

千年カルテのデータを活用した 診断支援予測モデルの開発

株式会社ログビー

都立産業技術大学院大学医療情報研究所

松田敦義

アジェンダ

- ① COVID-19の重症度別の薬剤・入院時併存症の集計と死亡予測モデルの開発 中間結果
 - ・ 重症度別の薬剤・入院時併存症の集計
 - ・ 死亡予測モデルの開発
- ② 次世代医療基盤法に基づく匿名加工医療情報を利用した入院後敗血症の予測モデルの開発における課題と対策の検討

COVID-19の重症度別の薬剤・入院時併 存症の集計と死亡予測モデルの開発 中間結果

都立産業技術大学院大学 医療情報研究所	<u>松田敦義</u> 、鹿内緑久
宮崎大学医学部附属病院	荒木賢二、古畑宏樹

COVID-19患者の重症度別の薬剤・入院時併存症の集計

COVID-19患者の重症度の予測



トリアージや医師の負担軽減

本研究では、特徴量探索のため集計を実施



システム概要（対象データ）

医療機関	共同研究機関18施設
入院期間	2019年4月1日～2021年9月30日
患者数	COVID-19入院患者 延べ6,050人
データ種類	DPC（入院時併存症、投与薬剤 等）

システム概要（前処理）

共通

- 次の5項目をキーとしてDPCデータ（FF1, EF）を統合；
- ①施設ID ②患者ID ③入院年月日 ④診療年月日 ⑤データ取込み年月日
- ※ 複数インデックスによる検索効率化、また患者データの重複の防止
- 前処理①：薬剤データ区分：内服, 頓服, 皮下筋肉内, 静脈内に限定
- 前処理②：レセプト電算コード：薬剤に該当するコードに限定
- 前処理③：欠損値削除
- 前処理④：カテゴリデータのフラグ化

