ is based on the *SSDpp-numeric* (https://github.com/HMProenca/SSDpp-numeric). For the reporting of this study, we followed the TRIPOD statement (https://www.equator-network.org/reporting-guidelines/tripod-statement/).

4. Results

4.1. Subgroups

Table S4 describes the subgroups that were discovered with each of the SGD methods, and information-theoretic quality metrics are provided for each subgroup. The discovered subgroups vary largely between the three methods. Firstly, they differ in terms of the number of subgroups (APRIORI-SD: 5, PRIM: 6, SSD++: 62). Secondly, the subgroups themselves also differ. In PRIM and APRIORI-SD, subgroups mostly concern a small number of variables (age, haematological malignancy, chronic cardiovascular insufficiency and chronic respiratory insufficiency, cardiopulmonary resuscitation, for APRIORI-SD; age, number of chronic comorbidities, lowest bicarbonate, referring specialism, origin of admission, gender, wave of infection, highest serum urea, lowest thrombocytes, lowest creatinine, lowest systolic blood pressure, for PRIM). SSD++ has discovered most subgroups with highest variation in terms of variables per subgroup and total number of variables.

4.2. Evaluation – information-theoretic quality

Table 2 shows the results of each method in terms of information-theoretic quality metrics. We observe large differences between the methods for *coverage* (average highest for APRIORI-SD, overall for SSD++) and *significance* (highest for APRIORI-SD, meaning that its groups have higher interest). The findings on the discovered subgroups (Table S4) are included in the metrics with the measured *number of subgroups* and their *average length* (i.e., number of variables).

4.3. Evaluation – predictive value

Table 3 shows the predictive performance of the discovered subgroups in terms of (a) whether a subgroup was selected in the variable selection with stepwise regression (with and without clinical variables), and (b) log likelihood ratio tests. These results show that the majority of the subgroups survived backward selection (but only about half for PRIM and APRIORI-SD when using subgroups together with clinical variables), which is indicative of additional predictive value over the patient variables. The log likelihood ratio tests also show significant added predictive value of the subgroups discovered by PRIM and SSD++, however not for APRIORI-SD.

4.4. Evaluation – clinical significance

Fig. 1 summarizes the clinical significance of the discovered subgroups. For each method, the agreement shows the number of subgroups (with respect to the number of discovered subgroups) on which the clinicians agreed on whether the group was clinically relevant or not. The fit outlines the number of subgroups which were considered clinically relevant with respect to the number of subgroups for which there was agreement. The average fit averages the number of subgroups judged clinically relevant by each clinician and shows it with respect to the number of discovered subgroups. For one subgroup identified by SDD++, one clinician was undecided and it was counted as not clinically relevant (unfit). Overall, the intensivists found the majority of subgroups (n = 66, 91%) fitting for triage on ICU admission of COVID-19 patients. APRIORI-SD resulted in 5 out of 5 fitting subgroups for both intensivists; SSD++ in the same 58 out of 62 for both intensivists; from the subgroups discovered with PRIM, only one subgroup was considered fit by both intensivists. For APRIORI-SD and SSD++, the agreement was high (5 and 59 same ratings, respectively); ratings on PRIM were less homogenous, there was agreement on only two subgroups.

When asked for an overall evaluation of the subgroups, both intensivists mentioned it was interesting for SDD++ to discover subgroups not only with a very high probability of dying, but also subgroups with very low mortality probabilities. However, given that the SDD++ groups were relatively small, it was suggested that performance metrics would be provided of the groups (we computed the performance metrics per group, but these were not shown to clinicians during their evaluation). Rules in the form of "not equal to", which were common in PRIM, were considered unintuitive. The APRIORI-SD subgroups were not considered very distinctive since the length of stay is long, but the mortality is around 0.5. Concerning the used variables, the *origin of admission,* i.e., the location just before the ICU admission (home, emergency room, ward, other hospital, etc.), was considered vague and not so clinically meaningful. Also, the variable indicating the *infection wave* was considered not useable in practice since new patients cannot belong to past infection waves.

