1 Prediction of intensive care unit mortality based on missing events

2 Tatsuma Shoji (1)^{*1,*2}, Hiroshi Yonekura (2)^{*2}, Sato Yoshiharu (1), Yohei Kawasaki (3)

3 *1: Corresponding Author; *2: Contributed equally

4

5 (1) DNA Chip Research Inc.

6 1-15-1 Kaigan, Suzue Baydium 5F, Minato-ku, Tokyo 105-0022, Japan

7

8 (2) Department of Anesthesiology and Pain Medicine, Fujita Health University Bantane

9 Hospital

10 2-6-10 Otoubashi, Nakagawa-ku, Nagoya City, Aichi, 454-8509, Japan

11

12 (3) Faculty of Nursing, Japan Red Cross College of Nursing, Tokyo, Japan

13 4-1-3 Hiroo, Shibuya-ku, Tokyo 150-0012, Japan

14

15 **E-mail:** Shoji Tatsuma: t-shoji@dna-chip.co.jp

16 Yonekura Hiroshi: hiroshi.yonekura@fujita-hu.ac.jp

17 Sato Yoshiharu: yo-sato@dna-chip.co.jp

18 Kawashiki yohei: ykawasaki@chiba-u.jp

19 **Abstract**

20 **Background**

21 The increasing availability of electronic health records has made it possible to
 22 construct and implement models for predicting intensive care unit (ICU) mortality using
 23 machine learning. However, the algorithms used are not clearly described, and the
 24 performance of the model remains low owing to several missing values, which is
 25 unavoidable in big databases.

26 **Methods**

27 We developed an algorithm for subgrouping patients based on missing event
 28 patterns using the Philips eICU Research Institute (eRI) database as an example. The
 29 eRI database contains data associated with 200,859 ICU admissions from many
 30 hospitals (>400) and is freely available. We then constructed a model for each subgroup
 31 using random forest classifiers and integrated the models. Finally, we compared the
 32 performance of the integrated model with the Acute Physiology and Chronic Health
 33 Evaluation (APACHE) scoring system, one of the best known predictors of patient
 34 mortality, and the imputation approach-based model.

35 **Results**

Subgrouping and patient mortality prediction were separately performed on two groups: the sepsis group (the ICU admission diagnosis of which is sepsis) and the non-sepsis group (a complementary subset of the sepsis group). The subgrouping algorithm identified a unique, clinically interpretable missing event patterns and divided the sepsis and non-sepsis groups into five and seven subgroups, respectively. The integrated model, which comprises five models for the sepsis group or seven models for the non-sepsis group, greatly outperformed the APACHE IV or IVa, with an area under the receiver operating characteristic (AUROC) of 0.91 (95% confidence interval 0.89–0.92) compared with 0.79 (0.76–0.81) for the APACHE system in the sepsis group and an AUROC of 0.90 (0.89–0.91) compared with 0.86 (0.85–0.87) in the non-sepsis group. Moreover, our model outperformed the imputation approach-based model, which had an AUROC of 0.85 (0.83–0.87) and 0.87 (0.86–0.88) in the sepsis and non-sepsis groups, respectively.

Conclusions

We developed a method to predict patient mortality based on missing event patterns. Our method more accurately predicts patient mortality than others. Our results indicate that subgrouping, based on missing event patterns, instead of imputation is essential and effective for machine learning against patient heterogeneity.

54 **Trial registration**

55 Not applicable.

56 **Keywords**

57 ICU, patient mortality, Acute Physiology and Chronic Health Evaluation scoring,

58 machine learning, missing values

59

60 **Background**

61 Accurate prognostication is central to medicine [1] and at the heart of clinical

62 decision-making. Sepsis is a systemic response to infection, with the highest mortality

63 rate in the field of intensive care. For nearly a decade, sepsis-related mortality has

64 remained at 20%–30%, with unsatisfactory improvements [2]. Prognosis remains a

65 challenge for physicians because of the high heterogeneity of clinical phenotypes [3].

66 Because accurate diagnoses can improve the physician’s decision-making abilities, it is

67 one of the essence for medical practice of sepsis to improve prognosis accuracy [4].