前処理による データ数増減

6,050件→ 3,366件

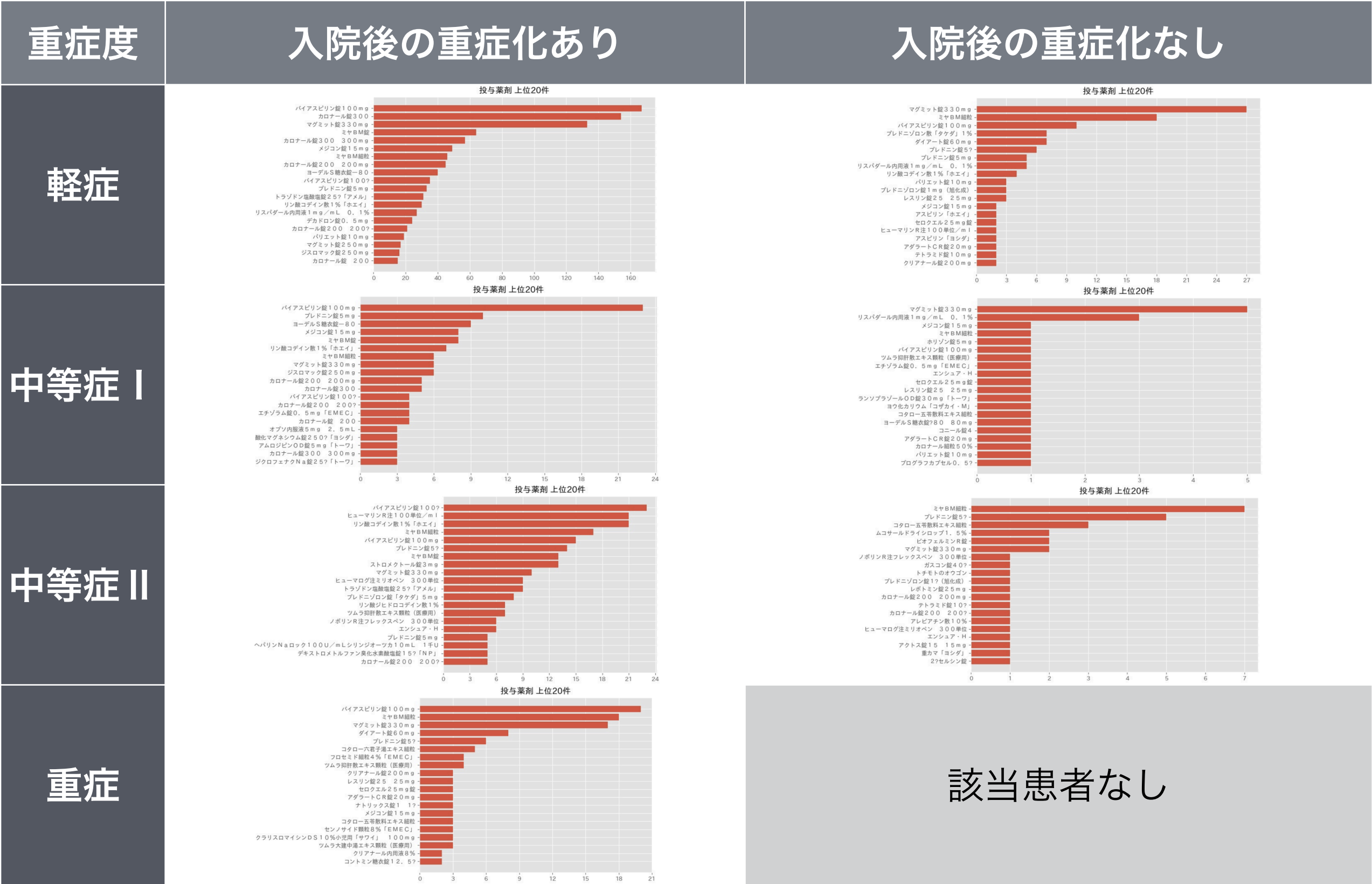
システム概要（重症度の求め方, 重症化の定義）

重症度の求め方	軽症：以下の項目に該当なし 中等症Ⅰ：呼吸困難、肺炎所見 中等症Ⅱ：酸素投与が必要 重症：ICUに入室、人工呼吸器が必要
重症化患者の定義	入院時点では非重症であったが、入院後に重症に転じた患者

重症度分類データは以下の研究成果を利用

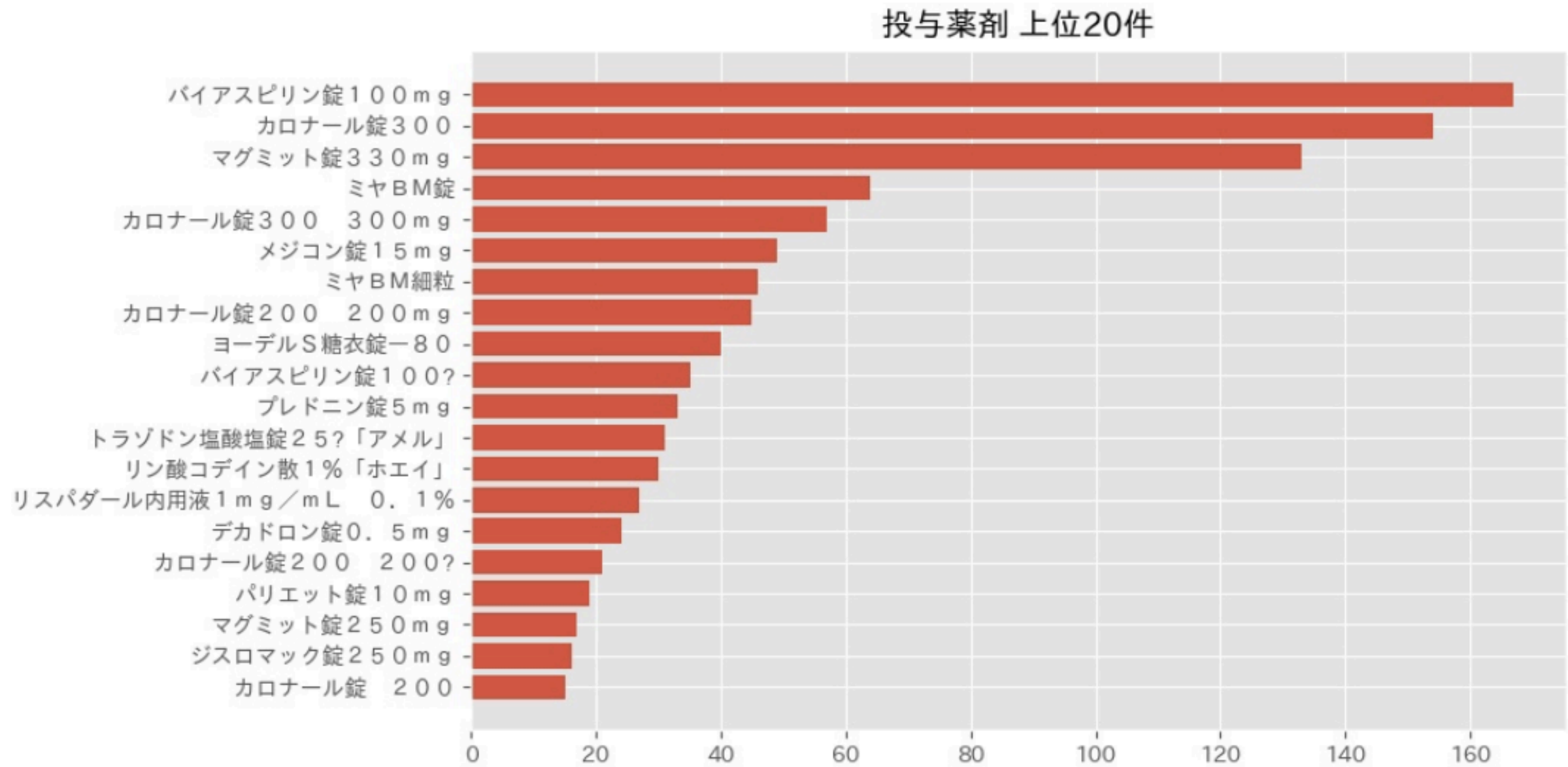
Furuhata H & Araki K. Validation of a specialized evaluation system for COVID-19 in Japan: A retrospective, multicenter cohort study. Journal of Infection and Chemotherapy 2023;29:294–301. <https://doi.org/10.1016/j.jiac.2022.12.004>.