5. Discussion

In this study, we performed quantitative and qualitative analyses of patient subgroups that were discovered automatically and have the form of rules (conditions) on the patient features.

Table 2
Information-theoretic evaluation of SGD models.

Method	Coverage		Support		Significance	Accuracy (WRAcc)		Confidence	Redundancy	Number of SG	Length of SG
	Average	Overall	Average	Overall	Average	Average	Maximum	Average	(%)		
APRIORI-SD	**0.20**	0.21	**0.10**	0.10	**608.43**	**0.0434**	**0.05**	**0.49**	20.00	5	1.80
PRIM	0.12	0.43	0.05	0.18	233.00	0.0192	0.04	0.45	**0.00**	6	3.83
SSD++	0.03	**0.75**	0.01	**0.21**	147.95	−0.0005	0.02	0.25	3.23	62	6.10

SG stands for subgroups. The number of subgroups for APRIORI-SD was pre-set as it is one of the model parameters. The best result for each measure is highlighted in bold.

Table 3
Stepwise AIC backward regression model with subgroups (a) and ANOVA of LinR models with and without subgroups (b).

Method	Group	(a) selected by stepwise AIC		(b) ANOVA (versus no groups)	p-value
		Only groups	Groups and clinical variables	Pr(>F)	
APRIORI-SD	all			0.0754	
	1	yes	yes	0.5638	
	2	yes	yes	0.3119	
	3	yes	yes	0.3360	
	4	yes	no	N/A	
	5	yes	no	N/A	
PRIM	all			0.0158	<0.05
	1	yes	no	N/A	
	2	yes	no	N/A	
	3	yes	yes	0.3383	
	4	yes	no	N/A	
	5	yes	yes	0.0286	<0.05
	6	yes	yes	0.0673	
SSD++	all			1.17E-130	<0.001
	1	yes	yes	3.89E-04	<0.001
	2	yes	no	N/A	
	3	yes	yes	1.85E-03	<0.01
	4	yes	yes	1.27E-06	<0.001
	5	yes	yes	3.94E-06	<0.001
	6	yes	yes	7.69E-06	<0.001
	7	yes	yes	4.90E-07	<0.001
	8	yes	yes	3.94E-04	<0.001
	9	yes	yes	4.91E-02	<0.05
	10	yes	yes	3.62E-04	<0.001
	11	yes	yes	1.91E-06	<0.001
	12	yes	yes	2.88E-06	<0.001
	13	yes	yes	2.37E-01	
	14	yes	yes	1.48E-05	<0.001
	15	yes	no	N/A	
	16	yes	yes	1.20E-06	<0.001
	17	yes	yes	2.97E-05	<0.001
	18	no	no	N/A	
	19	yes	yes	9.80E-05	<0.001
	20	yes	yes	4.19E-06	<0.001
	21	yes	yes	1.63E-04	<0.001
	22	yes	yes	7.48E-07	<0.001
	23	yes	yes	2.37E-07	<0.001
	24	yes	yes	1.75E-04	<0.001
	25	yes	yes	9.07E-07	<0.001
	26	yes	yes	3.81E-05	<0.001
	27	yes	yes	1.84E-07	<0.001
	28	yes	yes	9.61E-05	<0.001
	29	yes	yes	2.49E-03	<0.001
	30	yes	yes	3.61E-06	<0.001
	31	yes	yes	4.32E-05	<0.001
	32	yes	yes	1.86E-07	<0.001
	33	yes	yes	5.63E-05	<0.001
	34	yes	yes	1.72E-03	<0.01
	35	yes	yes	4.63E-04	<0.001
	36	yes	yes	1.68E-02	<0.05
	37	yes	yes	2.87E-04	<0.001
	38	yes	yes	9.08E-09	<0.001
	39	yes	yes	6.10E-03	<0.01
	40	yes	yes	1.18E-02	<0.05
	41	yes	yes	1.92E-07	<0.001
	42	yes	yes	5.80E-06	<0.001
	43	yes	yes	7.33E-06	<0.001
	44	yes	yes	6.12E-06	<0.001
	45	yes	yes	4.99E-05	<0.001
	46	yes	yes	8.92E-04	<0.001
	47	yes	yes	1.29E-05	<0.001
	48	yes	yes	6.34E-05	<0.001
	49	yes	yes	7.37E-07	<0.001
	50	yes	yes	1.27E-08	<0.001
	51	yes	yes	3.06E-04	<0.001
	52	yes	no	N/A	
	53	yes	yes	6.12E-06	<0.001
	54	yes	yes	3.51E-07	<0.001
	55	yes	yes	2.67E-06	<0.001
	56	yes	yes	6.23E-06	<0.001
	57	yes	yes	3.20E-02	<0.05
	58	yes	yes	3.31E-05	<0.001
	59	yes	no	N/A	
	60	yes	yes	3.65E-04	<0.001
	61	yes	yes	1.95E-03	<0.01
	62	yes	yes	7.47E-05	<0.001