68 Most large prognostic studies have developed clinical scoring systems for

69 objective risk stratification in the early phase of hospital admission using physiological

70 measurements, medical history, and demographics to predict the likelihood of survival

71 [5]. Among these scoring systems, Acute Physiology and Chronic Health Evaluation

(APACHE) scoring [6] is one of the best known, which has been validated for application at approximately 24 h after intensive care unit (ICU) admission. The APACHE scoring system generates a point score based on worst values of 12 variables during the initial 24 h after ICU admission. The APACHE II score was published in 1985 [5]; APACHE IV and IVa are the latest versions [7]. Built on the study of a more recent patient population and standard-of-care, APACHE IV or IVa are recommended as scoring systems over APACHE II and III. However, these indicators lack the precision required for use at the individual level. Therefore, efforts have been made to increase the performance of these indicators through the use of computational techniques, such as machine learning.

Machine learning classifiers may be advantageous for outcome prediction because they can handle large numbers of variables and learn non-linearities [8, 9]. Random forest is an example of modern machine learning algorithms [10]. Machine learning requires large volumes of data and generating a complex model. However, the increasing availability of electronic health records has made constructing and implementing the models possible. The availability of large training sets [11, 12] has made investigation of such approaches feasible. Studies on the application of machine learning to intensive care datasets have been performed [13, 14, 15, 16, 17, 18, 19, 20,

21]. Traditional scoring systems were outperformed by machine learning approaches in mortality prediction in medical ICUs. However, the algorithms are not clearly described and the produced models or databases are available commercially.

The Philips eICU Research Institute (eRI) is a non-profit institute established by Phillips that is governed by customers [22]. This freely accessible critical care dataset spans more than a decade and contains detailed information about individual patient care, including time-stamped, nurse-verified physiological measurements. As a distinctive feature, datasets are donated by >400, including teaching and non-teaching, hospitals. However, the eRI, which comprises 31 files according to clinical categories, includes much missing data. This makes the application of machine learning with eRI challenging.

In this study, we revealed unique characteristics of the distribution of missing values in the eRI database. By taking advantage of this characteristic, we successfully developed a more accurate and interpretable model against the APACHE system. Accurate prediction of patient state is critical in the critical care field. To address this problem, we propose the “missing-event-based prediction,” a new method for predicting ICU mortality.

107

108 **Methods**

109 **Dataset used in this study**

110 We used the eRI

111 (<https://www.usa.philips.com/healthcare/solutions/enterprise-telehealth/eri>)

112 database, which is an open database and provides anonymous data, including

113 demographic information, vital signs measurements, laboratory test results, drug

114 information, procedural information, fluid balance reports, hospital length of stay data,

115 and data on in-hospital mortality, donated by >400 member institutions. The eRI

116 contains data associated with 200859 ICU admissions with more than 100 variables. In

117 this study, we carefully selected 58 clinically important variables to construct a model,

118 which is easy to interpret clinically. Selected variables are summarized in Table S1 (see

119 Additional file 1). Briefly, the variables were as follows: (I) Listed from laboratory

120 measurements, including white blood cell count, hematocrit, bilirubin, creatinine,

121 sodium, albumin, blood urea nitrogen, glucose, arterial pH, fraction of inspired oxygen,

122 arterial oxygen pressure, and arterial blood carbon dioxide pressure. (II) Listed from

123 routine charted data, including temperature, respiratory rate, heart rate, mean arterial

124 blood pressure, urine output, and Glasgow Coma Scale (including score for eye, motor,

125 and verbal responses). (III) Listed from information taken at the time of ICU admission,

126 including age, gender, height, weight, time from hospitalization to ICU admission, type
 127 of hospital, bed count of the hospital, hospital ID, and diagnoses names at ICU
 128 admission. (IV) Comorbidities, including myocardial infarction within 6
 129 months, diabetes, hepatic failure, dialysis, immunosuppressive disease, lymphoma,
 130 leukemia, metastatic cancer, cirrhosis, acquired immune deficiency syndrome,
 131 and history of intubation and mechanical ventilation. Intervention required at admission,
 132 including catheter intervention for myocardial infarction, coronary artery bypass
 133 grafting with or without internal thoracic artery grafts, and use of thrombolytics. (V)
 134 APACHE scoring system, including not only the APACHE score but also actual ICU
 135 and in-hospital mortality, predicted ICU and in-hospital mortality, length of ICU stay,
 136 length of hospital stay, and ventilation duration. Among these variables, diagnoses
 137 names at ICU admission were used to define the sepsis or non-sepsis group (see the
 138 Definition of sepsis and non-sepsis groups in Methods section for details), actual ICU
 139 and in-hospital mortality were used as response variables to construct the models,
 140 predicted ICU and in-hospital mortality using the APACHE IV or IVa were used as
 141 benchmarks against our model, and others (53 parameters in total) were used as
 142 explanatory variables for machine learning.
 143