基礎集計結果（重症度別の投与薬剤の分布）



基礎集計結果（重症度別の投与薬剤の分布）

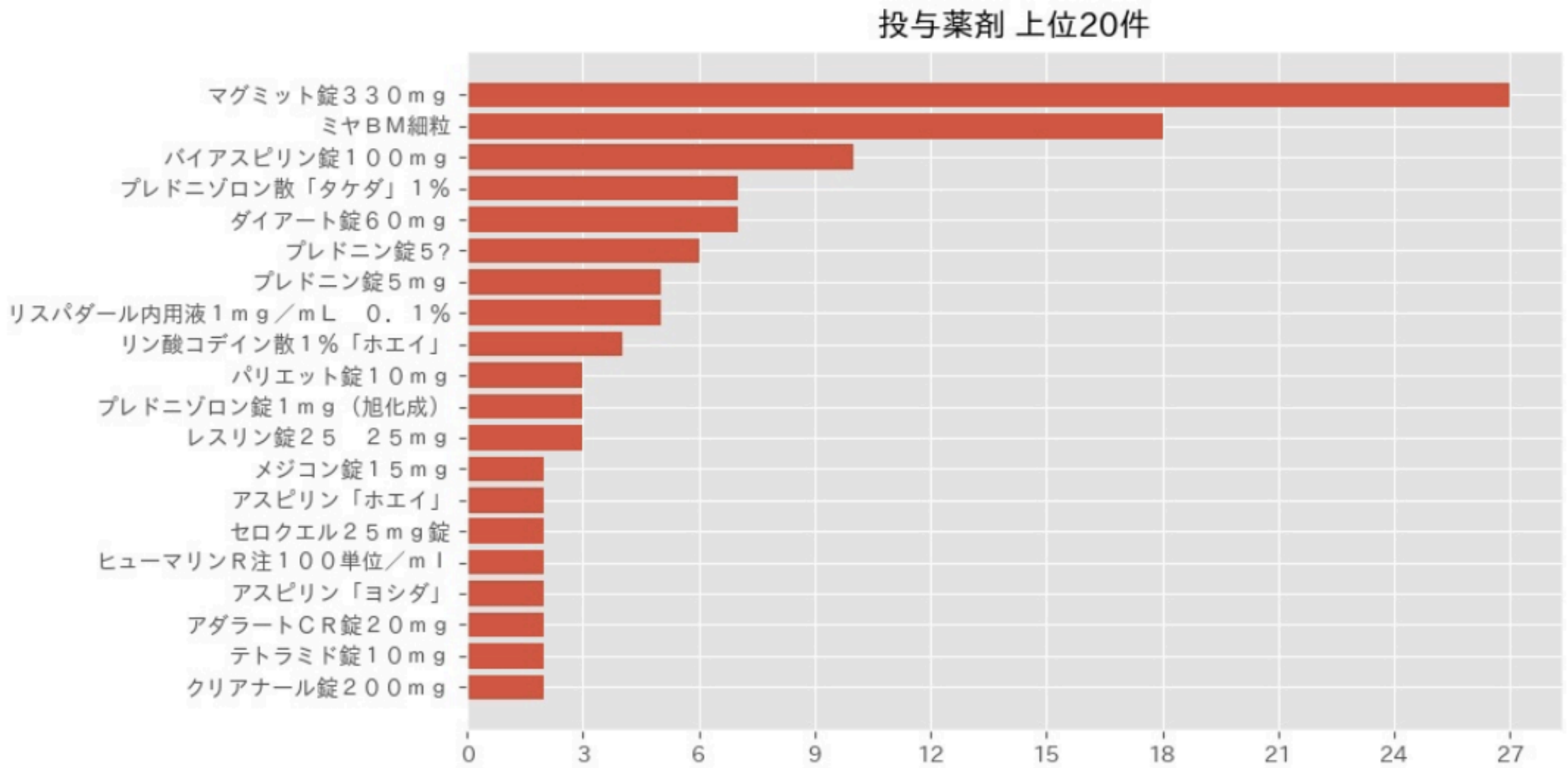
軽症（入院後の重症化あり）



バイアスピリン、カロナール、ミヤBM、マグミット