The p-value is omitted when it corresponds to a non-statistically-significant result.

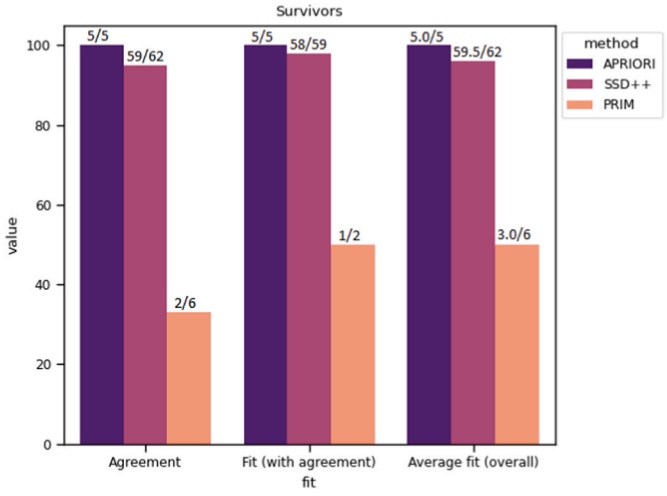

Fig. 1. Expert evaluation on clinical relevance of the obtained subgroups.

Findings – For the quantitative analyses, we observed that the tested algorithms yield different results in terms of (i) the total number of discovered subgroups (ranging between 5 and 62), (ii) the number of selected variables (overall and per subgroup), and (iii) the predictive value of the subgroups. Concerning the qualitative assessment (by means of evaluation of the clinical relevance of the subgroups by intensivists), we make three overall observations. Firstly, the subgroups make clinical sense. However, secondly, including the (past) infection waves does not make sense for the purpose of (future) triage. Lastly, although many (62) groups were discovered with the SSD++ algorithm, there is much potential use for these subgroups – either in an automated way, for example for flagging, or as clinical decision rules. As for the clinical utility of the subgroups, APRIORI-SD and PRIM are considered less effective because the subgroups do not have added predictive value and the subgroups are deemed clinically less relevant. Especially APRIORI-SD subgroups were not good: the mortality in each group (i.e. the number of non-survivors divided by the number of patients in the group) was about 0.5 (many patients, especially the non-survivor belonged to multiple subgroups) whereas SGD is supposed to find distinctive groups (either with high or low outcome probabilities), which means the algorithm proved not effective. Finally, SDD++ resulted best from both clinical and numerical (predictive power and redundancy), although it was second best in terms of information theoretic measures after APRIORI-SD.

Strengths – The conducted analysis was very extensive by evaluating many quantitative measures (in general on algorithmic performance) as well as qualitative aspects of found subgroups (by means of expert

consultation with questionnaires). Furthermore, the predictive performance of the subgroups was assessed extensively (by evaluating the subgroups as patient features, log likelihood tests, and stepwise feature selection) and separately from the internal validation during model development. Such an extensive and systematic approach as we undertook facilitates the use of algorithms in clinical practice. The analysed use case of ICU triage of COVID-19 patients included real world data. In this study, this case was rather illustrational and an example in support of the study's aim how to analyse and use SGD algorithms. However, since we used real world data, a follow-up clinical study for ICU COVID-19 triage with found subgroups can be readily undertaken (although it may depend on the virus variant, vaccination status and vaccine).