144 **Definition of sepsis and non-sepsis groups**

145 Inclusion criteria for the sepsis group were as follows: (I) extraction by
146 diagnoses names at ICU admission, namely “Sepsis, cutaneous/soft tissue,” “Sepsis, GI,”
147 “Sepsis, gynecologic,” “Sepsis, other,” “Sepsis, pulmonary,” “Sepsis, renal/UTI
148 (including bladder),” and “Sepsis, unknown;” (II) selection by documentation of
149 prognosis; and (III) exclusion by cases with any missing data in Acute Physiology Score
150 (APS)-related variables and prognosis information. Inclusion criteria for the non-sepsis
151 group were almost the same as the one used for the sepsis group. Briefly, the
152 complementary subset of (I) was first selected. Then, (II) and (III) were applied to the
153 resulting subset. Thus, 4226 and 23170 cases were defined as sepsis and non-sepsis
154 groups, respectively. To reproduce our results and access patient lists, follow the jupyter
155 notebook at
156 https://github.com/tatsumashoji/ICU/1_the_sepsis_group_and_non_sepsis_group.ipynb.

157

158 **Subgrouping based on missing data**

159 Subgroups were defined according to the diagram shown in Fig. S1 (see
160 Additional file 2). Briefly, (I) patient lists, containing no missing data for any pattern of
161 52 variables (derived from 53 variables) were first generated. Then, (II) the patient list,

162 which had a size not too small and not too large among 53 lists, was defined as
 163 subgroup #1. For the other subgroups, we repeated (I) and (II) with the other patients.
 164 To reproduce this subgrouping, follow the jupyter notebook opened at
 165 https://github.com/tatsumashoji/ICU/2_subgrouping_sepsis.ipynb for the sepsis group
 166 and https://github.com/tatsumashoji/ICU/3_subgrouping_non_sepsis.ipynb for the
 167 non-sepsis group.

168

169 **Generation and performance of our model**

170 To construct the model for each group, we used the random forest classifier
 171 implemented with “scikit-learn (0.24.1)” [23]. Briefly, we first selected 80% data as a
 172 training dataset for each group so that the ratio of “ALIVE” and “EXPIRED” cases
 173 were the same between the two datasets. After hyperparameters for the random forest
 174 were determined using the grid search algorithm, the actual model was generated, and
 175 the mean and standard deviation of accuracy were checked through 5-fold
 176 cross-validation (see Table S2-S5 in Additional file 1). Finally, patient mortalities in the
 177 test dataset were predicted by the generated model and compared to those from
 178 APACHE IV or IVa by drawing receiver operating characteristic (ROC) curves and
 179 calculating the area under the ROC (AUROC). The confidence intervals for the

180 AUROC were calculated as described [24]. For calibration plots, we used the module
181 “sklearn.calibration.calibration_curve.” All results in Fig. 2 can be reproduced by
182 running “https://github.com/tatsumashoji/ICU/4_sepsis_prediction.ipynb.”

183

184 **Imputation of missing values**

185 For imputation of missing values, we used multivariate imputation algorithms
186 implemented with “sklearn.impute.IterativeImputer,” which uses the entire set of
187 available feature dimensions to estimate missing values. We used the 0.24.1 version of
188 scikit-learn. To reproduce results shown in Fig. 3, follow the jupyter notebook opened
189 at https://github.com/tatsumashoji/ICU/5_imputation.ipynb.

