



MIMIC-IIIデータベースを用いた 肺炎患者死亡と敗血症発症リスク の予測研究

東北大学情報科学研究所
応用情報科学専攻 木下・大林・西研究室
修士学生 顔子昂

2019 3.8



目次

- 1 イントロダクション
- 2 死亡と敗血症発症リスク予測
- 3 特徴分析
- 4 結論と感想

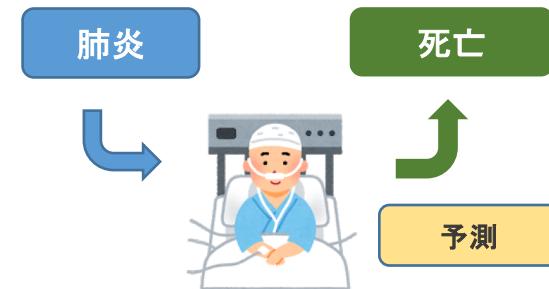
研究目的

➤ 対象:

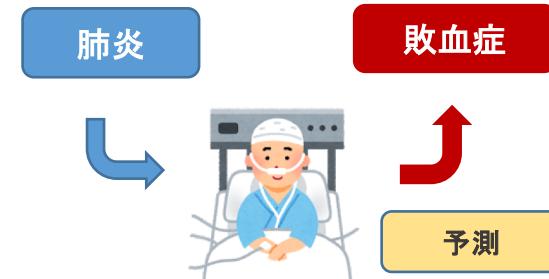
- 初回の入院時に肺炎と診断されたICU患者。

➤ 目標:

1. 肺炎と診断されて入院1日以内のデータを使って、40日以内に死亡するかどうかを予測する。



2. 肺炎と診断されて、入院1日以内のデータを使って、敗血症が発症するかどうかを予測する。



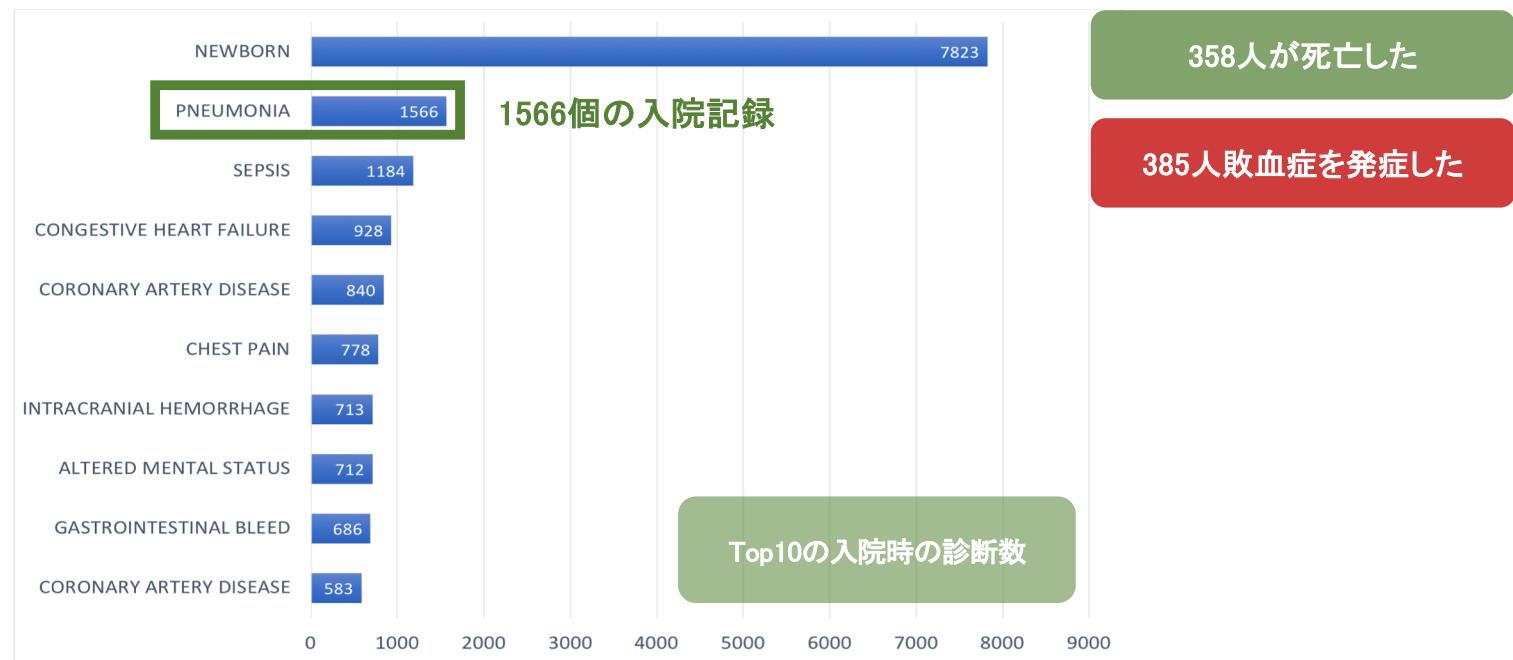
対象患者の抽出

テーブル	内容
Patients	患者背景、死亡時刻など
Admissions	入院記録など



複数回肺炎で
入院した場合、
1回目のみを対象

肺炎患者1419名



モデル訓練と予測プロセス

入力特徴

テーブル	内容
Labevents	病理学分析用の測定データ
Output events	EMR測定データ
Microbiology events	微生物検査データ
Diagnoses_icd	在院期間の診断データ

ラベルデータ

肺炎患者1419人

生存: 1061人
死亡: 358人発症しなかった: 1034人
発症した: 385人

入院後1日以内
の測定データを
訓練する

予測モデル

Logistic regression
SVM
Random forest

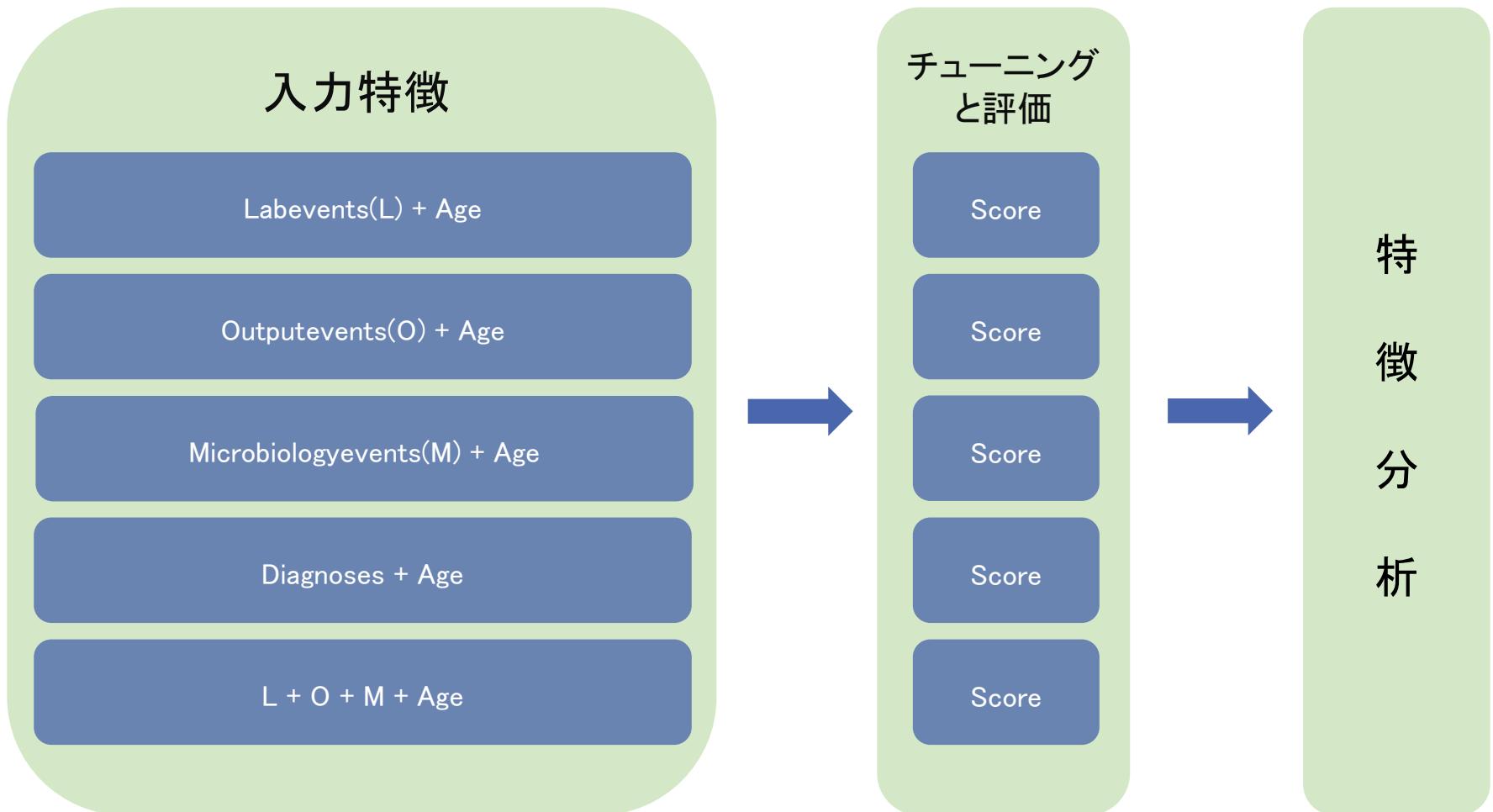
予測

40日以内に死亡するか

敗血症発症リスクがあるか

新しい
肺炎患者

モデリング全体像





目次

- 1 イントロダクション
- 2 死亡と敗血症発症リスク予測
- 3 特徴分析
- 4 結論と感想

Labevents特徴抽出

- アイデア: 入院後1日以内の病理学分析用データは予測に有効か。
- 中間発表でLabeventsテーブル中の19個の特徴、及び死亡と敗血症発症との相関関係を展示了。
- 今回はLabeventsテーブルの中に、少ないnull測定値を持つ順番によって、35個の特徴に広めて、z-scoreで標準化した。また前回より、もっと良い予測結果に達した。