Limitations – Three main factors limit generalizing the results of this study. Firstly, the use of a single country dataset is limiting, mainly as to which subgroups were found for the specific prediction task. Secondly, the subgroups were evaluated by (only) two intensivists (albeit from different institutions). Establishing broader common ground on the subgroups (possibly revised after external validation) may require a larger evaluation panel. Finally, we evaluated three SGD algorithms that we considered representative as explained above, but the sheer number of algorithms could warrant a more extensive analysis including more algorithms, and possibly related algorithms like association rules and clustering/phenotyping.

Implications – We showed that SGD methods can potentially be used in clinical practice. Our in-depth evaluation, which included clinical validation of the discovered subgroups, showed that SGD allows clinicians to identify clinically relevant subgroups for COVID-19 patients. SGD methods can be implemented in clinical decision support systems and our methodology can be used to validate SGD methods, also in another setting and for other outcomes. Subgroups can be interpreted as rules, which can be implemented in a clinical decision support system to identify high-risk patients. For instance, a newly admitted patient can be mapped to a subgroup by which the derived prognosis can be taken into account in treatment decision and can also be discussed with the patient and the family.

There are several ways to use the found subgroups in clinical practice. Such use may range from an automated algorithm for flagging patients who may have low survival probabilities, to use of the rules describing the subgroups in triage protocols. For direct clinical application of the found subgroups, one may need to consider that the threshold levels as used in subgroups are often extreme values (e.g., A-a gradient >450) and these may not occur often enough to justify inclusion in clinical practice. Concerning use for triage, the involved intensivists mentioned that thinking in subgroups or rules is the other way around from their usual way of thinking. For example, the intensivists think which patients do have a mortality of 80–100%, to which the answer is 80+ year old COVID-19 patients with >2 comorbidities, while subgroups are defined also on low risk of mortality. Noteworthy, some subgroups do not seem to represent ICU patients that are considered typical given by the variables that were used in the rules. However, typical patients vary during a pandemic. ICU patients in the first wave might have been a medium care or general ward patients in subsequent waves. Furthermore, age, creatinine and renal replacement therapy are known predictors of high mortality, but combining these variables with other variables to assess mortality remains difficult. Subgroup analyses can generate patient groups that are not considered as an important subgroup in clinical practice but yet help in rethinking the influence of variable on the outcome and generate new hypotheses.

Our study has implications for researchers and practitioners. We demonstrated how to assess the numerical and clinical quality of SGD methods to help clinicians to perform complex subgroup analysis as which SGD method is best applicable. SGD can pave the way to personalized medicine as our approach can ease the implementation of SGD in clinical decision support systems. The fact that APRIORI-SD was best in our case study according to information theoretic measures but was not for the other evaluations shows that our deeper evaluation results in a better choice of the best SGD method.

6. Conclusion

Automated patient subgroup discovery methods find clinical subgroups that are relevant both when assessed quantitatively (yield added predictive value) and qualitatively (intensivists consider the subgroups significant). Different methods yield different subgroups with varying degrees of predictive performance and clinical quality.

As future work, we propose to conduct further external validation studies to address the limitation that only one dataset was used. To establish broader common ground on the clinical relevance and validity of the subgroups by a larger evaluation panel, the qualitative analysis should be assessed in a broader Delphi study. Finally, several specific findings about the subgroups (e.g., non-typical ICU patients and particular variable interactions) need further follow-up. Future research is needed to explore which algorithm gives most benefit in other settings.

Funding

This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) COVID-19 Programme in the bottom-up focus area 1 "Predictive diagnostics and treatment" for theme 3 "Risk analysis and prognostics" (project number 10430 01 201 0011: IRIS). The funder had no role in the design of the study or writing the manuscript.

