JMDC

医療ビッグデータと機械学習

# 新型コロナウイルス感染時の重症化リスクファクターに関する分析結果

2021年7月9日
株式会社JMDC

# 目次



# 1. 当解析の意義

ワクチン接種が開始された今も、新型コロナウイルス感染症は世界中で猛威を奮っております。その現状の中で、医療ビッグデータを預かっている会社として何か社会に貢献できることがないかと、社内の医療従事者、機械学習エンジニア、コンサルタント、データサイエンティストなどさまざまな分野の専門家が集まり議論をし、これまで溜まってきたデータを活用して、新型コロナウイルス感染時の重症化のリスクファクターを解析することとしました。重症化リスクの高い方が予防をより意識いただくことで、個人の感染を減らし、周囲への伝播も防ぎ、医療機関リソースの圧迫を改善することができれば幸いです。

# 目次



# 2. 手法・結果の概要

## 手法

**データベース**
弊社DPC調査データに基づく
約1,400万人分の医療機関データ

**対象**
2021年2月末までに新型コロナを
理由とした入院患者 7,373 名 ※1

**重症化の定義**
上記入院患者のうちICU入室者を
重症化と定義、185名が該当

**調査項目**
・DPC調査データ様式１各項目
・医薬品投薬歴から判断される
各種既往歴 ※2

## 結果

以下の要素が重症化リスクに有意に関わることが判明

- **肥満　｜重症化リスク1.8倍（p-value=0.013）**
［定義］BMI 25以上。対照群はBMI 18.5～24.9。
- **喫煙　｜重症化リスク1.6倍（p-value=0.042）**
［定義］喫煙指数（１日の喫煙本数×喫煙年数）400以上。
対照群は喫煙指数 0～399。
- **糖尿病｜重症化リスク3.4倍（p-value<0.001）**
［定義］糖尿病治療薬の服薬者。対照群は非服薬者。
- **高血圧｜重症化リスク1.6倍（p-value=0.007）**
［定義］高血圧治療薬の服薬者。対照群は非服薬者。

※1：感染者ではなく入院患者を母集団としている
※2：今回用いた入院患者データのは8割が初診であったため、
既往歴は入院中の当薬歴から推察可能なもののみを対象としている

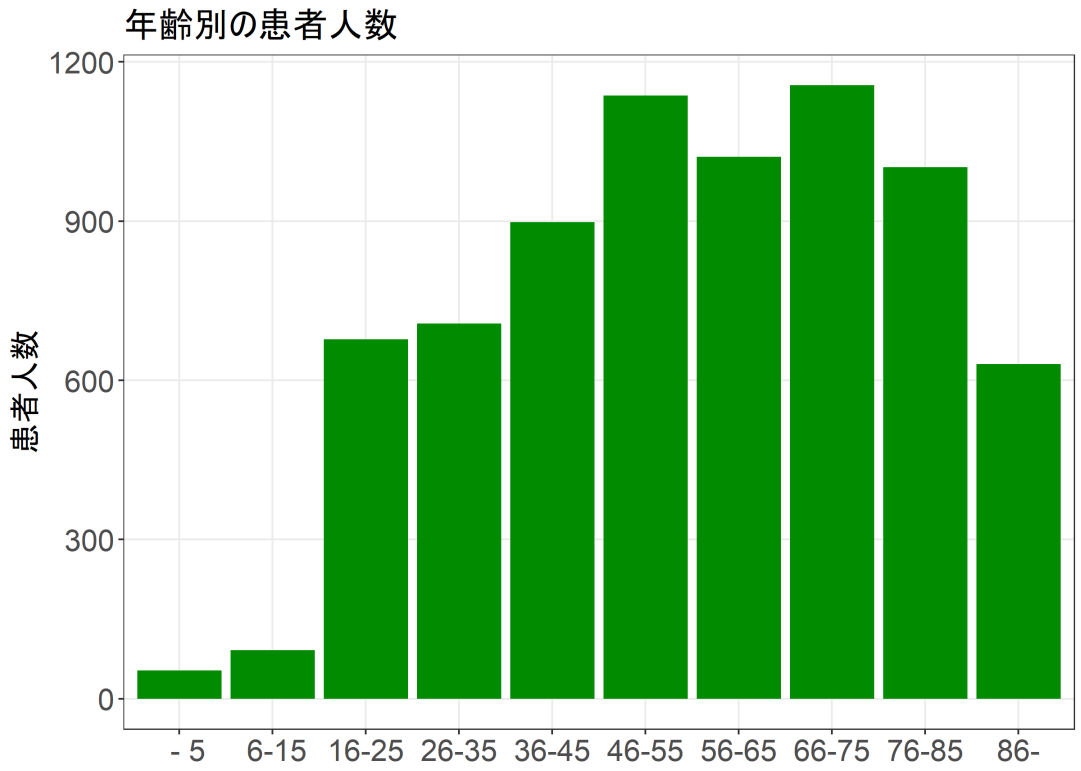
# 目次



# 3-1. 分析対象者の属性



## 年齢別分布




## 補足等

- 年齢別｜高齢に偏らず左図のとおり分布
- 男女比｜男性 55% ：女性 45%
- 初診率｜約 8 割 （入院先での初診割合）

# 3-2. 調査項目について

| 大項目 | 小細目 | 分析上の表記 |
|---|---|---|
| DPC調査<br>様式1 | ・患者年齢 | AGE |
| | ・患者性別コード | GENDER |
| | ・bmi | BMI |
| | ・喫煙指数 | SMOKE_INDEX |
| | ・入院日 | ADM_DATE |
| | ・救急車による搬送 | AMBL |
| | ・入院時_adl01～adl10 | ADL01～ADL10 ※ |
| 既往歴<br>入院期間中の<br>処方歴から判断 | ・糖尿病｜ATC ‘A10’ | DRUG_DM_FLG |
| | ・高血圧｜ATC ‘C02', 'C03’, 'C07’ 'C08', 'C09', 'C11' | DRUG_HP_FLG |
| | ・脂質異常症｜ATC 'C10', 'C11’ | DRUG_HL_FLG |