特徴/Null値の数	特徴/Null値の数
Creatinine 10	Basophils 71
Urea Nitrogen 10	Magnesium 77
Sodium 10	Calcium, Total 86
White Blood Cells 10	Phosphate 92
Chloride 10	Monocytes 109
Bicarbonate 10	pH 176
Anion Gap 10	Lactate 179
Hematocrit 10	Eosinophils 202
Potassium 10	INR(PT) 202
MCV 12	PT 203
Hemoglobin 12	PTT 219
Platelet Count 12	Lymphocytes 240
RDW 12	Specific Gravity 279
MCHC 12	Creatine Kinase (CK) 575
Red Blood Cells 12	pO2 584
MCH 12	pCO2 584
Neutrophils 61	Base Excess 584
	Calculated Total CO2 585

Outputevents特徴抽出

- アイデア: 入院後1日以内の電子診療記録データは予測に有効か。
- 肺炎患者が受けた診療項目を調べた時、恐らく診療方針によって肺炎患者それぞれが受けた診療項目は大きい違いがあるので、データがまばらである。
- 具体的な測定値の代わりに、ある項目を受けた場合に1、受けなかった場合0に分けて、受けた数が一番多い6個の特徴を選択した。

特徴(受けた数)		
Foley(607)	Urine Out Foley(362)	Pre-Admission(212)
Void(189)	Pre-Admission Output Pre-Admission Output(106)	Urine Out Void(71)

Microbiologyevents特徴抽出

- アイデア: 入院後1日以内の原因菌検査データは予測に有効か。
- 検査して出た原因菌を1、出なかった原因菌を0に分けて、原因菌別患者数が10以上の15個の特徴を抽出した。
- 最後に、抗菌薬の感受性検査^[1]を通して、感性であれば1、耐性であれば0に分けて、1個の特徴(抗体有無)を得た。それと、合わせて16個の特徴を得た。

特徴 / 検査して出た原因菌別患者数

CLOSTRIDIUM DIFFICILE	13
PROTEUS MIRABILIS	16
CORYNEBACTERIUM SPECIES (DIPHTHEROIDS)	16
GRAM NEGATIVE ROD #2	17
GRAM POSITIVE BACTERIA	22
KLEBSIELLA PNEUMONIAE	32
ENTEROCOCCUS SP.	34
STREPTOCOCCUS PNEUMONIAE	40
PSEUDOMONAS AERUGINOSA	48
GRAM NEGATIVE ROD(S)	66
STAPHYLOCOCCUS, COAGULASE NEGATIVE	66
ESCHERICHIA COLI	70
POSITIVE FOR METHICILLIN RESISTANT STAPH AUREUS	70
STAPH AUREUS COAG +	127
YEAST	172

[1] <https://www.jslm.org/books/guideline/06.pdf>

Diagnoses特徴抽出

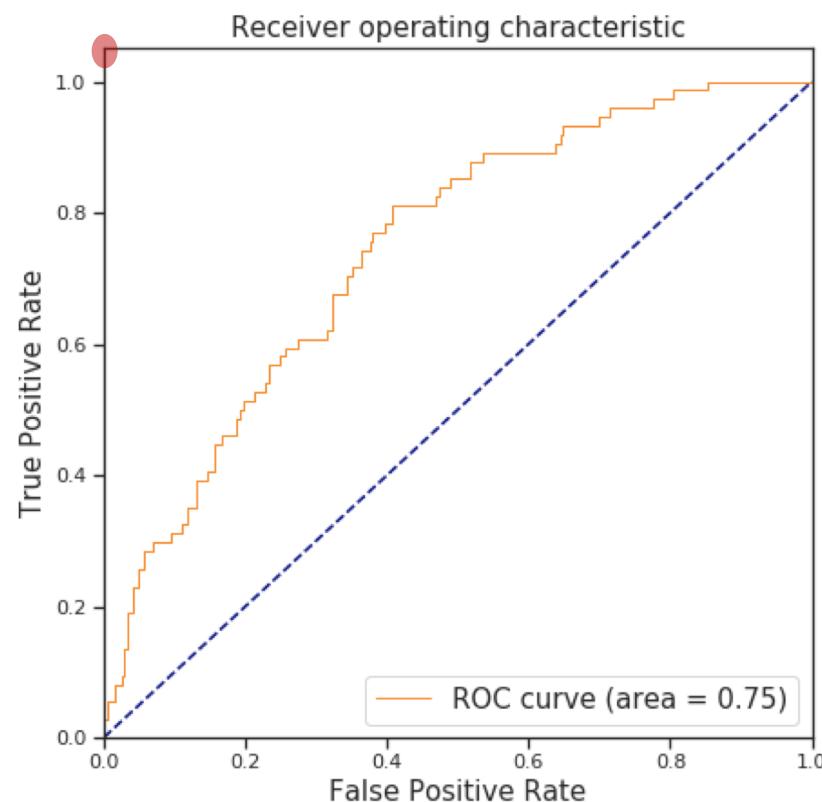
- アイデア: 入院後1日以内の診断データは予測に有効か。
- 問題: 診断された日時を記載されなかったので、1日以内の診断データを限定できない。
- 試みとして、患者の診断履歴はデータベース上にicd9コードで表現されていて、また簡略化のために、診断を19種のカテゴリにまとめなおして特徴としてモデルに取り入れた。

特徴 / 診断された人数

perinatal period	0
childbirth..complications	2
congenital anomalies	32
skin/subcutaneous tissue	233
injury/poisoning	270
neoplasms	317
musculoskeletal	348
sepsis	385
nervous system	567
digestive system	579
mental disorders	594
blood/blood-form organs	709
infectious/parasitic	710
ill-defined conditions	800
genitourinary system	831
external injury/supplmnt	994
metabolic/immunity disorders	1146
circulatory system	1221
respiratory system	1357

*ROC曲線紹介

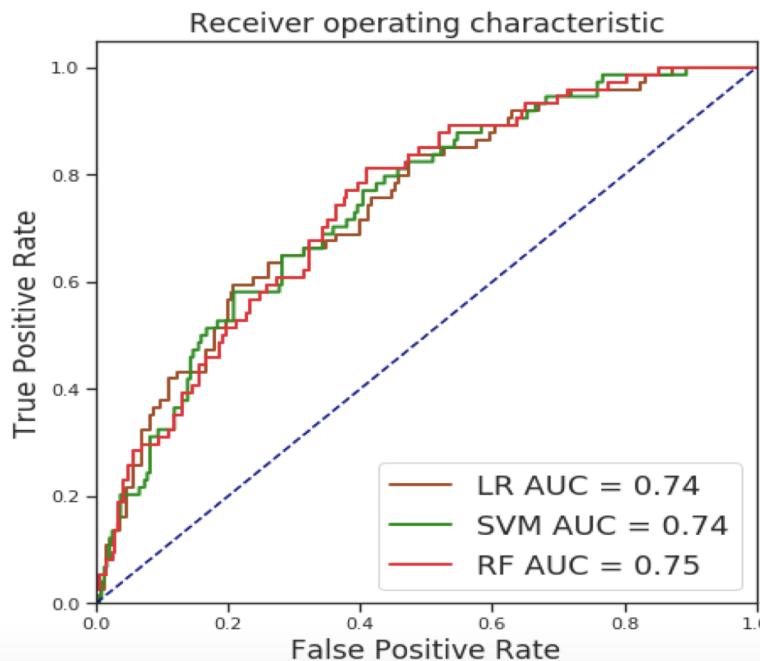
- ROC曲線(Receiver Operating Characteristic curve)は受信者動作特性曲線で、ここに予測方法の良否を区別するために使用する。
- TPRは全部の正例に正しく予測される確率で、FPRは全部の負例に間違って予測される確率である。赤い点は完璧な予測を表示する。
- この曲線の点は違うしきい値によって分類をする結果を表示する。曲線の下面積(AUC)は大きければ大きいほど、予測結果は良くなると判断できる。



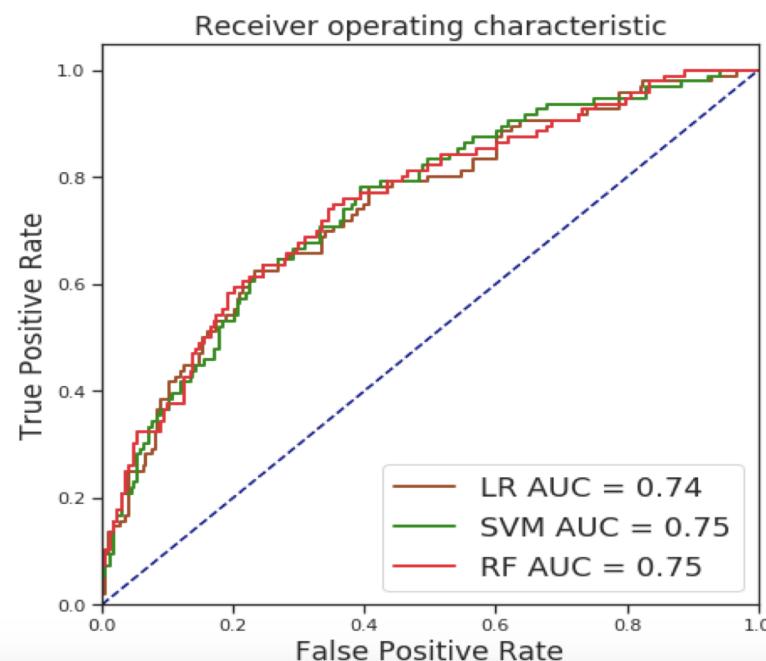
Labevents + Ageだけによる予測

- 1419個のデータから、ランダムで1100個を学習データ、319個をテストデータに分割する。
- 40日以内に死亡した(敗血症を発症した)患者を正例(1)、生存した(発症しなかった)患者を負例(0)とする。
- Grid search、と5分割交差検証で、ハイパーパラメータをチューニングする。