Author contribution

IV Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft; **MYK** Software, Formal analysis, Writing - Review & Editing; **DD** Methodology, Writing - Review & Editing; **DdL** Methodology, Writing - Review & Editing; **NdK** Conceptualization, Methodology, Investigation, Writing - original draft, Supervision, Project administration; **MCS** Conceptualization, Writing - original draft, Methodology, Investigation, Supervision, Project administration. **IV** and **MCS** had full access to the data and have verified the data.

Other declarations

The investigators were independent from the funders; IV, and MCS had full access to the data, have verified the data, and take responsibility for the integrity of the data and the accuracy of the data analysis; the lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained. DD, NdK, DdL are board members of the NICE foundation that facilitate the data collection for this study.

Ethics approval and consent to participate

The study protocol was reviewed by the Medical Ethics Committee of the Amsterdam Medical Center, the Netherlands. This committee provided a waiver from formal approval (W20_273 # 20.308) and informed consent since this trial does not fall within the scope of the Dutch Medical Research (Human Subjects) Act.

Data and code availability

Data is available under stringent conditions as described on the NICE website https://www.stichting-nice.nl/extractieverzoek.jsp (in Dutch).

The code used for our analyses is publicly available at https://bitbucket.org/aumc-kik/subgroup-discovery/.

Declaration of competing interest

The authors declare that they have no conflict of interests.

Acknowledgements

We thank Sylvia Brinkman for her support with the data extraction and all participating ICUs for making this study possible.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107146.

References

[1] Available fromJ.P.T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V.A. Welch (Eds.), Cochrane Handbook for Systematic Reviews of Interventions Version 6.3 (Updated February 2022). Cochrane, 2022. : www.training.cochrane.org/handbook.

[2] S. Ventura, J.M. Luna, Subgroup discovery, in: Supervised Descriptive Pattern Mining, Springer, Cham, 2018, https://doi.org/10.1007/978-3-319-98140-6_4.

[3] S. Helal, Subgroup discovery algorithms: a survey and empirical evaluation, J. Comput. Sci. Technol. 31 (2016) 561–576, https://doi.org/10.1007/s11390-016-1647-1.

[4] H.M. Proença, P. Grünwald, T. Bäck, v. Leeuwen M, Discovering outstanding subgroup lists for numeric targets using MDL, in: F. Hutter, K. Kersting, J. Lijffijt, I. Valera (Eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2020, Lecture Notes in Computer Science, 12457, Springer, Cham, 2021, https://doi.org/10.1007/978-3-030-67658-2_2.

[5] C. Esnault, M.-L. Gadonna, M. Queyrel, A. Templier, J.-D. Zucker, Q-finder: an algorithm for credible subgroup discovery in clinical data analysis — an application to the international diabetes management practice study, Front. Artif. Intell. 3 (2020), 559927, https://doi.org/10.3389/frai.2020.559927.

[6] Chirag Nagpal, Dennis Wei, Bhanukiran Vinzamuri, Monica Shekhar, Sara E. Berger, Subhro Das, Kush R. Varshney, Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines, in: Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL '20), 19–29, Association for Computing Machinery, New York, NY, USA, 2020, https://doi.org/10.1145/3368555.3384456.

[7] D. Gamberger, N. Lavrač, G. Krstačić, Active subgroup mining: a case study in coronary heart disease risk group detection, Artif. Intell. Med. 28 (1) (2003) 27–57, https://doi.org/10.1016/S0933-3657(03)00034-4.

[8] N. Lavrac, P.K. Novak, I. Mozetic, V. Podpecan, H. Motaln, M. Petek, K. Gruden, Semantic subgroup discovery: using ontologies in microarray data analysis, Annu Int Conf IEEE Eng Med Biol Soc 2009 (2009) 5613–5616, https://doi.org/10.1109/IEMBS.2009.5333782.

[9] Zainab Al-Taie, Danlu Liu, Jonathan B. Mitchem, Christos Papageorgiou, Jussuf T. Kaifi, Wesley C. Warren, Chi-Ren Shyu, Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential, J. Biomed. Inf. 118 (2021), 103792, https://doi.org/10.1016/j.jbi.2021.103792.