190

191 **Results**

192 **Heterogeneity of the sepsis group**

193 Machine learning fails to predict the outcomes if input data consist of more
194 than two populations. Subgrouping of input data is essential before constructing the
195 model using machine learning. Therefore, we first investigated the distribution of each
196 parameter in the sepsis group. The histogram in Fig. 1a shows the distribution of some
197 variables recorded in “apacheApsVar.csv”, which contains the variables used to

198 calculate the Acute Physiology Score (APS) III for patients. More than two peaks were
 199 observed, indicating the necessity for subgrouping the sepsis group. Importantly, cases
 200 with missing data on any variable are excluded from the histograms in Fig. 1a. Thus, to
 201 examine how missing values were distributed in the sepsis group, correlation
 202 coefficients were calculated for all possible combinations of two variables after missing
 203 data were replaced with 0 and others with 1 (Fig. 1b). Surprisingly, perfect correlations
 204 were observed in some pairs in addition to the diagonal line. This suggests that missing
 205 events occur depending on other missing events. Clinically, this can be explained by
 206 different histories and backgrounds in the sepsis group. Thus, missing events contain
 207 information and may play an important role in subgrouping the sepsis group.

208

209 **Machine learning combining missing-event-based subgrouping approach**

210 **outperforms APACHE**

211 The unique distribution of missing values in the sepsis group led us to divide
 212 the group before machine learning. To account for the pattern of missing events when
 213 dividing the sepsis group, we defined subgroups such that each subgroup had the same
 214 missing pattern while the number of subgroups could as small as possible and the size
 215 of each subgroup could be as large as possible (Fig. 2a). For details, see the Methods

216 section. After defining five subgroups, we constructed models for each subgroup based
 217 on the random forest algorithm and calculated patient mortality. Then, we assessed the
 218 performance of each model by calculating AUROC and compared them to those from
 219 the APACHE IV or IVa system (Fig. 2b). Our model outperformed the APACHE
 220 systems, especially when integrating subgroups. Moreover, our model was more
 221 successful than APACHE in distinguishing patient mortality (Fig. 2c). Furthermore,
 222 results from calibration plots supported the ability of our model to predict patient
 223 mortality (Fig. 2d). Our model tends to output patient mortality higher than actual,
 224 whereas the APACHE system does not, indicating that the APACHE system
 225 underestimates patients' mortality.

226

227 **Comparison with the imputation approach**

228 A typical way of handling missing data is to impute them. Therefore, we
 229 constructed the model completely same as the way taken in Fig. 2 after imputing the
 230 dataset first, then compared to our model. Although the performance when imputed was
 231 slightly higher than the APACHE systems, it still remained lower than our method (Fig.
 232 3a), indicating that missing events are important information and our method, which
 233 accounts for the pattern of missing events, is reasonable. The imputation approach

234 outperforms the approach of the APACHE system but does perform as well as our
 235 model, as confirmed using the scatter plot (Fig. 3b). Our model distinguished patient
 236 mortality most precisely among all four models. Moreover, analyses of calibration plots
 237 supported the observation that our model is the most conservative, and thus, safer,
 238 because the model based on the imputation approach estimated lower patient mortality
 239 compared with our model (Fig. 3c).

240

241 **Application to the non-sepsis group**

242 To test the generalizability of our model, we applied our method to the
 243 non-sepsis group defined in the Methods section. We first generated seven subgroups
 244 using the same algorithm used for the sepsis group (Fig. 4a) and constructed models for
 245 each subgroup. Then, we compared the performance of the four models, namely our
 246 model, APACHE IV, APACHE IVa, and the model based on the imputation approach
 247 (Fig. 4b). Surprisingly, the distribution of missing values in the non-sepsis group was
 248 almost the same as that in the sepsis group, indicating that our subgrouping algorithm
 249 could be used for any group in addition to the sepsis group, and the performance of our
 250 model was the highest among all four models. Scatter plots and calibration plots also
 251 supported our model (Fig. S4, S5, see Additional file 2). These results strongly suggest

252 that missing events themselves are essential when predicting patient mortality in the
253 ICU.

254

255 **Discussion**

256 The main objective of this work was to present a more accurate and clinically
257 interpretable model for predicting patient mortality in the ICU and show the
258 effectiveness and potential of the missing-event-based prediction method. Our analysis
259 of the eRI confirmed that the larger the database across hospitals, the more the
260 heterogeneity (Fig. 1a). This is known as the domain shift problem, where traditional
261 models, including the APACHE system, developed in one geographical region or
262 healthcare system lose their ability to discriminate when applied outside of learning data.
263 The effect of prediction accuracy by volume and diversity of missing events clearly
264 increases in big data. We demonstrated the advantage of using a missing-event-based
265 method through comprehensive analysis and presented a more accurate and
266 interpretable model than others.