➤ 死亡予測ROC曲線



➤ 敗血症発症予測ROC曲線



特徴の種類別による予測結果

- 違う種類の特徴と三つの機械学習方法の予測結果比較



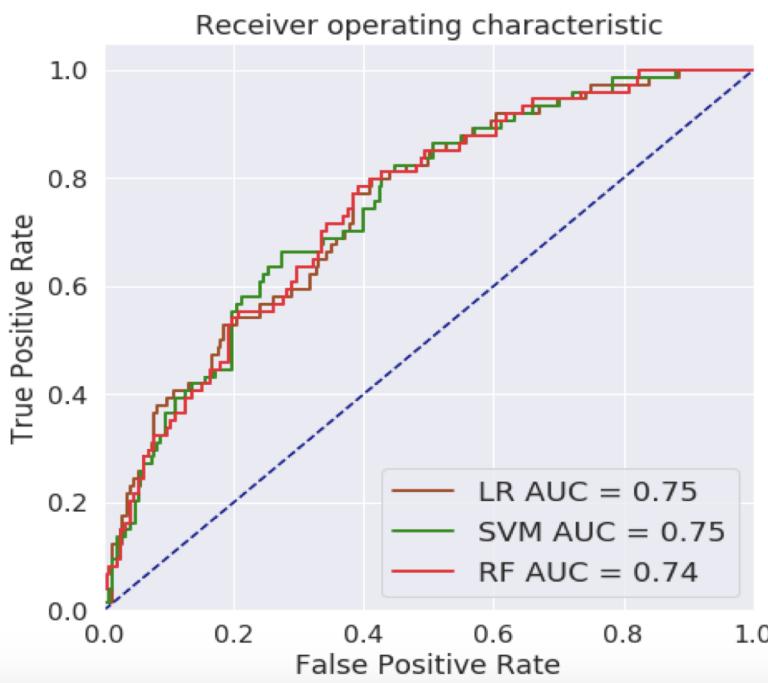
AUC	肺炎死亡予測			敗血症発症予測		
	Labevents + Age	SVM	Random forest	Labevents + age [0.75(RF)]	Diagnose + age [0.87(RF)]	Labevents + age [0.75(SVM or RF)]
Labevents + Age	0.74	0.74	<u>0.75</u>	0.74	<u>0.75</u>	<u>0.75</u>
Outpuťevents + Age	<u>0.64</u>	<u>0.64</u>	0.63	0.63	<u>0.64</u>	0.63
Microbiology events + Age	0.62	0.60	<u>0.63</u>	<u>0.60</u>	<u>0.60</u>	0.59
Diagnoses + Age	0.69	0.69	<u>0.70</u>	0.85	0.86	<u>0.87</u>
Max feature	Labevents + age [0.75(RF)]			Diagnose + age [0.87(RF)] Labevents + age [0.75(SVM or RF)]		

時間を限定できないので、
実際の予測に
使わなかった。

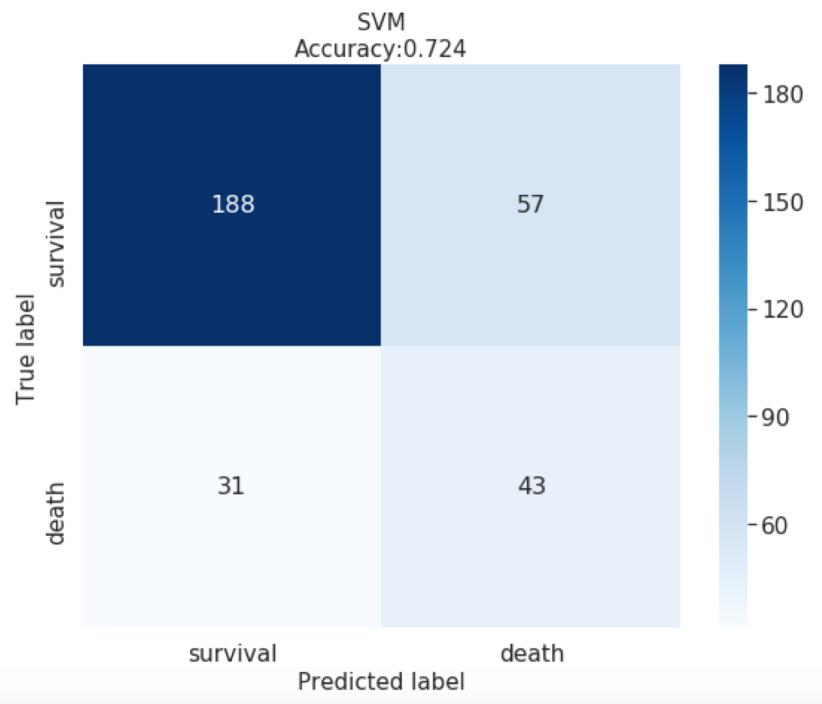
肺炎患者の死亡予測

- Labevents、OutpuťeventsとMicrobiologyeventsに関する特徴を合わせて58個の特徴で予測してみた。
- AUCはLRとSVM方法で0.75に達し、Labeventsの特徴と年齢のみを使う場合のAUCとは同じである。
- テストデータ中に245の負例から188例と、74の正例から43例を正しく予測できる。
- 元々は少ない正例を持つアンバランスデータなので、正例より負例の予測精度はもっと高い。

➤ 予測ROC曲線



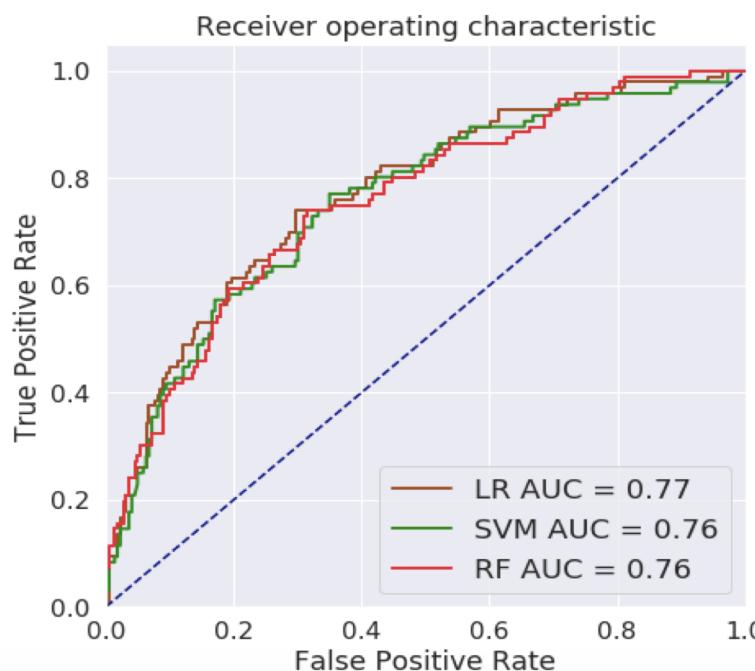
➤ 混同行列



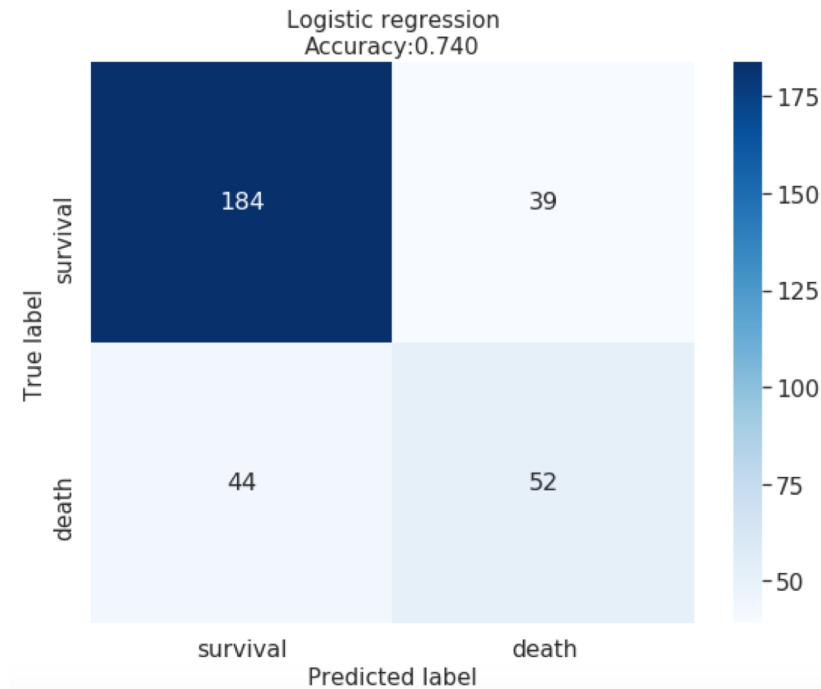
敗血症発症予測

- 肺炎患者の死亡予測で示すように、三つのテーブルから抽出した58個の特徴で予測してみた。
- AUCはLR方法で0.77に達し、Diagnosesの特徴と年齢のみを使う場合のAUCより低いが、任意のテーブルから抽出した特徴と年齢のみを使う場合のAUCより高い。
- テストデータ中に223の負例から184例と、96の正例から52例を正しく予測できる。
- 訓練データはアンバランスデータなので、正例より負例の予測精度はもっと高い。