※ ADL01｜食事、 ADL02｜移乗、 ADL03｜整容、 ADL04｜トイレ動作、 ADL05｜入浴、
ADL06｜平地歩行、 ADL07｜階段、 ADL08｜更衣、 ADL09｜排便、 ADL10｜排尿

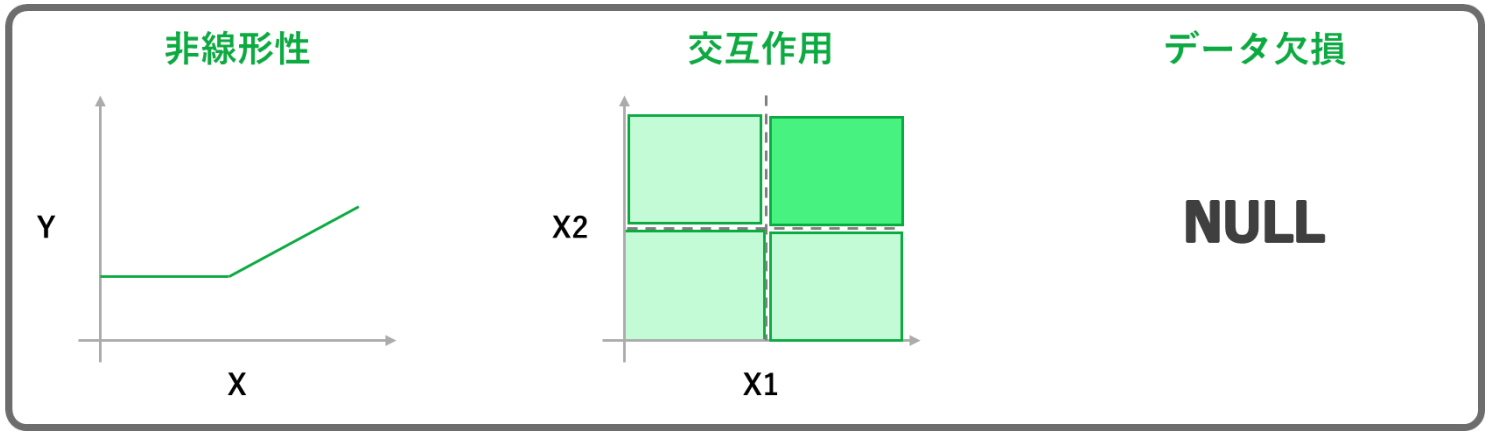
# 3-3. 機械学習アルゴリズム

## 本解析においては 「XGboost ※」 を解析アルゴリズムに適用した

機械学習を用いることで…

- 統計的分析手法で障壁となりがちな事項（下図）に、自動的な対応が可能
- 高精度な予測モデルの構築と同時に、特徴量（関連性のある要素）の重要度も調査が可能
- その特徴量について精緻な統計的検証を行うことで、効率的かつ的確な分析が可能