基礎集計結果（重症度別の投与薬剤の分布）

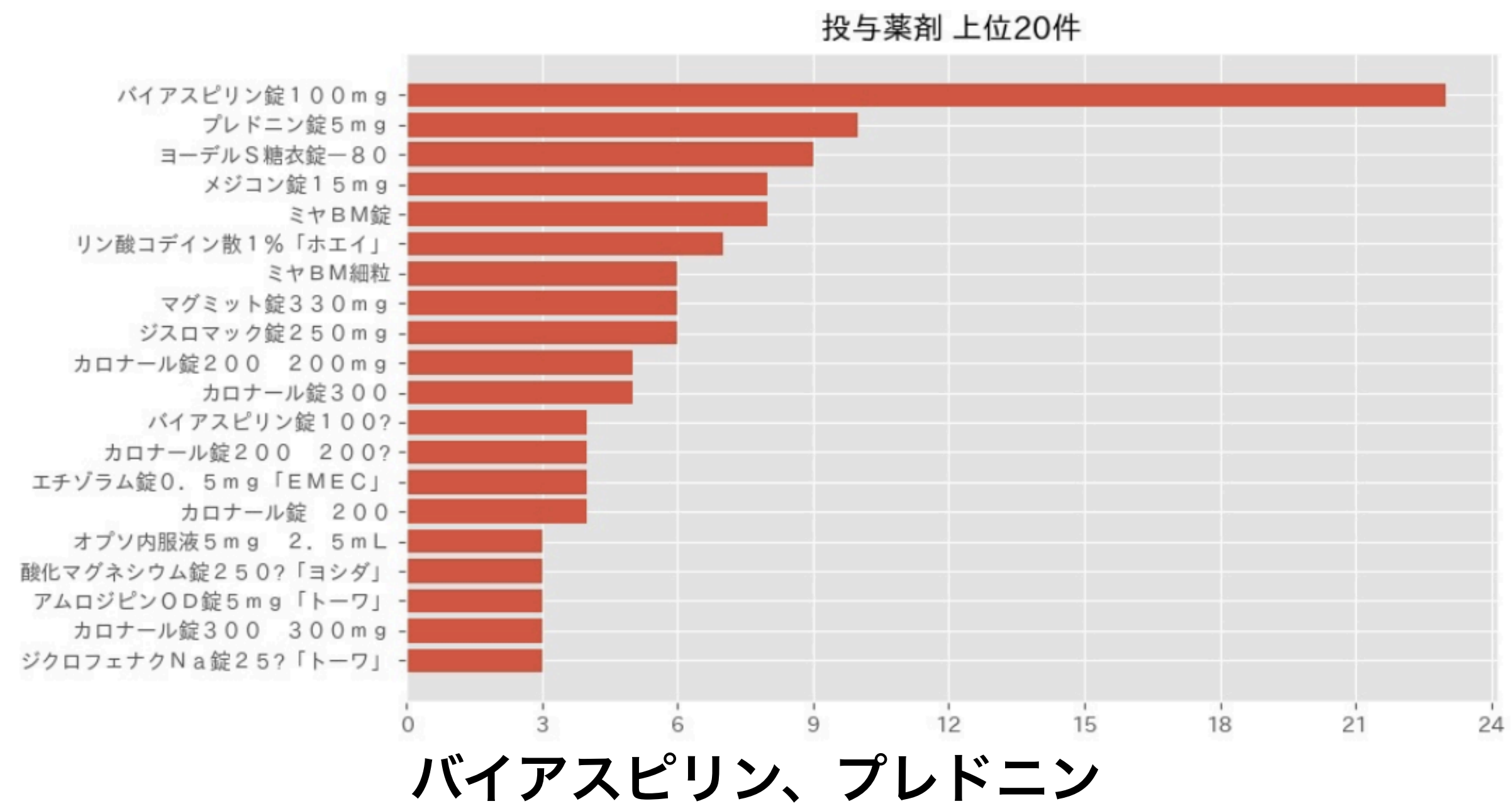
軽症（入院後の重症化なし）



バイアスピリン、ミヤBM、マグミット