➤ 予測ROC曲線



➤ 混同行列



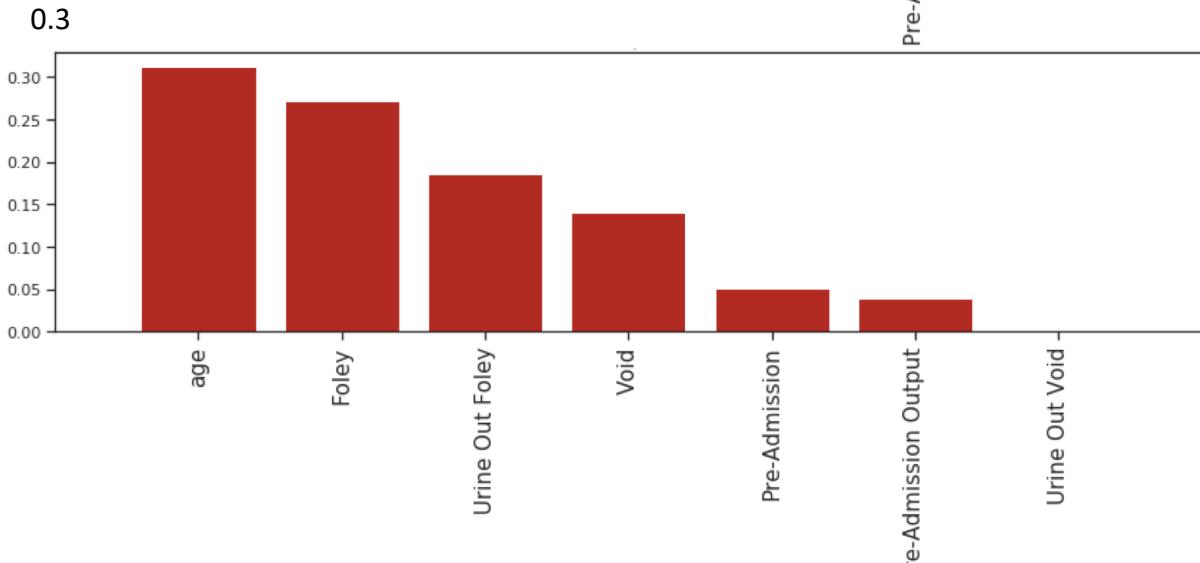
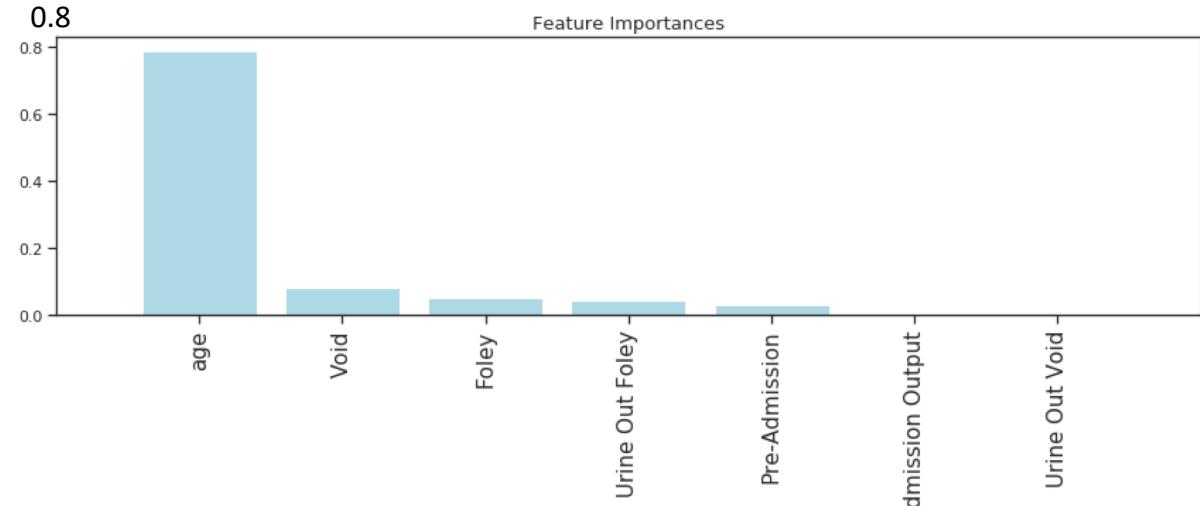


目次

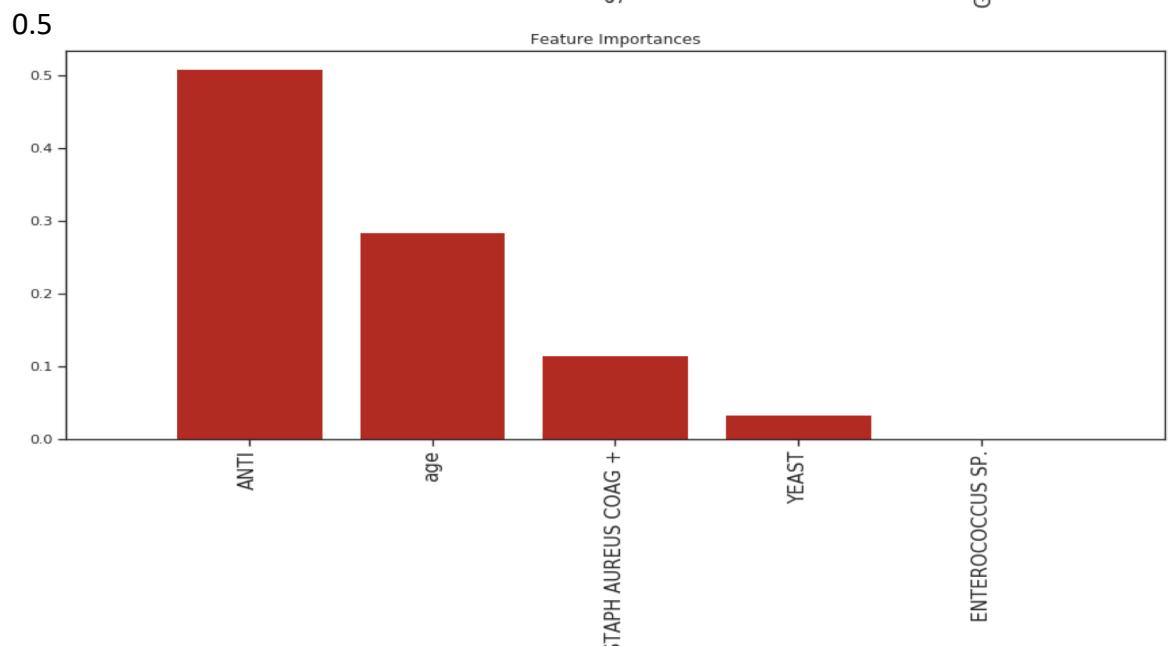
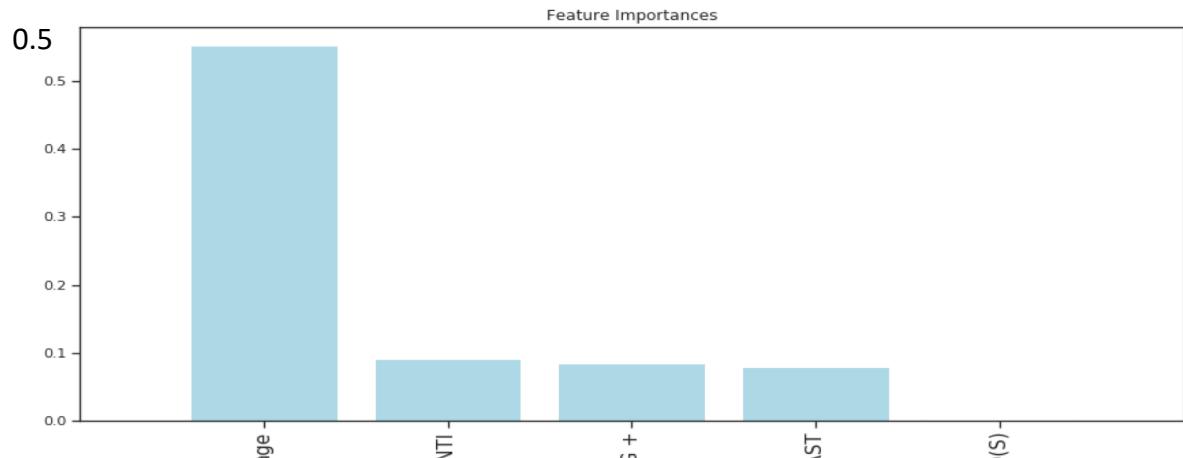
- 1 イントロダクション
- 2 死亡と敗血症発症リスク予測
- 3 特徴分析
- 4 結論と感想