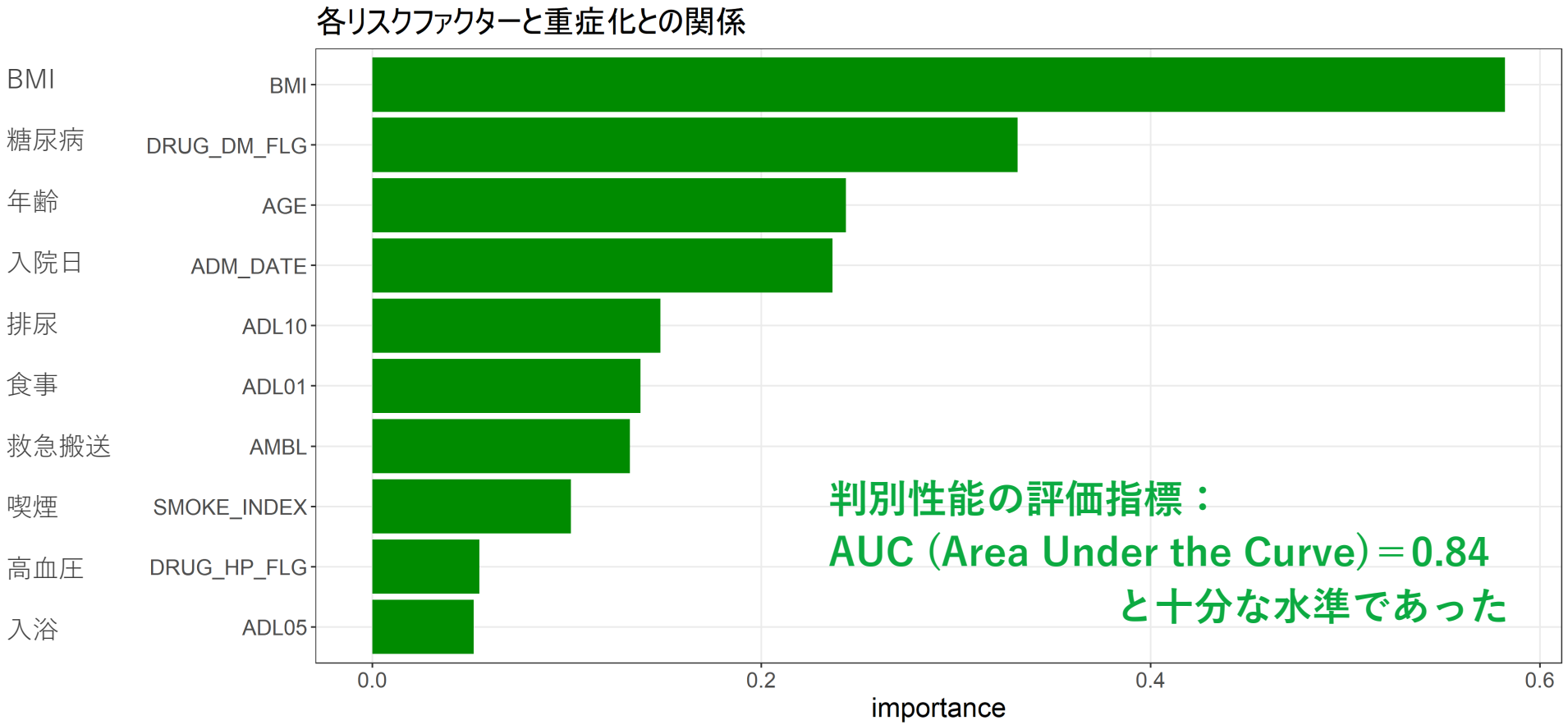
［参考］統計的分析手法において手間のかかる対処事項




※「eXtreme Gradient Boosting」の略。複数の決定木とアンサンブル学習の組み合わせで、非常に高い精度や汎用性があるといわれているモデル。

# 3-4. 機械学習による一次的結果（特徴量・重要度・判別性能）

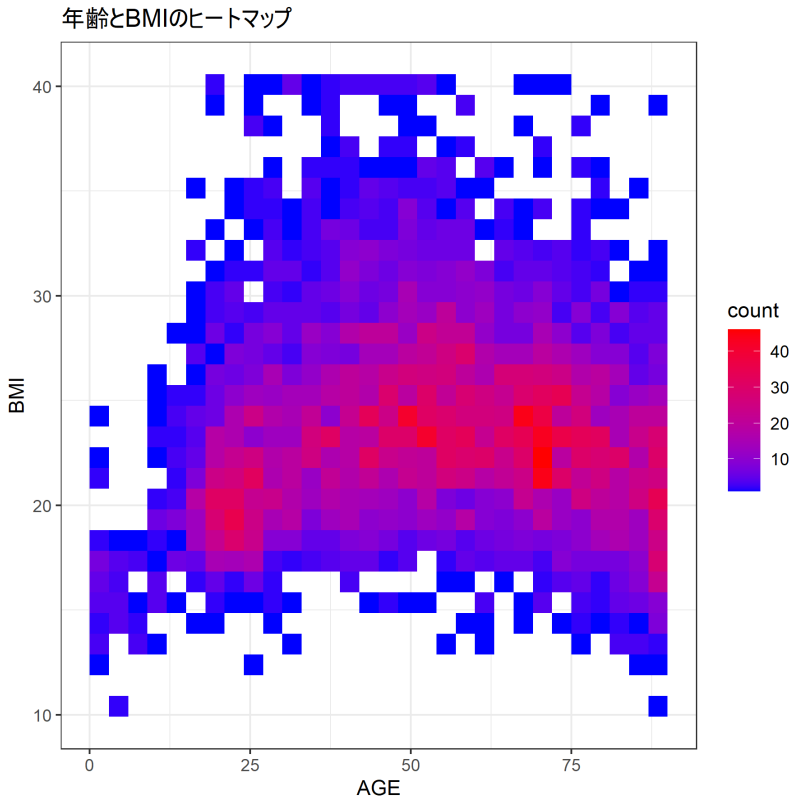
## 以下のとおり、特徴量(リスクファクター)および重要度を検出した

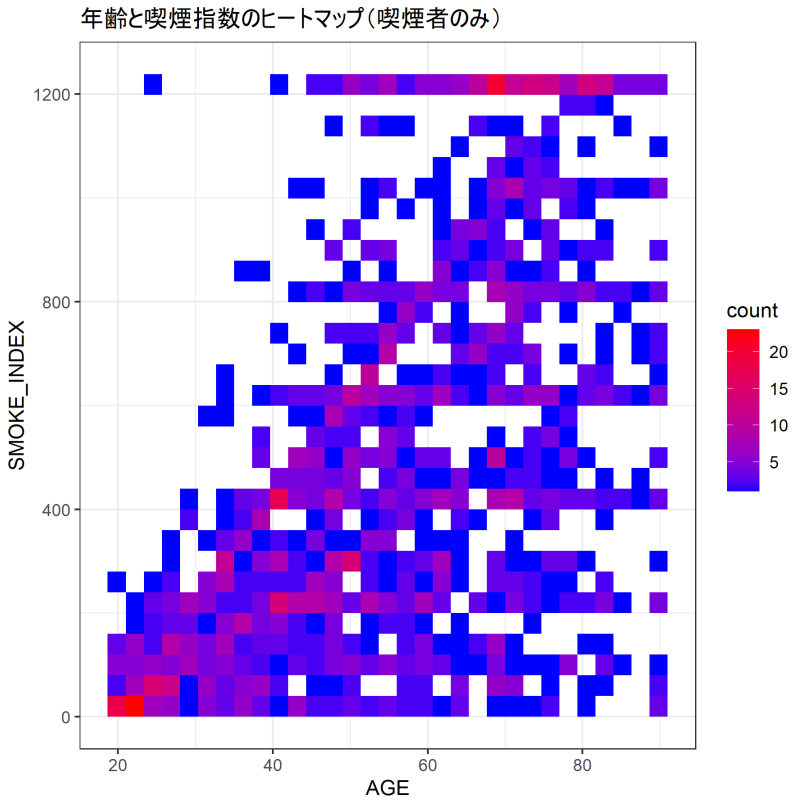



注：重要度(importance)はSHAP値を用いて算出。判別性能は5-fold stratified CVによるAUC merge

# 3-5. 精査①前提としての年齢調整

## BMI・喫煙指数の重症化への影響を判断するには、やはり年齢を考慮した調整が必要



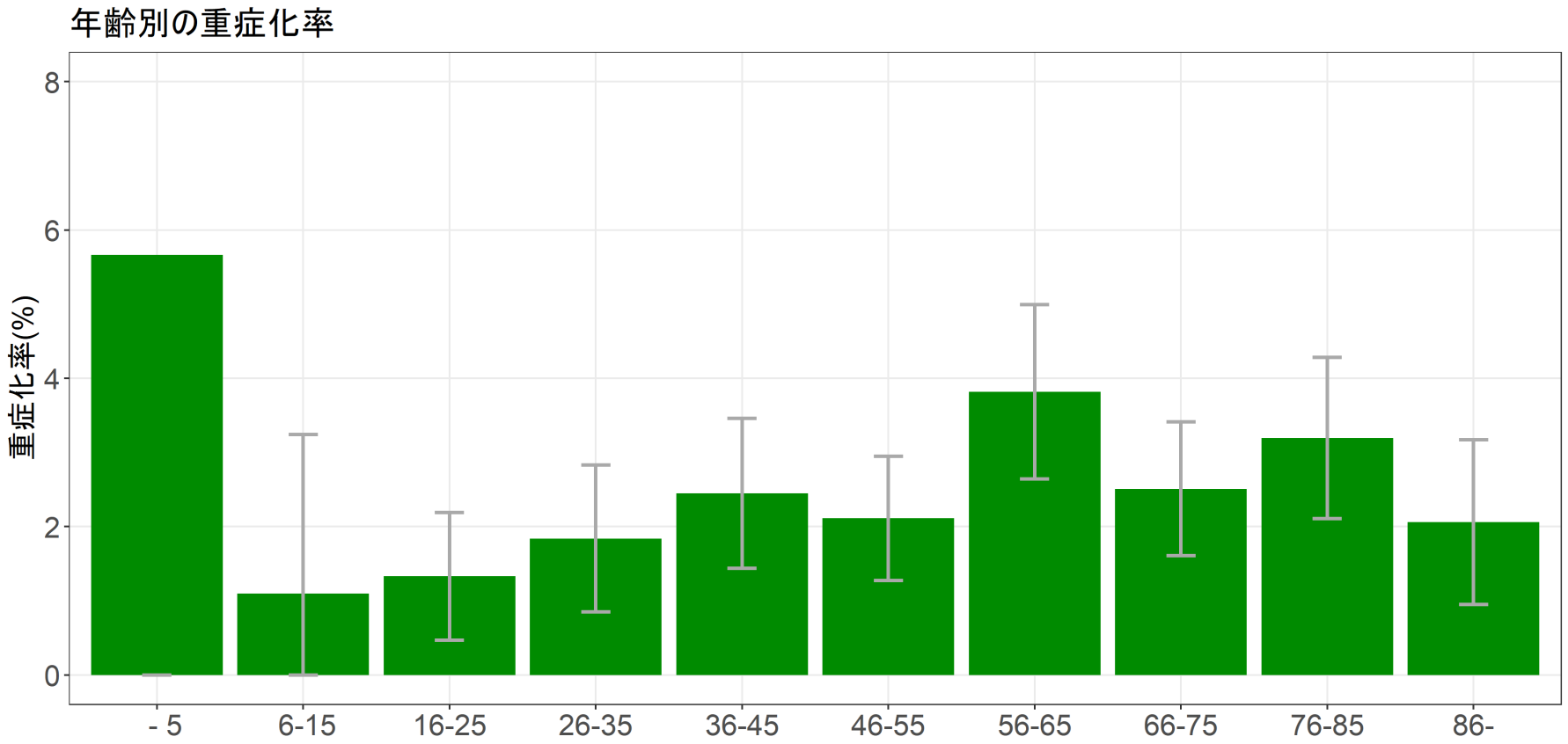




※ ヒートマップ作成において、BMIは上限を40、喫煙指数は上限を1200とした

# 3-5. 精査①前提としての年齢調整



## 年齢別の重症化率は乳幼児と高齢者でやや高い傾向（ただし乳幼児はｎ数が極端に少ない）




※エラーバーは95%信頼区間（以下本資料において同じ）