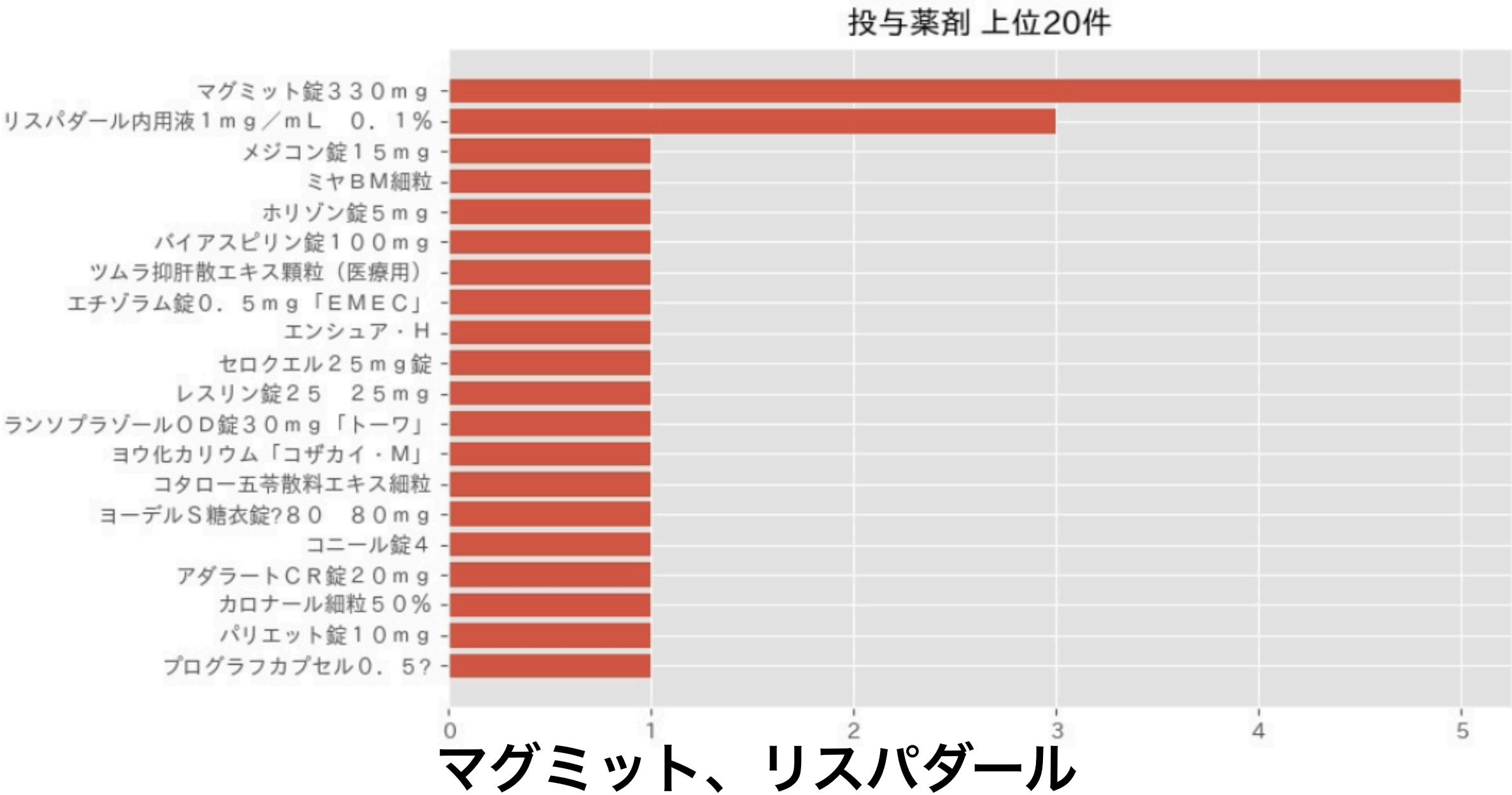
基礎集計結果（重症度別の投与薬剤の分布）

中等症Ⅰ（入院後の重症化あり）



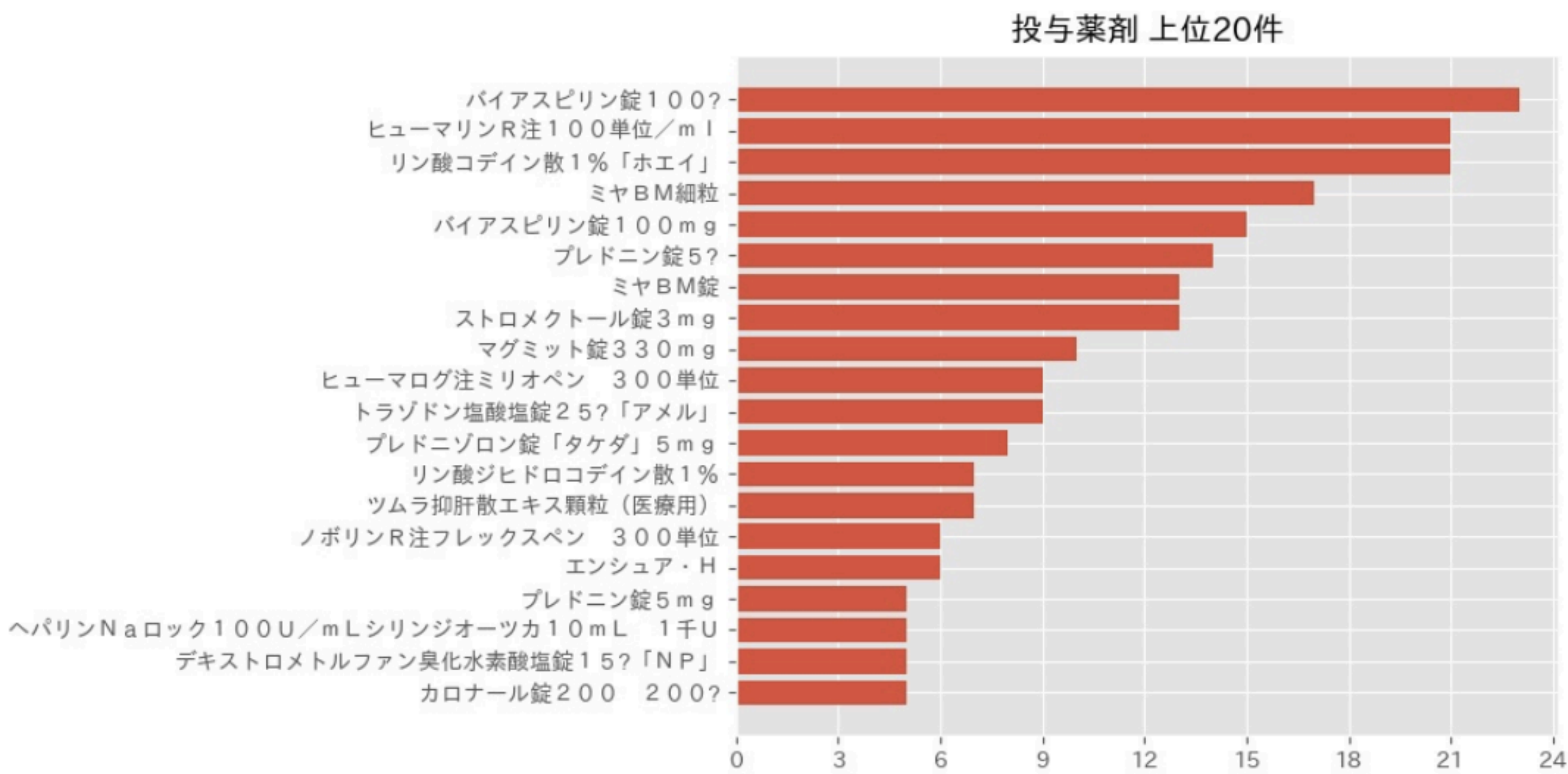
基礎集計結果（重症度別の投与薬剤の分布）