267 In the eRI database, missing events occurred depending on other missing
268 events (Fig. 1b). Thus, missing events were not random. Such cases would not be rare in
269 big data because the larger the dataset, the more complex the missing events. If missing

270 event was not regarded “missing at random,” the simple imputation of missing values
 271 would not be appropriate [25]. For example, Meiring et al. developed an algorithm to
 272 predict mortality over time in the ICU using CCHIC (a UK database) [20]. They
 273 imputed missing values using predictive mean matching through parallel
 274 implementation of multiple imputation by chained equations [26]. In this report, the
 275 discriminative power of the APACHE II score to predict outcomes on subsequent days
 276 reduced considerably. Moreover, Meiring et al. indicated that the longer the ICU stay,
 277 the smaller the size of the dataset, and therefore, an increase in the proportion of
 278 missing values. Thus, the weight of missing values impaired the accuracy of prediction.
 279 Handling missing values is important to generate a powerful model workable on
 280 complex clinical courses. Thus, our missing-event-based method is reasonable. Because
 281 the subgrouping pattern in this study was humanly impossible to detect, although easy
 282 to understand, our subgrouping algorithm could be the new *de facto* standard of
 283 database prescreening when constructing an accurate and interpretable model through
 284 machine learning using big data, including missing values. Our subgrouping algorithm
 285 can recognize missing events as an “informative missingness” instead of a dirty record,
 286 which big data cannot avoid including.

287 For constructing the model based on the random forest algorithm, we selected
288 53 variables as explanatory variables so that they were closely related to the APACHE
289 system. This selection is an important aspect of the interpretability of our model
290 because the APACHE system consists of clinically used variables. This feature is quite
291 important because the interpretable model addresses the problem of the “black box”,
292 which has hindered the use of this model as a clinical tool [27, 28, 29]. Given that our
293 model greatly outperformed the APACHE system, it can be considered a developed
294 version of the APACHE system, which detects the presence of more complex
295 interactions between covariates, leading to optimization both of clinical usability and
296 discrimination ability.

297 As tested with the non-sepsis group, our method is applicable to other cases,
298 indicating that it may not depend on diagnoses in the ICU. Under intensive care
299 conditions, avoiding the occurrence of missing events is a challenge; therefore, our
300 method can be useful in the clinic. Distinguish a generalizable method from a
301 generalizable predictive model for clinical applications of machine learning is important.
302 This study shows availability of subgrouping based on missing events as a generalizable
303 method and production of a promising predictor model as a clinical decision support
304 tool.

305

306 **Conclusions**

307 We developed a method to predict patient mortality based on information on
308 missing events. This method more accurately predicted patient mortality than others,
309 while maintaining clinical interpretability. Our results indicate that the subgrouping
310 process is important and effective for machine learning against patient heterogeneity.
311 By combining our method with other methods, such as the reinforcement learning, a
312 more realistic Artificial Intelligence clinician can be developed.

313

314 **List of abbreviations**

315 APACHE: Acute Physiology and Chronic Health Evaluation

316 ICU: Intensive Care Unit

317 eRI: eICU Research Institute

318 APS: Acute Physiology Score

319 ROC: Receiver Operating Characteristic

320 AUROC: Area Under the ROC

321

322 **Declarations**

323 **Ethics approval and consent to participate**

324 Not applicable.

325

326 **Consent for publication**

327 Not applicable.

328

329 **Availability of data and materials**

330 The datasets generated and/or analyzed during the current study are available in the

331 eICU repository, [<https://eicu-crd.mit.edu/gettingstarted/access/>].

332

333 **Competing interests**

334 The authors declare that they have no competing interests.

335

336 **Funding**

337 This work was supported by JSPS KAKENHI Grant Number JP 20K17834.

338

339 **Author contributions**

340 ST analyzed the eRI data, suggested, and implemented subgrouping algorithm and
 341 random forest model, prepared all of Figures and Tables and was a major contributor in
 342 writing the manuscript. YH selected the 53 explanatory variables in the point of clinical
 343 view, was the other major contributor in writing the manuscript as a clinician and
 344 supported the research grant. SY checked the programs that ST wrote, reviewed the
 345 manuscript and supported the research grant. KY contributed to the conception and
 346 overall design of the work, reviewed the results and manuscripts as a clinical statistician
 347 and supported the research grant. Finally, all four authors played an essential role in
 348 organizing this project.