[10] B. Ozdemir, W. Abd-Almageed, S. Roessler, X.W. Wang, iSubgraph: integrative genomics for subgroup discovery in hepatocellular carcinoma using graph mining and mixture models, PLoS One 8 (11) (2013), e78624, https://doi.org/10.1371/journal.pone.0078624.

[11] D. Gamberger, N. Lavrač, A. Krstačić, et al., Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis, Appl. Intell. 27 (2007) 205–217, https://doi.org/10.1007/s10489-007-0068-9.

[12] A. Abu-Hanna, B. Nannings, D. Dongelmans, A. Hasman, PRIM versus CART in subgroup discovery: when patience is harmful, J. Biomed. Inf. 43 (5) (2010) 701–708, https://doi.org/10.1016/j.jbi.2010.05.009.

[13] B. Nannings, A. Abu-Hanna, E. de Jonge, Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients, Int. J. Med. Inf. 77 (4) (2008) 272–279, https://doi.org/10.1016/j.ijmedinf.2007.06.007.

[14] O. Kholod, W.I. Basket, J.B. Mitchem, J.T. Kaifi, R.D. Hammer, C.N. Papageorgiou, C.R. Shyu, Immune-related gene signatures to predict the effectiveness of chemoimmunotherapy in triple-negative breast cancer using exploratory subgroup discovery, Cancers 14 (23) (2022 Nov 25) 5806, https://doi.org/10.3390/cancers14235806.

[15] O. Kholod, W. Basket, D. Liu, J. Mitchem, J. Kaifi, L. Dooley, C.R. Shyu, Identification of immuno-targeted combination therapies using explanatory subgroup discovery for cancer patients with EGFR wild-type gene, Cancers 14 (19) (2022 Sep 29) 4759, https://doi.org/10.3390/cancers14194759.

[16] A. Ibald-Mulli, J. Seufert, J.M. Grimsmann, M. Laimer, P. Bramlage, A. Civet, M. Blanchon, S. Gosset, A. Templier, W.D. Paar, F.L. Zhou, S. Lanzinger, Identification of predictive factors of diabetic ketoacidosis in type 1 diabetes using a subgroup discovery algorithm, Diabetes Obes. Metabol. (2023 Mar 3), https://doi.org/10.1111/dom.15039.

[17] T.R. Yordanov, R.R. Lopes, A.C.J. Ravelli, M. Vis, S. Houterman, H. Marquering, A. Abu-Hanna, NHR THI Registration committee, An integrated approach to geographic validation helped scrutinize prediction model performance and its variability, J. Clin. Epidemiol. 157 (2023 Feb 22) 13–21, https://doi.org/10.1016/j.jclinepi.2023.02.021.

[18] D. Gamberger, N. Lavrač, F. Železný, J. Tolar, Induction of comprehensible models for gene expression datasets by subgroup discovery methodology, J. Biomed. Inf. 37 (4) (2004) 269–284, https://doi.org/10.1016/j.jbi.2004.07.007.

[19] C.H. Olson, S. Dey, V. Kumar, K.A. Monsen, B.L. Westra, Clustering of elderly patient subgroups to identify medication-related readmission risks, Int. J. Med. Inf. 85 (1) (2016) 43–52, https://doi.org/10.1016/j.ijmedinf.2015.10.004.

[20] H.C. Chen, W. Zou, T.P. Lu, J.J. Chen, A composite model for subgroup identification and prediction via bicluster analysis, PLoS One 9 (10) (2014), e111318, https://doi.org/10.1371/journal.pone.0111318.

[21] L. Bondeelle, S. Chevret, S. Cassonnet, S. Harel, B. Denis, et al., Profiles and outcomes in patients with COVID-19 admitted to wards of a French oncohematological hospital: a clustering approach, PLoS One 16 (5) (2021), e0250569, https://doi.org/10.1371/journal.pone.0250569.