中等症Ⅰ（入院後の重症化なし）



基礎集計結果（重症度別の投与薬剤の分布）

中等症Ⅱ（入院後の重症化あり）

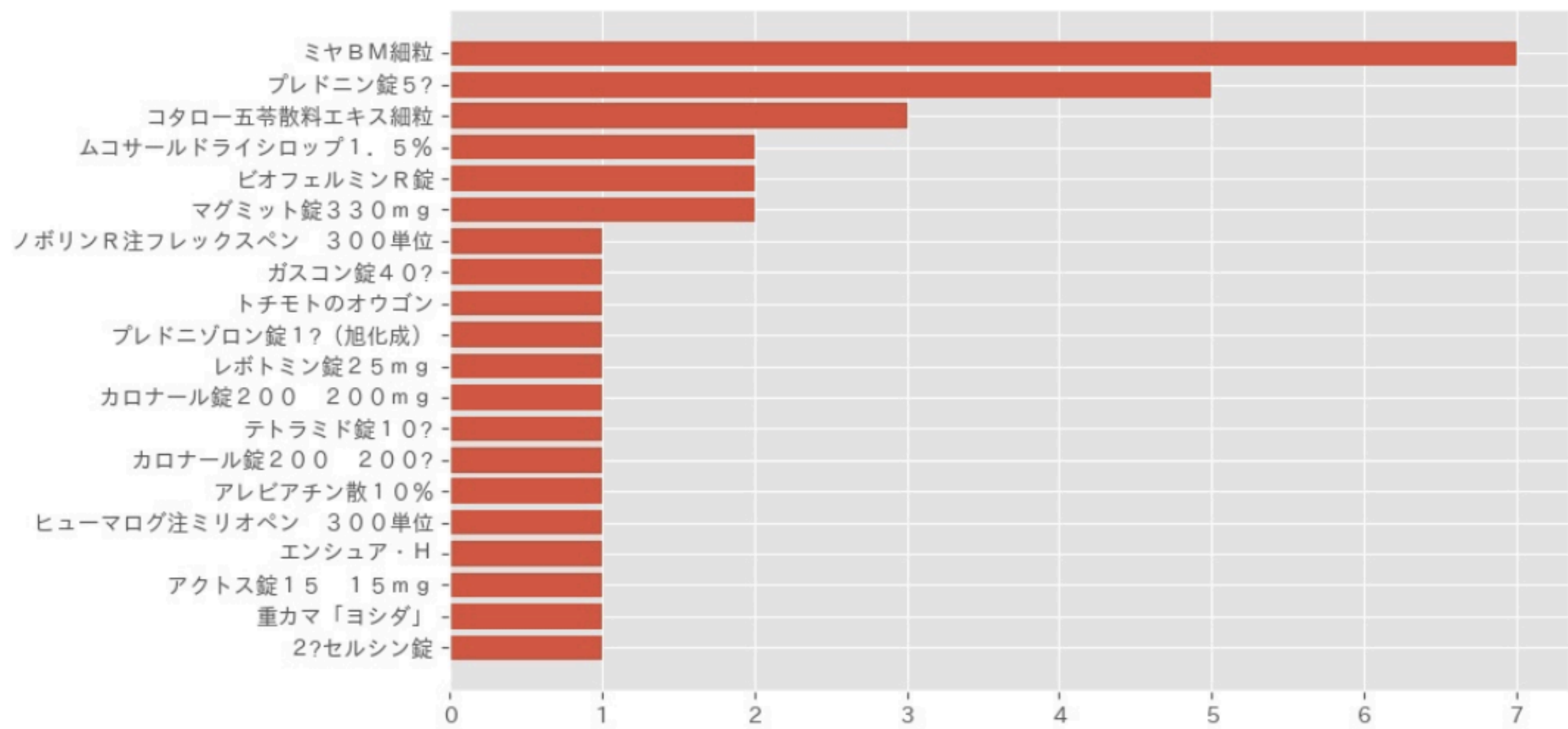


バイアスピリン、ヒューマリン、リン酸コデイン