特徴の種類別の重要性分析

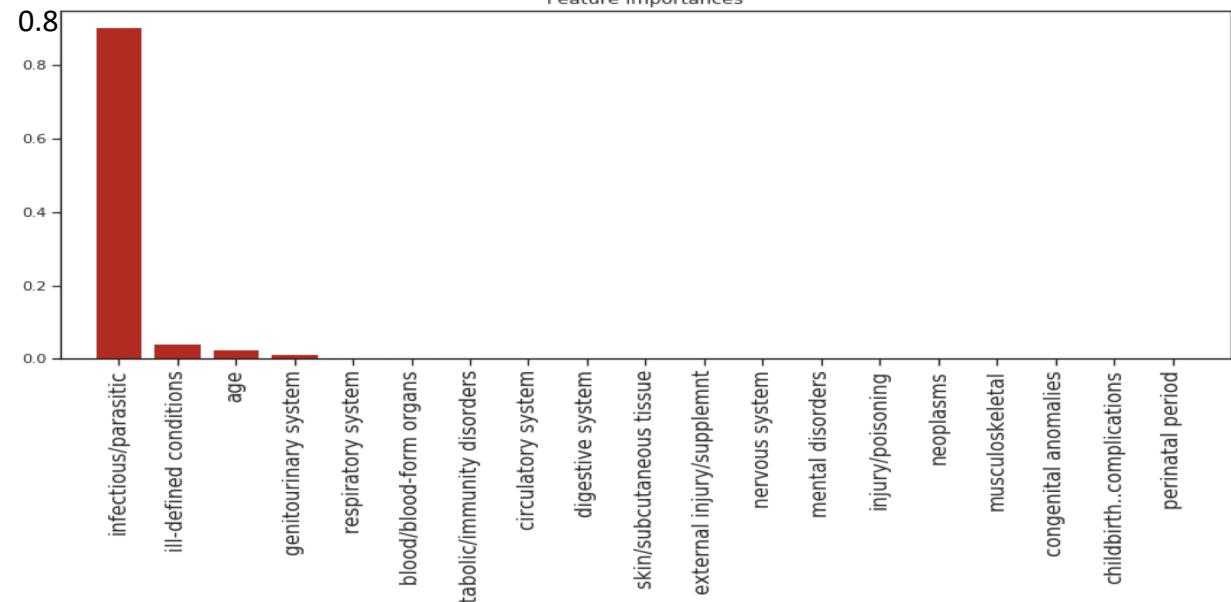
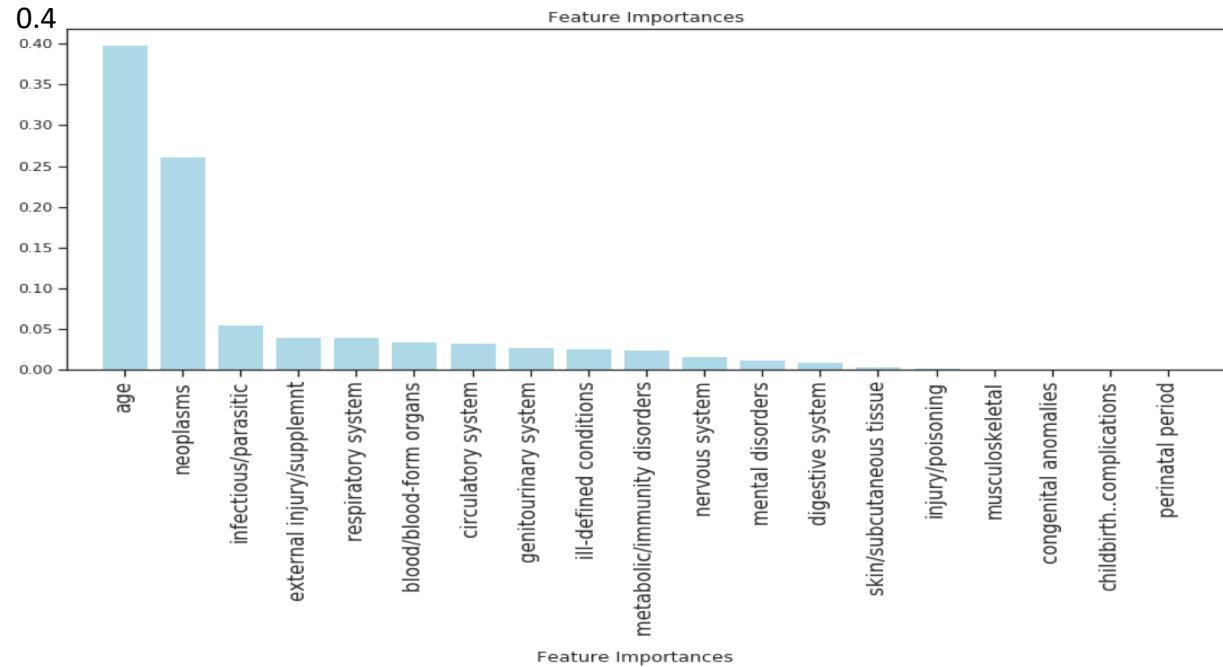
- 年齢と6個のOutpuvents特徴が死亡予測に対して重要かどうか分析する。
- 年齢は他の特徴に比べて重要性が特に高い。Outpuventsから抽出した特徴は死亡予測にほぼ影響しない。
- 敗血症発症予測に対しての特徴重要性を展示する。
- 年齢の重要性は0.3に達し、死亡予測より、敗血症発症の予測に年齢はそれほど重要ではなく、Foley、Urine Out Foleyなどの測定項目は相対的に重要である。



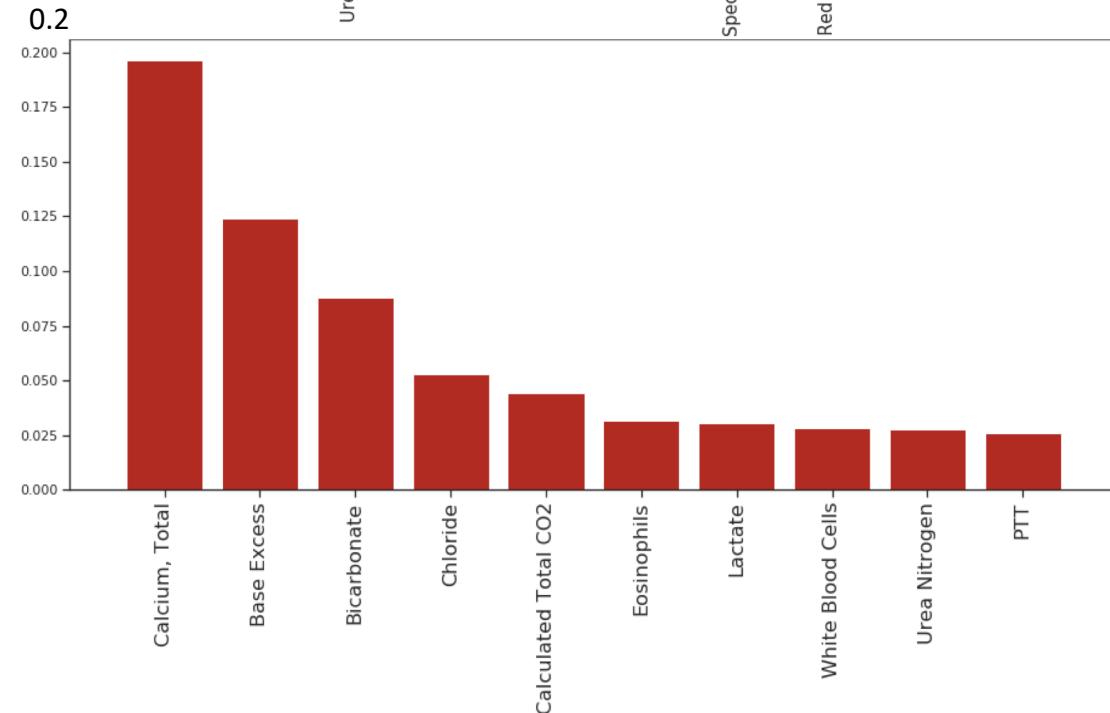
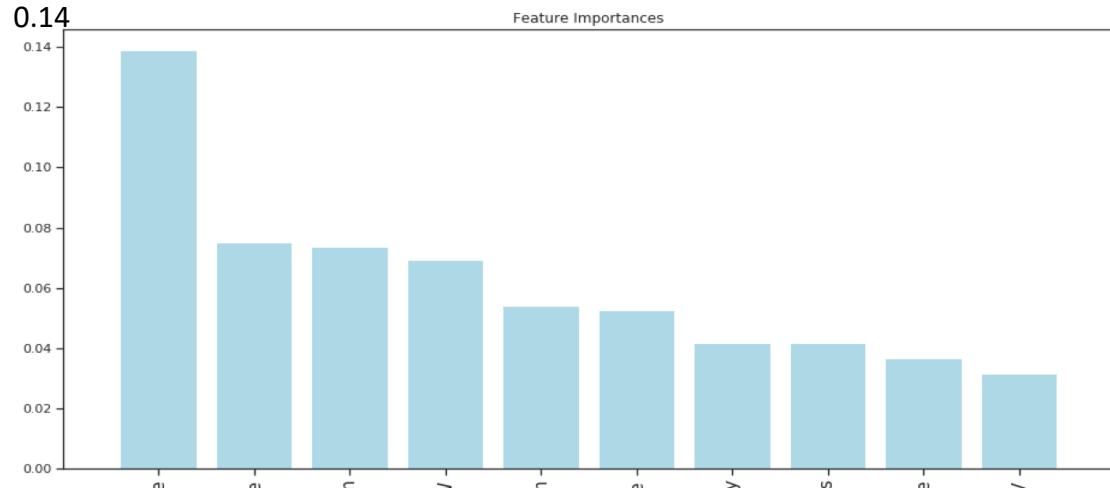
- 年齢とMicrobiologyeventsだけで死亡予測に1番重要な5個の特徴を展示する。
- 年齢の重要性は0.5以上で、抗体の有無と原因菌特徴より重要である。
- 原因菌の特徴はあまり重要ではないと判断できる。
- 敗血症発症予測に1番重要な5個の特徴を展示する。
- 抗菌薬の感受性検査によって判断される抗体有無が意外と1番重要な特徴であった。
- 原因菌の特徴もあまり重要ではない。