# 3-5. 精査①前提としての年齢調整

**以上を踏まえ、成人を対象に、56歳で２区分とするダミー変数によって年齢調整を実施**

- ✓ 調べたいファクターはBMI、喫煙指数、糖尿病、高血圧と成人対象の因子なので
  20歳未満のデータは、効果検証用データセットから除外する。

- ✓ 高齢者の重症化率が右肩上がりとはなっていないため、年齢を直接共変量とするのではなく、
  低中年齢と高齢の２区分ダミー変数によって年齢調整を行う。

- ✓ 閾値は構築されたXGBoostにおいて、AGEによる分岐がなされた葉のsplit_valueを参考にする。
  - 66歳、70歳、88歳の区分は高齢層の重症化率のブレを学習したものと推察される。
  - 重症化率のグラフと照らし合わせても妥当と考えられる56歳を閾値として採用する。

AGEのsplit_value分布

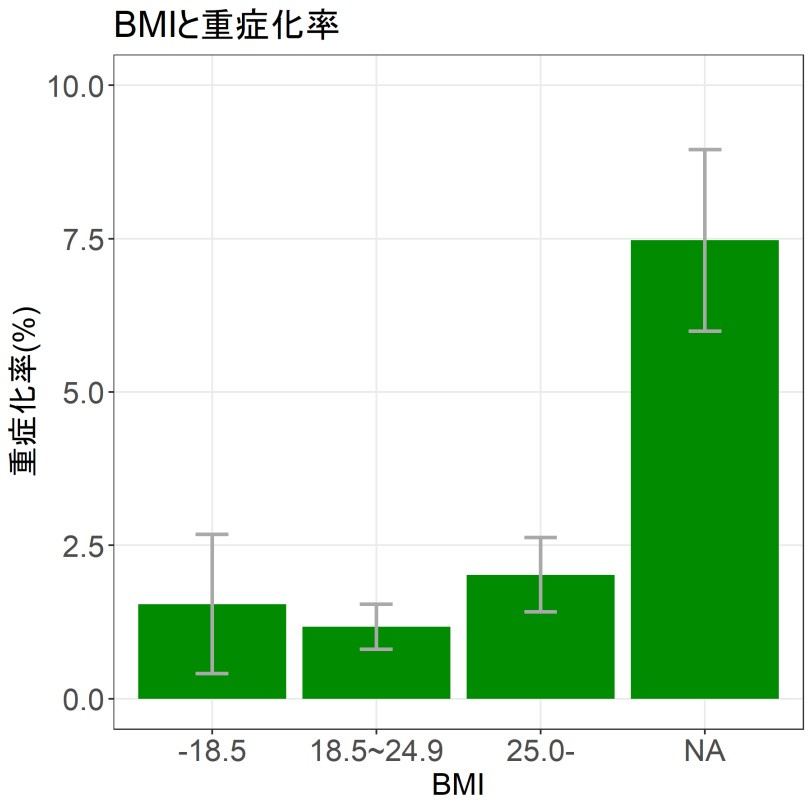
| split_value | 15 | 46 | 53 | 54 | 56 | 66 | 70 | 73 | 80 | 88 |
|---|---|---|---|---|---|---|---|---|---|---|
| 出現回数 | 4 | 4 | 4 | 4 | 5 | 9 | 5 | 4 | 4 | 5 |