基礎集計結果（重症度別の投与薬剤の分布）

中等症Ⅱ（入院後の重症化なし）

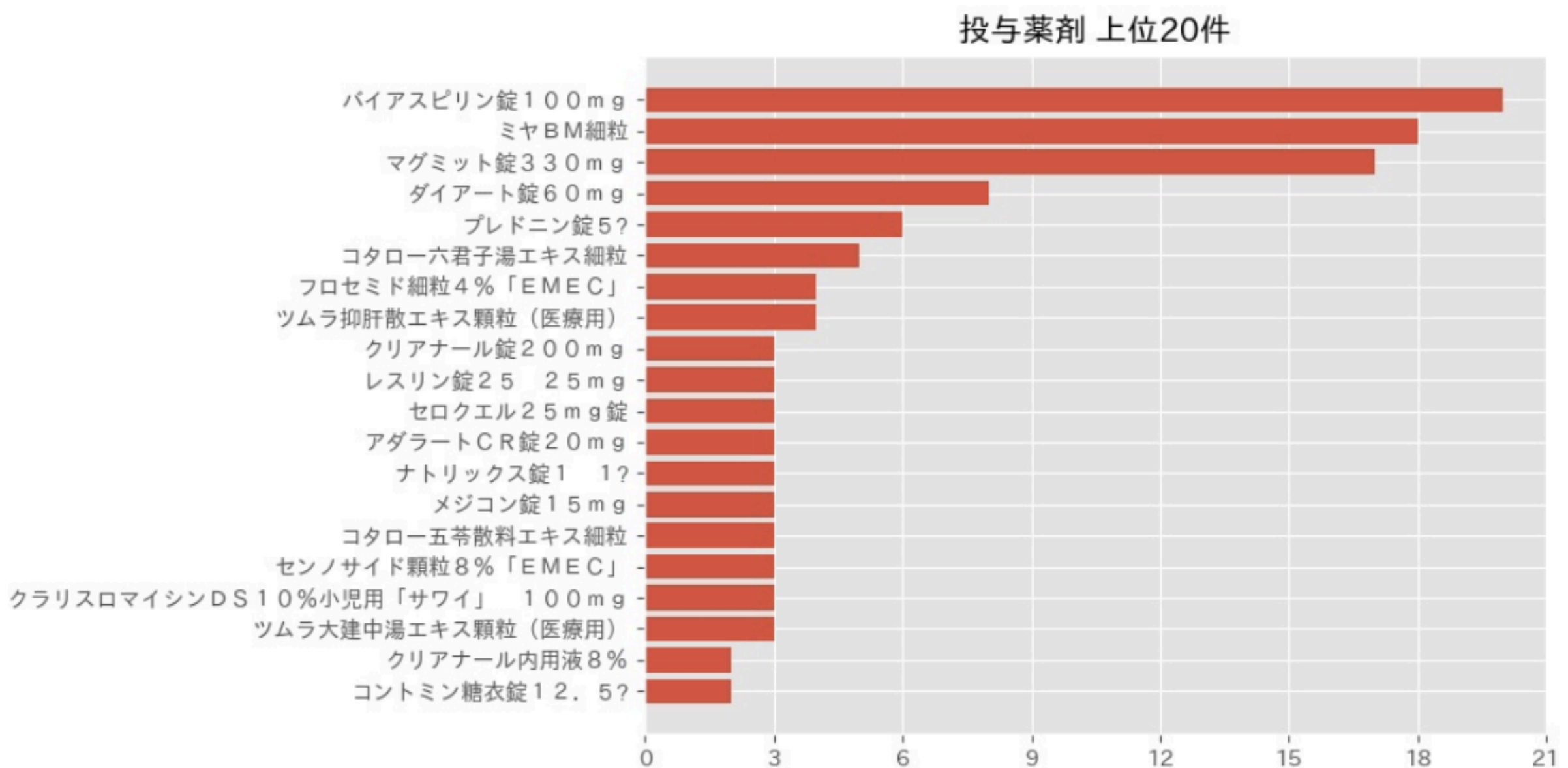
投与薬剤 上位20件



ミヤBM、プレドニン

基礎集計結果（重症度別の投与薬剤の分布）