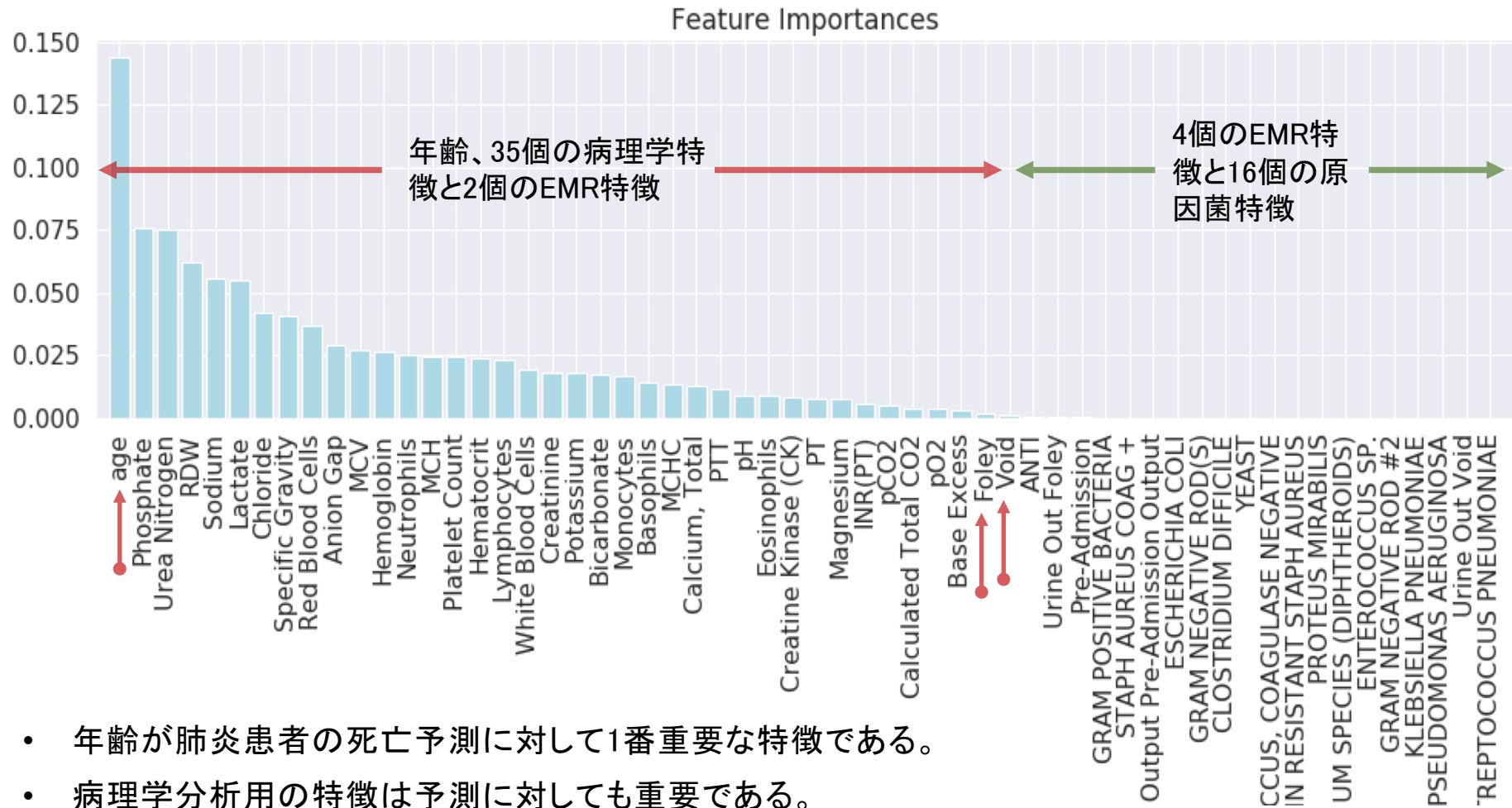
- 年齢とDiagnosesだけで死亡予測をする19個の特徴を展示する。
- 死亡予測に、年齢とNeoplasmsと診断される数は他の診断の数より重要である。
- 敗血症発症を予測する19個の特徴を展示する。
- Infection、parasiticと診断される数が非常に重要で、これもDiagnosesだけを使う場合に予測精度が高い理由と思われる。



- 年齢とLabeventsだけで死亡予測をする1番重要な10個の特徴を展示する。
- 年齢も1番重要であるが、重要性はただ0.14で、他の特徴(いくつかの化合物、赤血球など)も死亡予測に影響する。
- 敗血症発症を予測する1番重要な10個の特徴を展示する。
- 年齢の重要性が低いので、ここに含まられない。
- Calcium、Base Excess、Bicarbonateなどの重要性が相対的に高い。



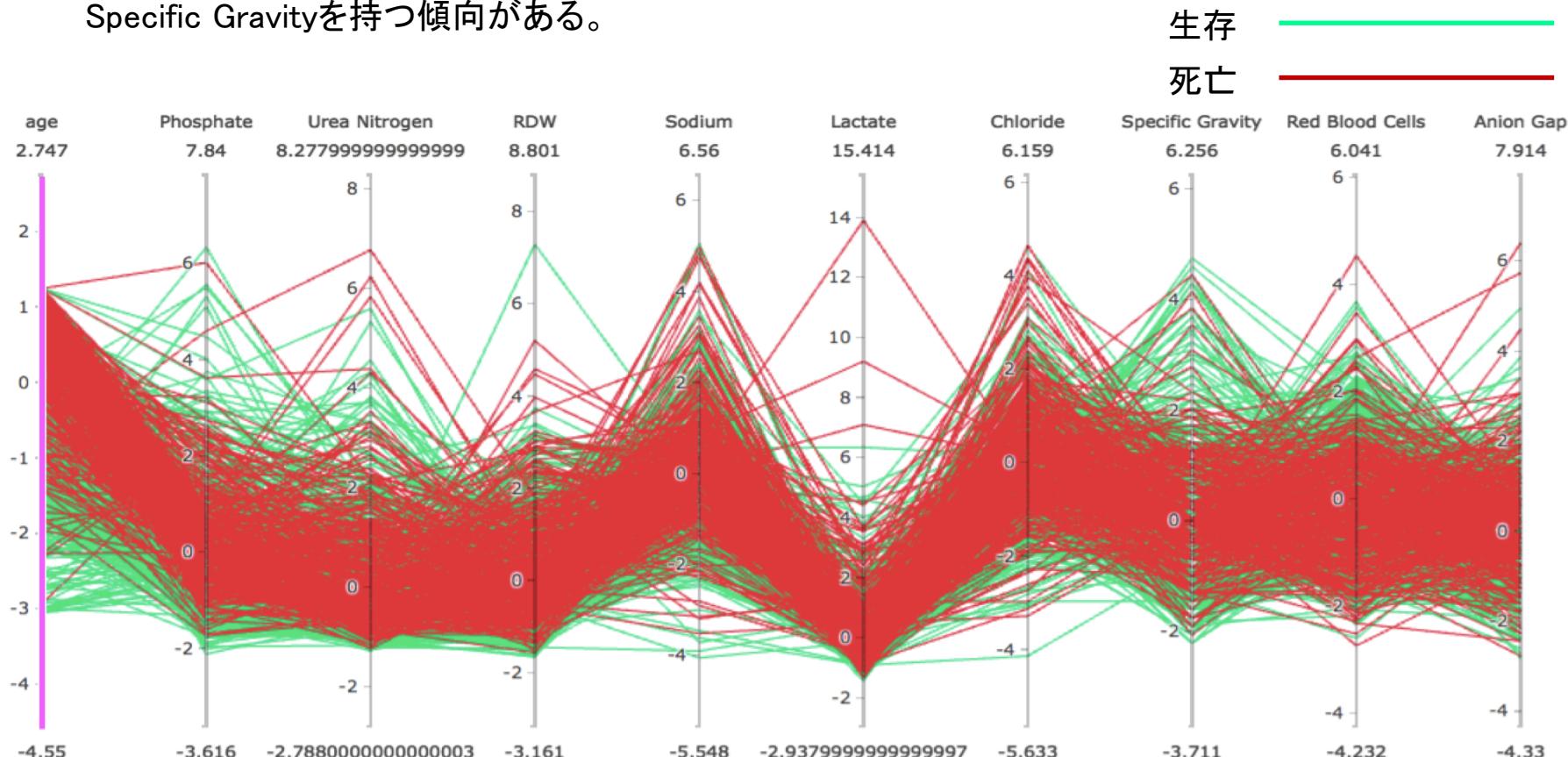
L + M + O + Ageで死亡を予測する特徴重要性



- 年齢が肺炎患者の死亡予測に対して1番重要な特徴である。
- 病理学分析用の特徴は予測に対しても重要である。
- 原因菌検査とEMRは予測に重要ではないと判断できる。