※出現回数4回以上のみ記載

# 3-5. 精査② BMI｜単純集計

## 20歳以上を対象としたBMI区分ごとの重症化率の集計結果は以下のとおり

### 分布




### 考察等

- NAの区分（測定不能）の重症化率が高い。
  入院時すでに重症化していて、身長・体重の測定が行われなかった患者の影響と考えられる。
- BMIの特徴量重要度が極端に高かったのは、
  このBMI欠損者の影響と考えられる。
- BMIが測定された者については、
  低体重(BMI<18.5)、肥満(BMI>=25)いずれも
  重症化率が高いことが推察される。

# 3-5. 精査② BMI｜重症化への影響

## 肥満（BMI ≧ 25） は有意なリスクファクターであることが確認された

- BMI 区分の設定
  - 人間ドック学会の基準に準じる。すなわち、
    低体重(BMI<18.5)、標準(18.5<=BMI<25)、肥満(BMI>=25)
  - BMIのsplit_value分布は右図のとおり。
    0はNAを区分するための閾値。14.3は過度に低いため不採用。
    19.1は低体重の閾値18.5に近い。高BMIの閾値が22.8および23と
    やや低いが、概ね近い水準なので人間ドック学会の基準を採用した。

BMIのsplit_value分布

| split_value | 0 | 14.3 | 19.1 | 22.8 | 23 |
|---|---|---|---|---|---|
| 出現回数 | 10 | 13 | 5 | 16 | 5 |

※出現回数5回以上のみ記載

- 検証結果

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5631     0.1861 -24.515   <2e-16 ***
AGE_elderly   0.2364     0.1597   1.480   0.1388
BMI_NA        1.8895     0.1952   9.679   <2e-16 ***
BMI_low       0.2579     0.4138   0.623   0.5330
BMI_high      0.5604     0.2243   2.499   0.0125 *
```

✓ 肥満(BMI>=25) は有意なリスク因子であり、
重症化リスクへの効果は1.8倍(=exp[0.5604])だった。

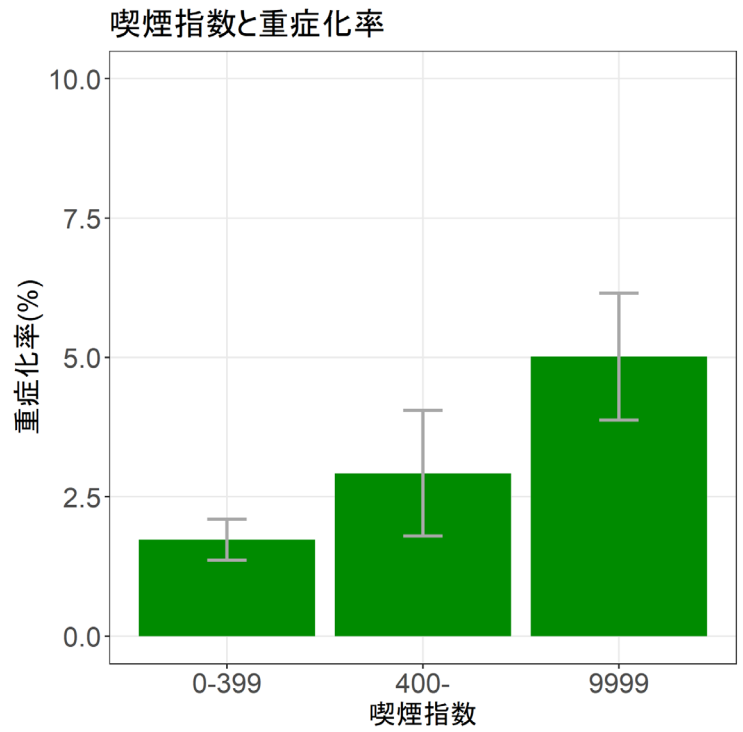
✓ 低体重(BMI<18.5)は有意とはならなかったが、
重症化リスクへの効果は1.3倍(=exp[0.2579])だった。

# 3-5. 精査③ 喫煙指数｜単純集計



## 20歳以上を対象とした喫煙指数区分ごとの重症化率の集計結果は以下のとおり

### 分布




### 考察等

- 9999の区分（不明）の重症化率が高い
  BMI同様、入院時すでに重症化している患者と相関の強いカテゴリーと考えられる。
- 喫煙指数が取得できた者については、
  ヘビースモーカー（喫煙指数400以上）の重症化率が高いことが推察される。
- 参考 ※

| 1日の平均喫煙本数×これまでの喫煙年数 | 喫煙によりがんになる危険度の目安 |
|---|---|
| 200以上 | 禁煙外来対象者 |
| 400以上 | 肺がんが発生しやすい状態 |
| 600以上 | 肺がんの高度危険値 |
| 1000以上 | 喫煙者の<br>喉頭がん発症者平均値 |
| 1200以上 | 肺がんに加え<br>喉頭がんの危険性が激高 |

※ Source｜https://www.pref.iwate.jp/_res/projects/default_project/_page_/001/021/231/kangoka.pdf

# 3-5. 精査③ 喫煙指数｜重症化への影響



**喫煙指数≧400の区分は有意なリスクファクターであることが確認された**

- 喫煙指数の区分の設定
  - split_value分布を参考に検討した。
    －100や60は低水準。
    喫煙年数が短い＝直近の健康状態が良好であることを示唆し、
    実際に重症化率が低いゾーン
    －3600はNAを区分するための閾値として設定
    －1590はヘビースモーカーだが高すぎると判断
  - 上記を踏まえて400を閾値とする。

SMOKE_INDEXのsplit_value分布

| split_value | 60 | 100 | 400 | 1590 | 3600 |
|---|---|---|---|---|---|
| 出現回数 | 4 | 6 | 3 | 3 | 5 |

※出現回数3回以上のみ記載

- 検証結果

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2242     0.1444 -29.251  < 2e-16 ***
AGE_elderly   0.3387     0.1584   2.138   0.0325 *
SMOKE_NA      1.0807     0.1648   6.557 5.48e-11 ***
HeavySmoker   0.4735     0.2330   2.032   0.0421 *
```