[22] N. Bruse, E.J. Kooistra, A. Jansen, R.B.E. van Amstel, N.F. de Keizer, J.N. Kennedy, C. Seymour, L.A. van Vught, P. Pickkers, M. Kox, Clinical sepsis phenotypes in critically ill COVID-19 patients, Crit. Care 26 (1) (2022 Aug 9) 244, https://doi.org/10.1186/s13054-022-04118-6.

[23] J. Bommelé, M. Kleinjan, T.M. Schoenmakers, W.J. Burk, R. van den Eijnden, et al., Identifying subgroups among hardcore smokers: a latent profile approach, PLoS One 10 (7) (2015), e0133570, https://doi.org/10.1371/journal.pone.0133570.

[24] U. Niemann, M. Spiliopoulou, B. Preim, T. Ittermann, H. Völzke, Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data, in: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), 2017, pp. 582–587, https://doi.org/10.1109/CBMS.2017.15.

[25] W. Qi, A. Abu-Hanna, v. Esch TEM, D. de Beurs, Y. Liu, L.E. Flinterman, M.C. Schut, Explaining heterogeneity of individual treatment causal effects by subgroup discovery: an observational case study in antibiotics treatment of acute rhinosinusitis, Artif. Intell. Med. 116 (2021), https://doi.org/10.1016/j.artmed.2021.102080.

[26] Z. Rong, Z. Liu, J. Song, L. Cao, Y. Yu, M. Qiu, Y. Hou, MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data, Comput. Biol. Med. 150 (2022), 106085, https://doi.org/10.1016/j.compbiomed.2022.106085.

[27] F. Nezhadmoghadam, J. Tamez-Peña, Risk profiles for negative and positive COVID-19 hospitalized patients, Comput. Biol. Med. 136 (2021), 104753, https://doi.org/10.1016/j.compbiomed.2021.104753.

[28] N. van de Klundert, R. Holman, D.A. Dongelmans, N.F. de Keizer, Data resource profile: the Dutch national intensive care evaluation (NICE) registry of admissions to adult intensive care units, Int. J. Epidemiol. 44 (6) (2015 Dec), https://doi.org/10.1093/ije/dyv291, 1850-1850h.

[29] D.G. Arts, N.F. De Keizer, G.J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework, J. Am. Med. Inf. Assoc. 9 (6) (2002) 600–611, https://doi.org/10.1197/jamia.m1087.

[30] S. van Buuren, K. Groothuis-Oudshoorn, MICE: multivariate imputation by chained equations in R, J. Stat. Software 45 (2011), https://doi.org/10.18637/jss.v045.i03.

[31] J.H. Friedman, N.I. Fisher, Bump hunting in high-dimensional data, Stat. Comput. 9 (1999) 123–143, https://doi.org/10.1023/A:1008894516817.

[32] B. Kavšek, N. Lavrač, V. Jovanoski, APRIORI-SD: adapting association rule learning to subgroup discovery, in: R. Berthold, M, H.J. Lenz, E. Bradley, R. Kruse, C. Borgelt (Eds.), Advances in Intelligent Data Analysis V. IDA 2003. Lecture Notes in Computer Science, 2810, Springer, Berlin, Heidelberg, 2003, https://doi.org/10.1007/978-3-540-45231-7_22.

[33] F. Herrera, C.J. Carmona, P. González, et al., An overview on subgroup discovery: foundations and applications, Knowl. Inf. Syst. 29 (2011) 495–525, https://doi.org/10.1007/s10115-010-0356-2.

[34] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: E. Parzen, K. Tanabe, G. Kitagawa (Eds.), Selected Papers of Hirotugu Akaike, Springer New York, New York, NY, 1998, pp. 199–213, https://doi.org/10.1007/978-1-4612-1694-0_15.

[35] F. Lemmerich, M. Becker, Pysubgroup: easy-to-use subgroup discovery in Python, in: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science, 11053, Springer, Cham, 2019, https://doi.org/10.1007/978-3-030-10997-4_46.