349

350 **Acknowledgments**

351 We thank Ryota Jin for discussions and comments to the manuscript.

352

353 **References**

- 354 1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and
 355 prognostic research: what, why, and how? BMJ. 2009;338:b375.

- 356 2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al.
357 Global, regional, and national sepsis incidence and mortality, 1990–2017:
358 analysis for the Global Burden of Disease Study. *Lancet*. 2020;395:200-11.
- 359 3. Seymour CW, Kennedy JN, Wang S, Chang CC, Elliott CF, Xu Z, et al.
360 Derivation, validation, and potential treatment implications of novel clinical
361 phenotypes for sepsis, *J Am Med Assoc*. 2019;321:2003.
- 362 4. Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB.
363 Clinical applications of artificial intelligence in sepsis: a narrative review.
364 *Comput Biol Med*. 2019;115:103488.
- 365 5. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. *Crit*
366 *Care*. 2010;14:207.
- 367 6. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of
368 disease classification system. *Crit Care Med*. 1985;13:818-29.
- 369 7. Zimmerman JE1, Kramer AA. Outcome prediction in critical care: the Acute
370 Physiology and Chronic Health Evaluation models. *Curr Opin Crit Care*.
371 2008;14:491-7.
- 372 8. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920-30.

373 9. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine
374 learning applications in cancer prognosis and prediction. *Comput Struct*
375 *Biotechnol J.* 2015;13:8-17.

376 10. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.

377 11. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al.
378 MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016;3:160035.

379 12. Harris S, Shi S, Brealey D, MacCallum NS, Denaxas S, Perez-Suarez D, et al.
380 Critical Care Health Informatics Collaborative (CCHIC): data, tools and
381 methods for reproducible research: a multi-centre UK intensive care database.
382 *Int J Med Inform.* 2018;112:82-9.

383 13. Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill
384 patients using artificial neural network synthesised by genetic algorithm. *Lancet.*
385 1996;347:1146-50.

386 14. Jaimes F1, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic
387 regression and neural networks to predict death in patients with suspected sepsis
388 in the emergency room. *Crit Care.* 2005;9:R150-6.

- 389 15. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamdlou H, et al. Using
390 electronic health record collected clinical variables to predict medical intensive
391 care unit mortality. *Ann Med Surg (Lond)*. 2016;11:52-7.
- 392 16. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et
393 al. Prediction of in-hospital mortality in emergency department patients with
394 sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med*.
395 2016;23:269-78.
- 396 17. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamdlou H, et al. Using
397 electronic health record collected clinical variables to predict medical intensive
398 care unit mortality. *Ann Med Surg (Lond)*. 2016;11:52-7.
- 399 18. Ward L, Paul M, Andreassen S. Automatic learning of mortality in a CPN model
400 of the systemic inflammatory response syndrome. *Math Biosci*. 2017;284:12-20.
- 401 19. Aushev A, Ripoll VR, Vellido A, Aletti F, Pinto BB, Herpain A, et al. Feature
402 selection for the accurate prediction of septic and cardiogenic shock ICU
403 mortality in the acute phase. *PLoS One*. 2018;13:e0199089.
- 404 20. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al.
405 Optimal intensive care outcome prediction over time using machine learning.
406 *PLoS One*. 2018;13:e0206862.

- 407 21. García-Gallo JE, Fonseca-Ruiz NJ, Celi LA, Duitama-Muñoz JF. A machine
408 learning-based model for 1-year mortality prediction in patients admitted to an
409 Intensive Care Unit with a diagnosis of sepsis. *Med Intensiva*. 2020;44:160-70.
- 410 22. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU
411 Collaborative Research Database, a freely available multi-center database for
412 critical care research. *Sci Data*. 2018;5:1-13.
- 413 23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.
414 Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-30.
- 415 24. Cortes C, Mohri M. Confidence intervals for the area under the ROC curve. *Adv*
416 *Neural Inform Process Syst*. 2015;17:305-12.
- 417 25. Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations
418 for the primary analysis of continuous endpoints in longitudinal clinical trials.
419 *Drug Inf J*. 2008;42:303-19.
- 420 26. Buuren SV, Groothuis-Oudshoorn K. mice: multivariate imputation by chained
421 equations in R. *J Stat Softw*. 2011;45:1-67.