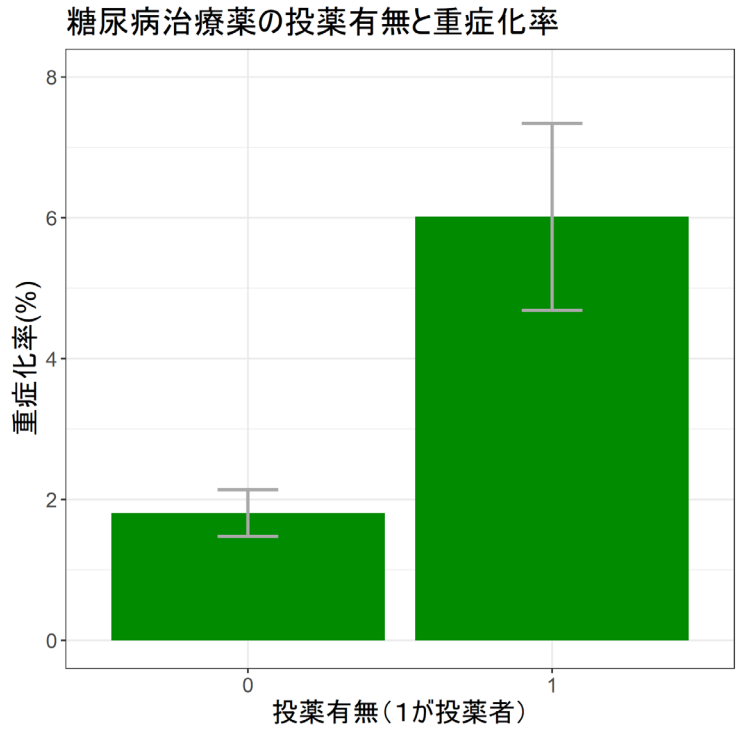
✓ 喫煙習慣(喫煙指数>=400) は有意なリスク因子であり、
重症化リスクへの効果は1.6倍(=exp[0.4735])だった。

# 3-5. 精査④ 糖尿病｜重症化への影響



**入院中の当薬歴から判断した糖尿病は、有意なリスクファクターであることが確認された**

## 分布




## 検証結果

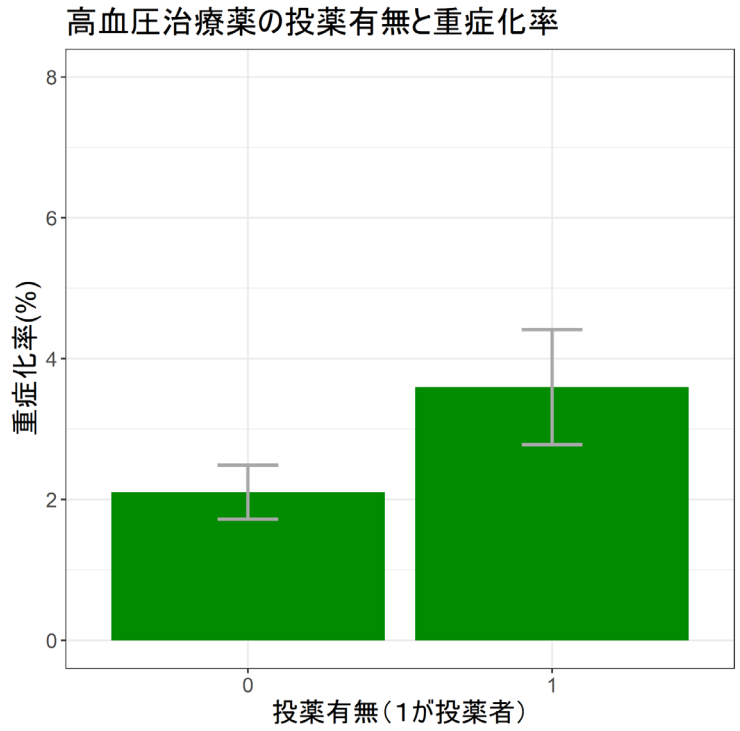
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.03421    0.12877 -31.328  < 2e-16 ***
AGE_elderly  0.06933    0.16637   0.417    0.677
DRUG_DM_FLG  1.22915    0.16401   7.494 6.66e-14 ***
```

✓ 糖尿病は有意なリスク因子であり、重症化リスクへの効果は3.4倍(=exp[1.22915])だった。

# 3-5. 精査⑤ 高血圧｜重症化への影響

**入院中の当薬歴から判断した高血圧は、有意なリスクファクターであることが確認された**

## 分布




## 検証結果

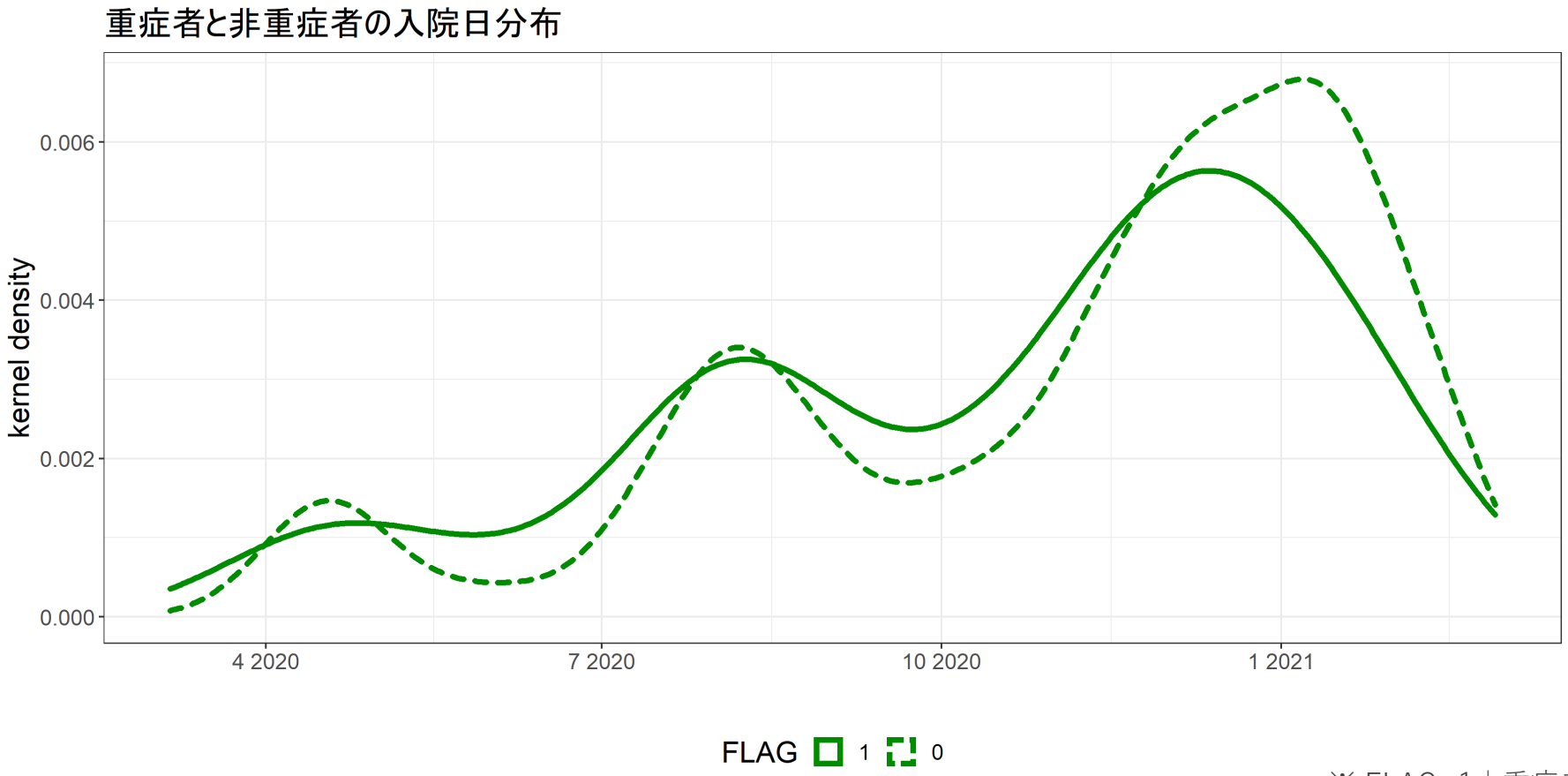
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9270     0.1262 -31.108  < 2e-16 ***
AGE_elderly   0.2073     0.1735   1.195  0.23216
DRUG_HP_FLG   0.4580     0.1712   2.675  0.00748 **
```