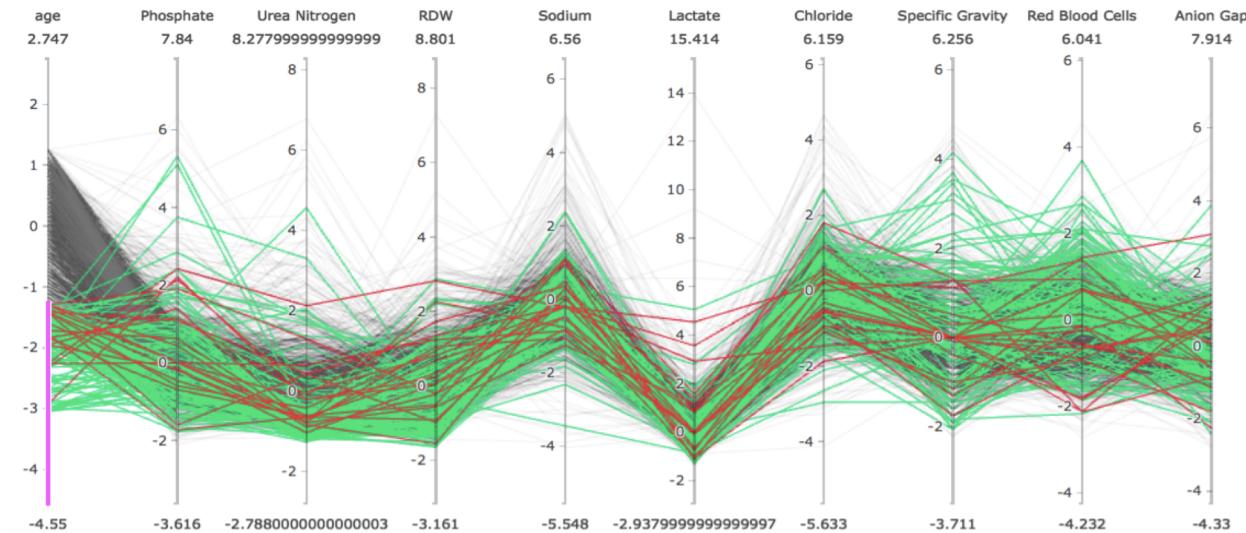
肺炎患者死亡予測の特徴相関性調査

- 肺炎患者の死亡予測に1番重要な10個の特徴を抽出し相関性を調べた。
- 一見すると、40日以内に生存しやすい患者は相対的に低い年齢、高い Specific Gravityを持つ傾向がある。



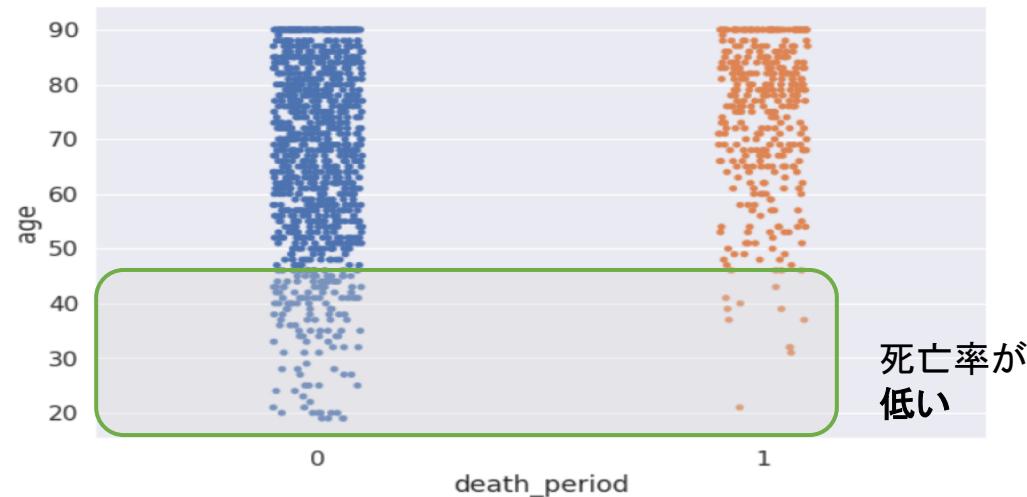
➤ 特徴追跡

- 低い年齢を選択し追跡して、少ない赤の線が展示されたので、年齢が低い患者が生存しやすいと観測できる。

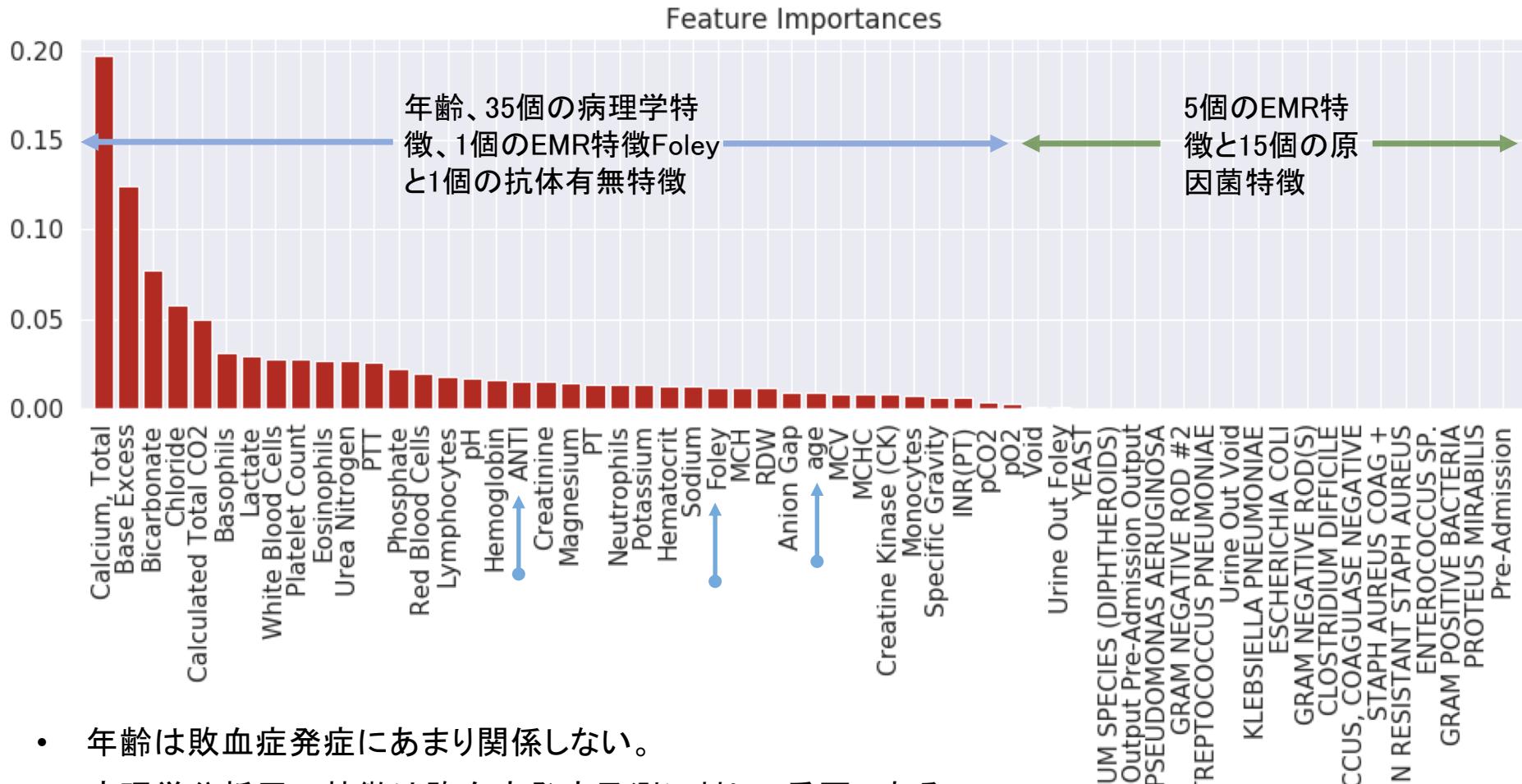


➤ 特徴が生存と40日以内の死亡に関する分布

- 年齢が生存と死亡に関する分布を展示する。0は生存、1は死亡を表示する。
- 患者が45歳ぐらい以下であれば、肺炎を発症するリスクは相対的に低く、また死亡率も低い。