- 422 27. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL,
423 et al. A computational approach to early sepsis detection. *Comput Biol Med.*
424 2016;74:69-73.
- 425 28. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, et al.
426 High-performance detection and early prediction of septic shock for alcohol-use
427 disorder patients. *Ann Med Surg.* 2016;8:50-5.
- 428 29. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of
429 sepsis in the intensive care unit with minimal electronic health record data: a
430 machine learning approach. *JMIR Medical Informatics.* 2016;4:e5909.

431

432

433 **Figure Legends**

434 **Fig. 1 Distribution of Acute Physiology Score (APS)-related variables.**

435 **a** The distribution of APS-related variables in the sepsis group. **b** Joint matrix for
436 checking the missing values. Correlation coefficients in some combinations were picked
437 up and visualized using the heatmap where higher coefficients were lighter and lower
438 were darker.

439

440 **Fig. 2 The performance of our model.**

441 **a** Schematic of subgrouping results. The sepsis group was divided into five subgroups,
 442 namely subgroup #1 (884 cases), #2 (1037 cases), #3 (509 cases), #4 (1230 cases), and
 443 #5 (566 cases). Colors indicate information about missing values. For example, cases in
 444 subgroup #1 have no missing values for all variables except “hospitaladmitoffset”. **b**
 445 Receiver operating characteristic (ROC) curves for each subgroup produced using our
 446 model (red), Acute Physiology and Chronic Health Evaluation (APACHE) IV (green),
 447 and APACHE IVa (blue). The integrated version is shown on extreme right. For the
 448 confidence interval of AUROC, see Table S6 in Additional file 1. **c** Scatter plots of
 449 predicted mortality; APACHE IV vs. APACHE IVa (left), APACHE IV vs. our model
 450 (middle), and APACHE IVa vs. our model (right). Colors indicate actual mortality
 451 (green for “ALIVE” and red for “EXPIRED” cases). For scatter plots of each subgroup,
 452 see Fig. S2 in Additional file 2. **d** Calibration plots for our model (left), APACHE IV
 453 (middle), and APACHE IVa (right). For the plot for each subgroup, see Fig. S3 in
 454 Additional file 2.

455

456 **Fig. 3 Performance of our model against the model based on the imputation**

457 **approach.**

458 **a** Receiver operating characteristic (ROC) curves for the four models, namely our
459 model (red), Acute Physiology and Chronic Health Evaluation (APACHE) IV (green),
460 APACHE IVa (blue), and imputation (orange). For the confidence interval of AUROC,
461 see Table S7 in Additional file 1. **b** Scatter plots of predicted mortality; imputation vs.
462 APACHE IV (left), imputation vs. APACHE IVa (middle), and imputation vs. our
463 model (right). Colors indicate actual mortality (green for “ALIVE” and red for
464 “EXPIRED” cases). **c** Calibration plots for the model based on the imputation approach.

465

466 **Fig. 4 Performance in the non-sepsis group.**

467 **a** Schematic image for the result of subgrouping. The non-sepsis group was divided into
468 seven subgroups, namely #1 (3703 cases), #2 (4112 cases), #3 (1414 cases), #4 (930
469 cases), #5 (9487 cases), #6 (786 cases), and #7 (2738 cases). Colors indicate the
470 information about missing cases. For example, cases in the subgroup #1 have no
471 missing values for all variables except “hospitaladmitoffset”. **b** Receiver operating
472 characteristic (ROC) curve for each subgroup using our model (red), Acute Physiology
473 and Chronic Health Evaluation (APACHE) IV (green), and APACHE IVa (blue). The

474 integrated version is shown on extreme right with the result from the imputation
475 approach. For the confidence interval of the area under the ROC (AUROC), see Table
476 S8 in Additional file 1.