✓ 高血圧は有意なリスク因子であり、重症化リスクへの効果は1.6倍(=exp[0.4580])だった。

# 参考｜重要度上位のその他リスクファクター

# 参考① 入院日

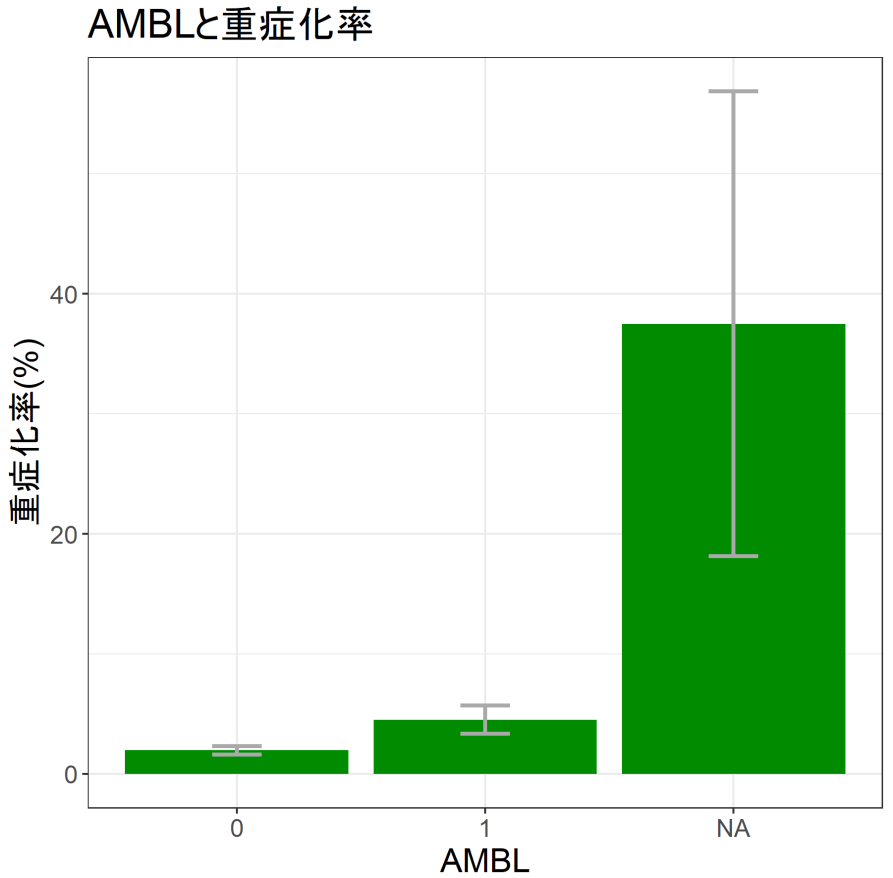
第1波～第3波の各ピーク付近ではICU病床の不足が示唆される




※ FLAG=1｜重症者

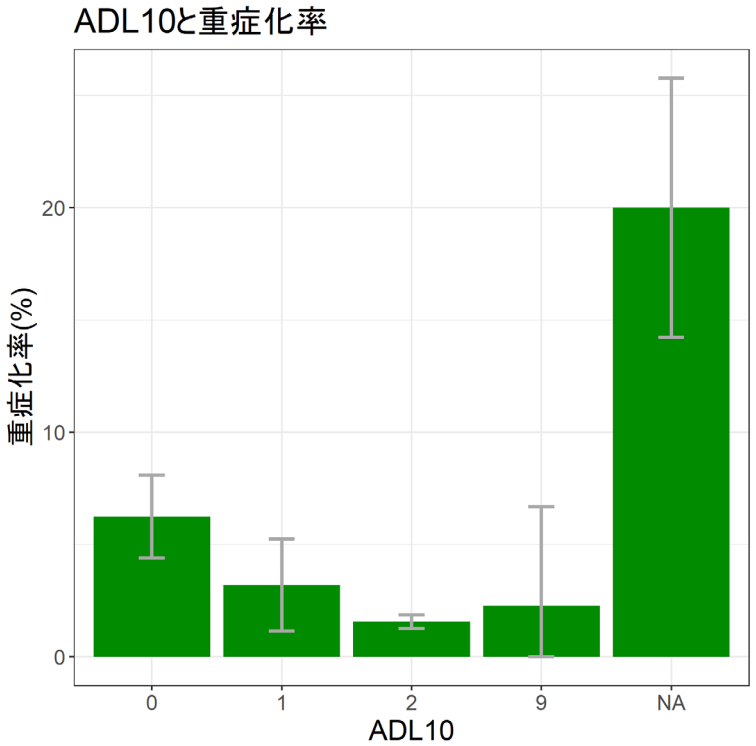
# 参考② 救急搬送

救急搬送者（AMBL=1）の重症化率がやや高い（4.5%）が、それ以上に欠損者の重症化率が高い

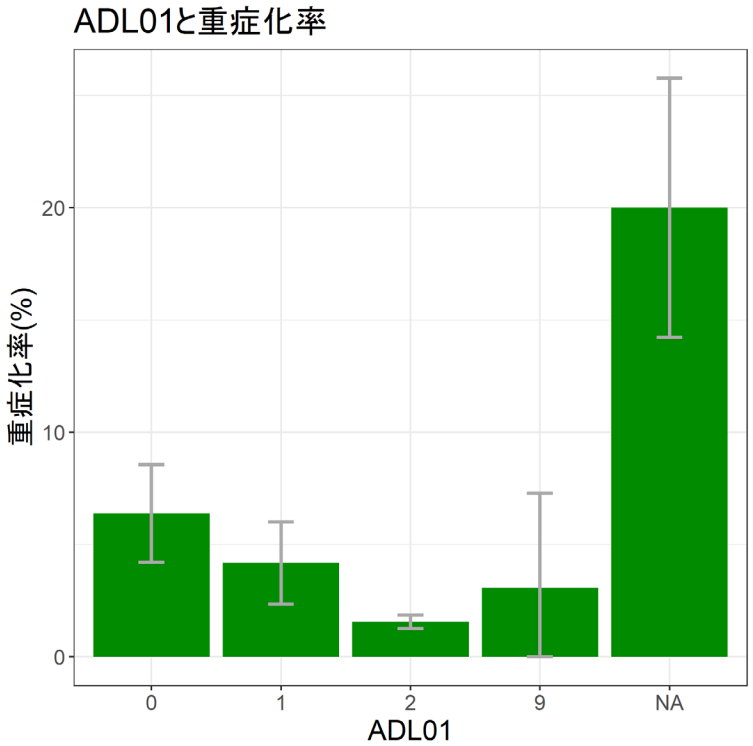


# 参考③ ADL

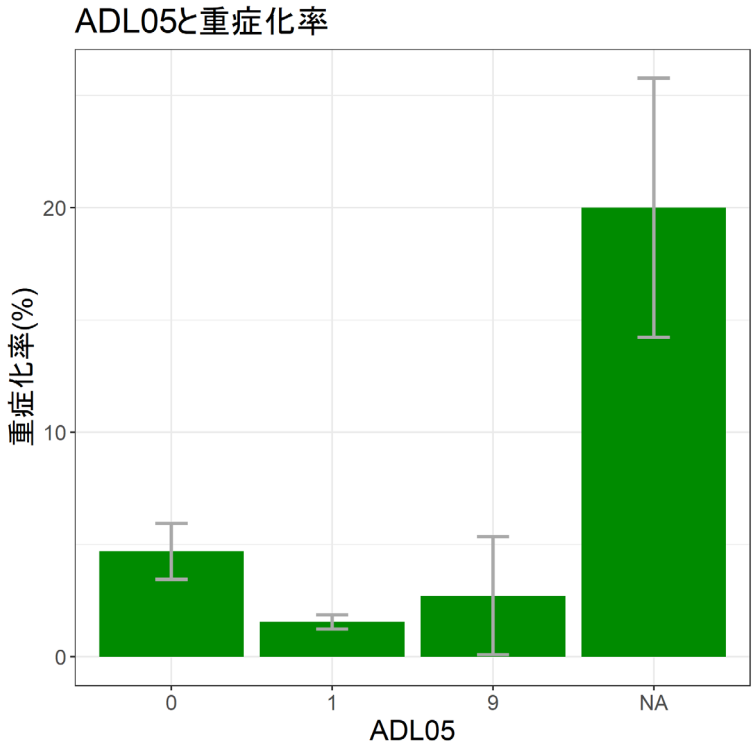
ADLも欠損者の重症化率が高い（各ADL項目の欠損者は同一人物）
なお、非欠損者では介助度が高いほど重症化率が高い傾向




※ ADL10(排尿)
0=失禁、1=時々失敗、2=自立、9=不明




※ ADL01(食事)
0＝全介助、1=一部介助、2=自立、9=不明




※ ADL05(入浴)
0＝全介助、1=自立、9=不明

本資料は、株式会社JMDC（以下「当社」といいます。）が、日本国内の一般の方を対象として作成した資料であり、医療従事者及び日本国外の方に対する情報提供を目的として作成されたものではありません。

本資料に記載される分析対象事実に関する評価については、本資料の作成時点における当社の判断又は考えにすぎず、分析対象事実以外の事実を考慮し、又は本資料の作成時点以降の新たな事実又は技術を考慮した場合、実際の評価の内容が本資料記載の内容又はそこから推測される内容と大きく異なる可能性があります。

今後の状況の変化等が本資料の内容に影響を与える可能性がありますが、当社は、本資料を更新、修正又は確認する義務を負うものではありません。

本資料の内容は事前の通知なく変更されることがあります。新型コロナウイルスに関する最新の情報については、厚生労働省その他の信頼のおける機関の情報を確認し、また医師その他の医療専門家の診断を受けることを推奨いたします。

JMDC

jmdc-pr@jmdc.co.jp

株式会社JMDC
経営管理部 広報担当
03-5733-5010