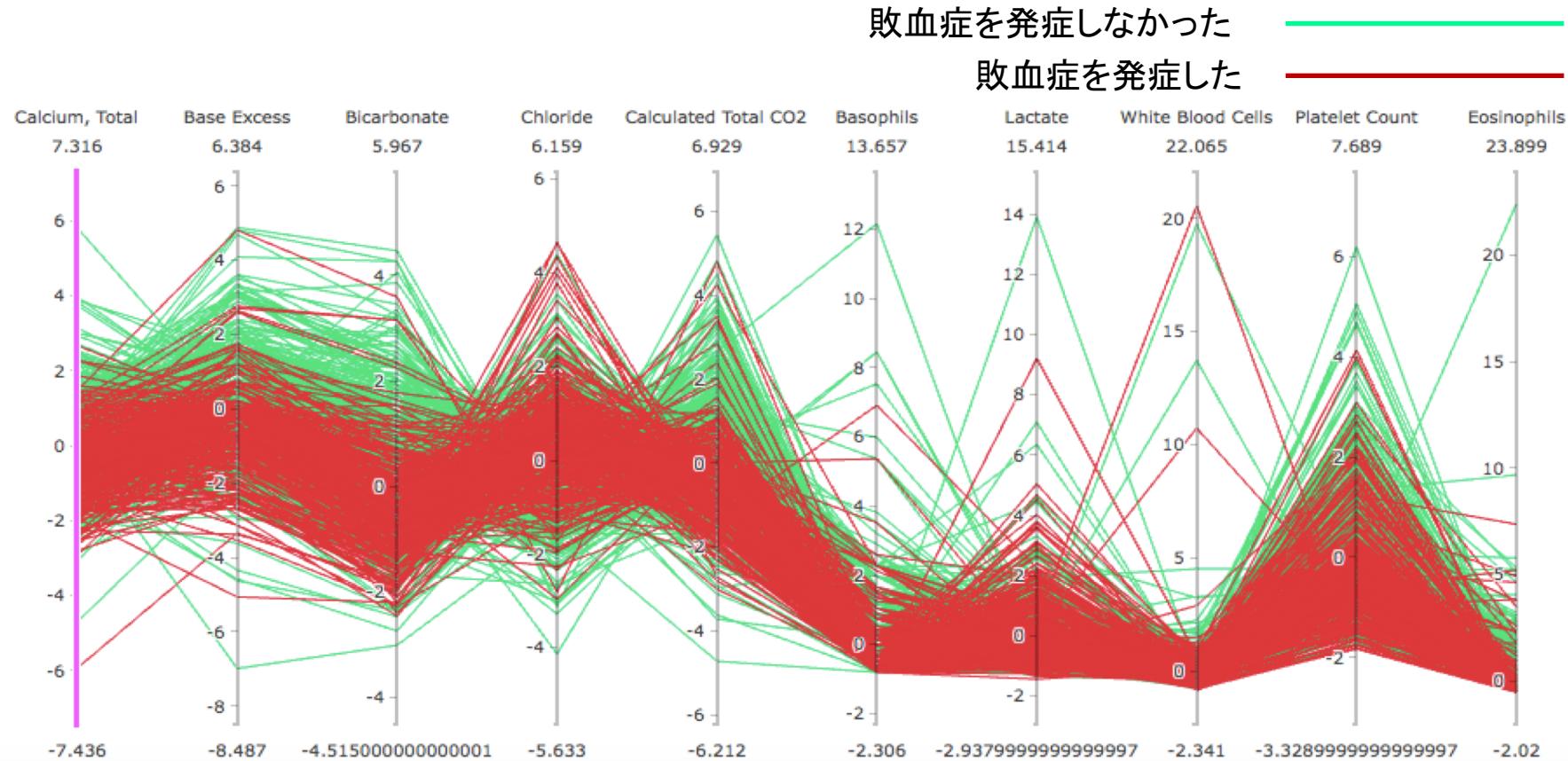
L + M + O + Ageで敗血症発症を予測する特徴重要性



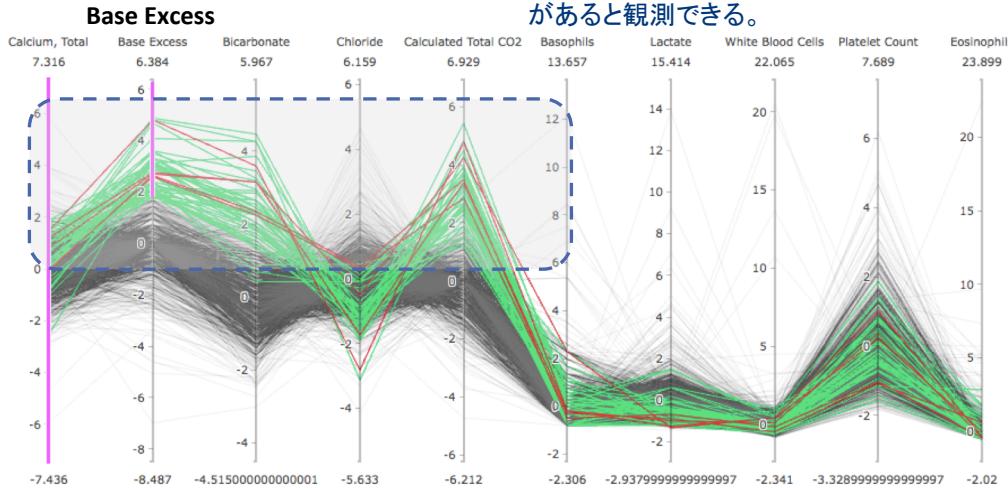
- 年齢は敗血症発症にあまり関係しない。
- 病理学分析用の特徴は敗血症発症予測に対して重要である。
- 原因菌検査とEMRは予測に重要ではないと判断できる。

敗血症発症予測の特徴相関性調査

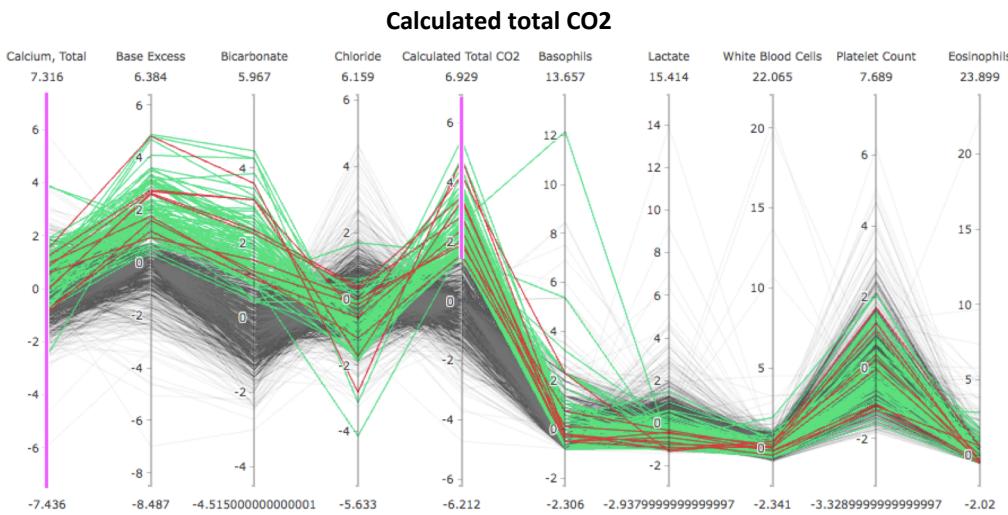
- 敗血症発症予測に1番重要な10個の特徴を抽出し相関性を調べた。
- 一見すると、敗血症を発症しにくい患者は相対的に高いCalcium、高いBase Excess、高いCO₂、と高いBicarbonateを持つ傾向がある。



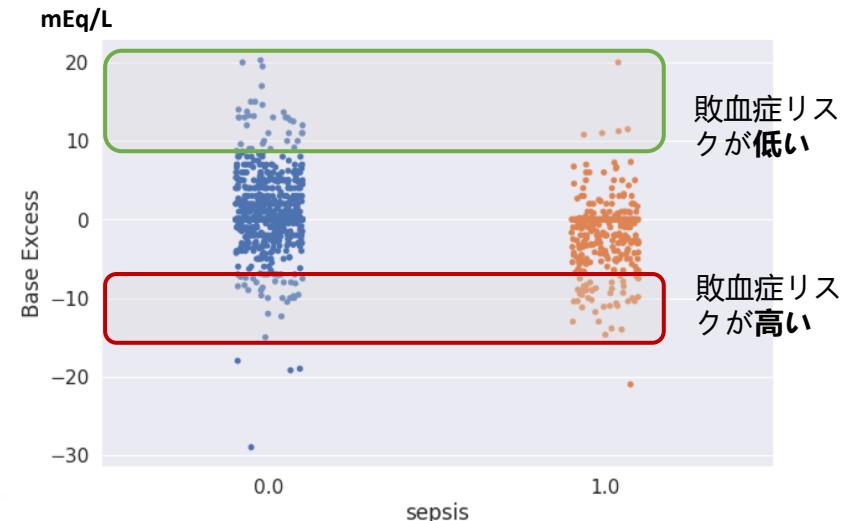
➤ Base Excess特徴追跡



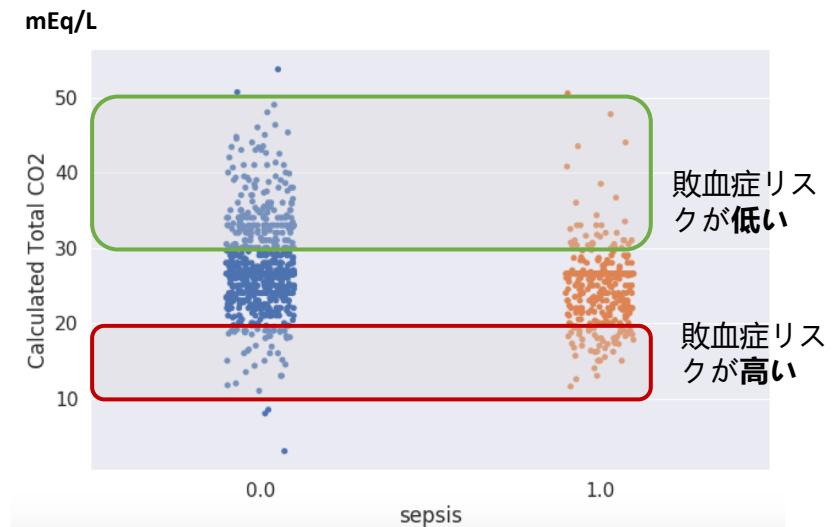
➤ Total CO₂特徴追跡



➤ 敗血症発症に関するBase Excessの分布



➤ 敗血症発症に関するCO₂の分布





目次

- 1 イントロダクション
- 2 死亡と敗血症発症リスク予測
- 3 特徴分析
- 4 結論と感想

結論

- MIMIC-IIIデータベースの肺炎患者の入院後1日以内のデータを用いて、肺炎患者の死亡と敗血症発症予測を進めた。
- 四つの種類の特徴をまとめて、三つの機械学習方法で予測して、最後の40日以内の死亡予測に0.75のAUC、敗血症発症予測に0.77のAUCに達して、有効な予測ができるが、1日分のデータを用いるだけで、予測結果それほど高くないです。
- 最後の特徴分析に、特徴重要性といくつかの特徴の分布の分析を通して、死亡予測に年齢とLabevents、敗血症発症予測にLabevents測定データが重要です。また、特徴の追跡を通して、死亡と敗血症発症予測にいくつかの特徴の間に、高い相関性を観測できた。
- もし将来は機会があれば、これらの特徴相関性をより深く解析したい。また2、3日以内の測定データ、またMIMIC Waveform データベースのチャートデータを用いると、予測精度がもっと高くなると思われる。



ご静聴
ありがとうございました

2019 3.8