477

478 **Supplementary Information**

479 **Additional file 1**

480 File name: Additional_file_1.xlsx

481 File format: Microsoft excel.

482 Title of data: Tables S1–S8.

483 Description of data: Supplementary tables.

484

485 **Additional file 2**

486 File name: Additional_file_2.pptx

487 File format: Microsoft power point.

488 Title of data: Figures S1–S5.

489 Description of data: Supplementary figures.

490

491 **Supplementary Figure Legends**

Fig. S1 Flowchart for generating subgroups.

The blue box at the top indicates the starting point of the flowchart. Subgroups for the sepsis or non-sepsis group are generated by following steps of the flowchart. The set P is the defined patient list ($P = 4226$ for the sepsis group and $P = 23170$ for the non-sepsis group). V is the set of explanatory variables (i.e., the size of the V is 53). N_{min} and N_{max} are set at 500 and 2000, respectively, for the sepsis group and 500 and 10000, respectively, for the non-sepsis group. 2^V denotes the power set of set V . V_{dj}^c denotes the complementary subset for V_{dj} . $n(A)$ denotes the size of set A .

Fig. S2 Scatter plot for the predicted mortality of each subgroup.

Scatter plots for predicted mortality; Acute Physiology and Chronic Health Evaluation (APACHE) IV vs. APACHE IVa (left column), APACHE IV vs. our model (middle column), and APACHE IVa vs. our model (right column). Colors indicate actual mortality (green for “ALIVE” and red for “EXPIRED” cases).

Fig. S3 Calibration plots for each subgroup.

Calibration plots for our model (left column), Acute Physiology and Chronic Health Evaluation (APACHE) IV (middle column), and APACHE IVa (right column).

510

511 **Fig. S4 Scatter plots of predicted mortality in the non-sepsis group.**

512 Scatter plots of predicted mortality for each subgroup and integrated version; Acute
513 Physiology and Chronic Health Evaluation (APACHE) IV vs. APACHE IVa (left
514 column), APACHE IV vs. our model (middle column), and APACHE IVa vs. our
515 model (right column). Three panels at the bottom show the comparison with the model
516 generated based on the imputation approach. Colors indicate actual mortality (green for
517 “ALIVE” and red for “EXPIRED” cases).

518

519 **Fig. S5 Calibration plots in the non-sepsis group.**

520 Calibration plots in the non-sepsis group for each subgroup and the integrated version.
521 Columns on the left, middle, and right show results from our model, Acute Physiology
522 and Chronic Health Evaluation (APACHE) IV, and APACHE IVa, respectively. The
523 bottom panel shows results using the model generated based on the imputation
524 approach.

525

526 **Supplementary Table Legends**

527 **Table S1 Working dataset used in this study.**

528 The column on the left corresponds to file names, which can be downloaded from the
 529 eRI. We used 4 out of 31 files. The column in the middle shows the names of variables
 530 selected in this study. The column on the right indicates the role of variables in this
 531 study, where “key” indicates the unique key used for merging each .csv file and “x” and
 532 “y” indicate explanatory and response variables for machine learning, respectively,
 533 Acute Physiology and Chronic Health Evaluation (APACHE) indicates predicting
 534 patient mortality using APACHE IV or IVa, which was used as a benchmark against our
 535 model, and “classification” indicates the variable that is used to define the sepsis or
 536 non-sepsis group.

537

538 **Table S2 Results from cross-validation for each subgroup from the sepsis group.**

539

540 **Table S3 Results from cross-validation for imputed data from the sepsis group.**

541

542 **Table S4 Results from cross-validation for each subgroup from the non-sepsis**

543 **group.**

544

545 **Table S5 Results from cross-validation for imputed data from the non-sepsis**

546 **group.**

547

548 **Table S6 Confidence interval of area under the receiver operating characteristic**

549 **(AUROC) for each subgroup and the integrated version.**

550

551 **Table S7 Confidence interval of area under the receiver operating characteristic**

552 **(AUROC) for all four models.**

553

554 **Table S8 Confidence interval of area under the receiver operating characteristic**

555 **(AUROC) for each subgroup and the integrated version in the non-sepsis group.**

556

Figure 1

(A)

(B)

Figure 2

(A)

(B)

(C)

(D)

Figure 3

Figure 4

(A)

The non-sepsis group (23,170 cases)

(B)