重症（入院後の重症化あり）



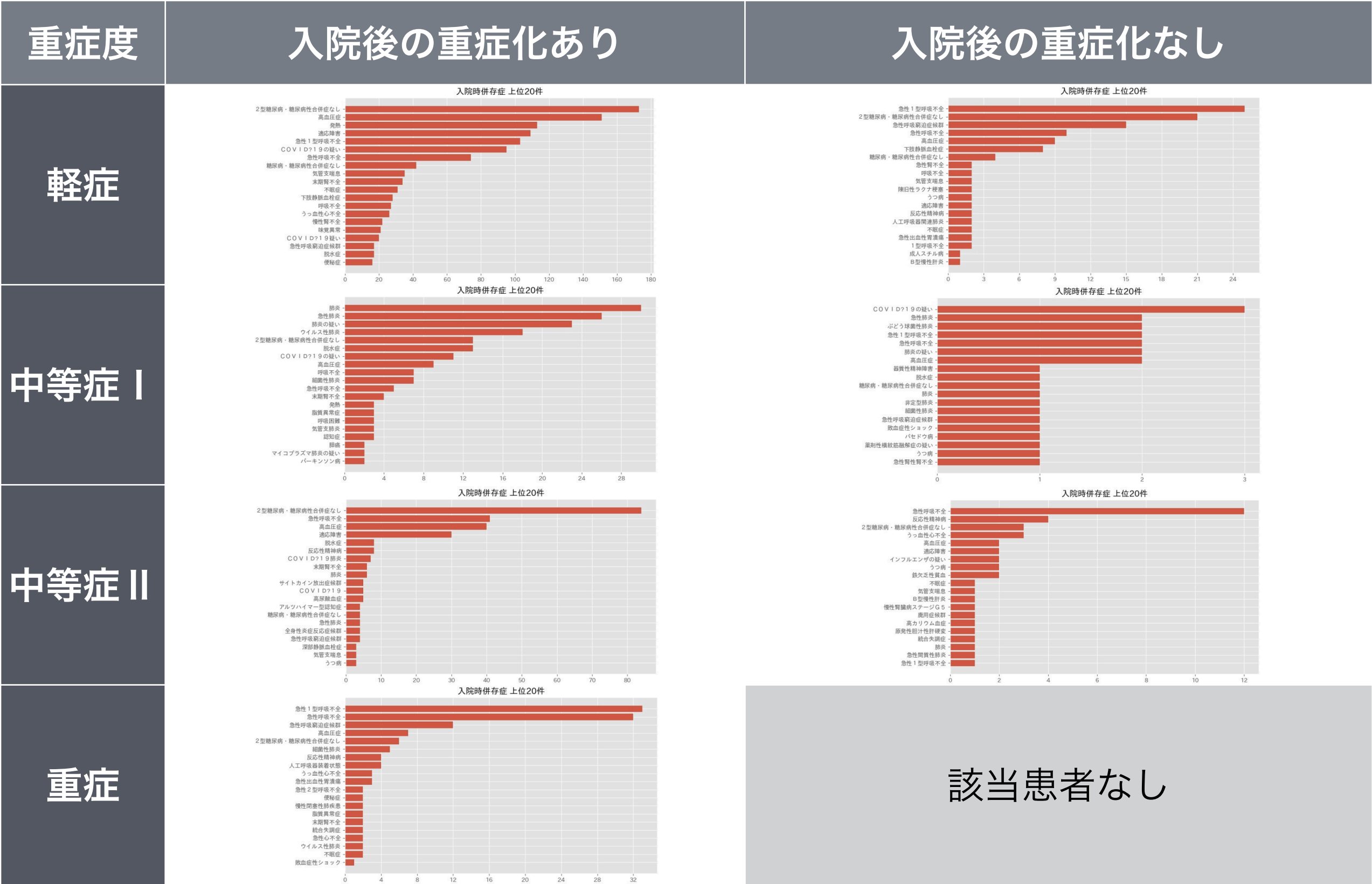
バイアスピリン、ミヤBM、マグミット、ダイアート

基礎集計結果（重症度別の投与薬剤の分布）

重症（入院後の重症化なし）

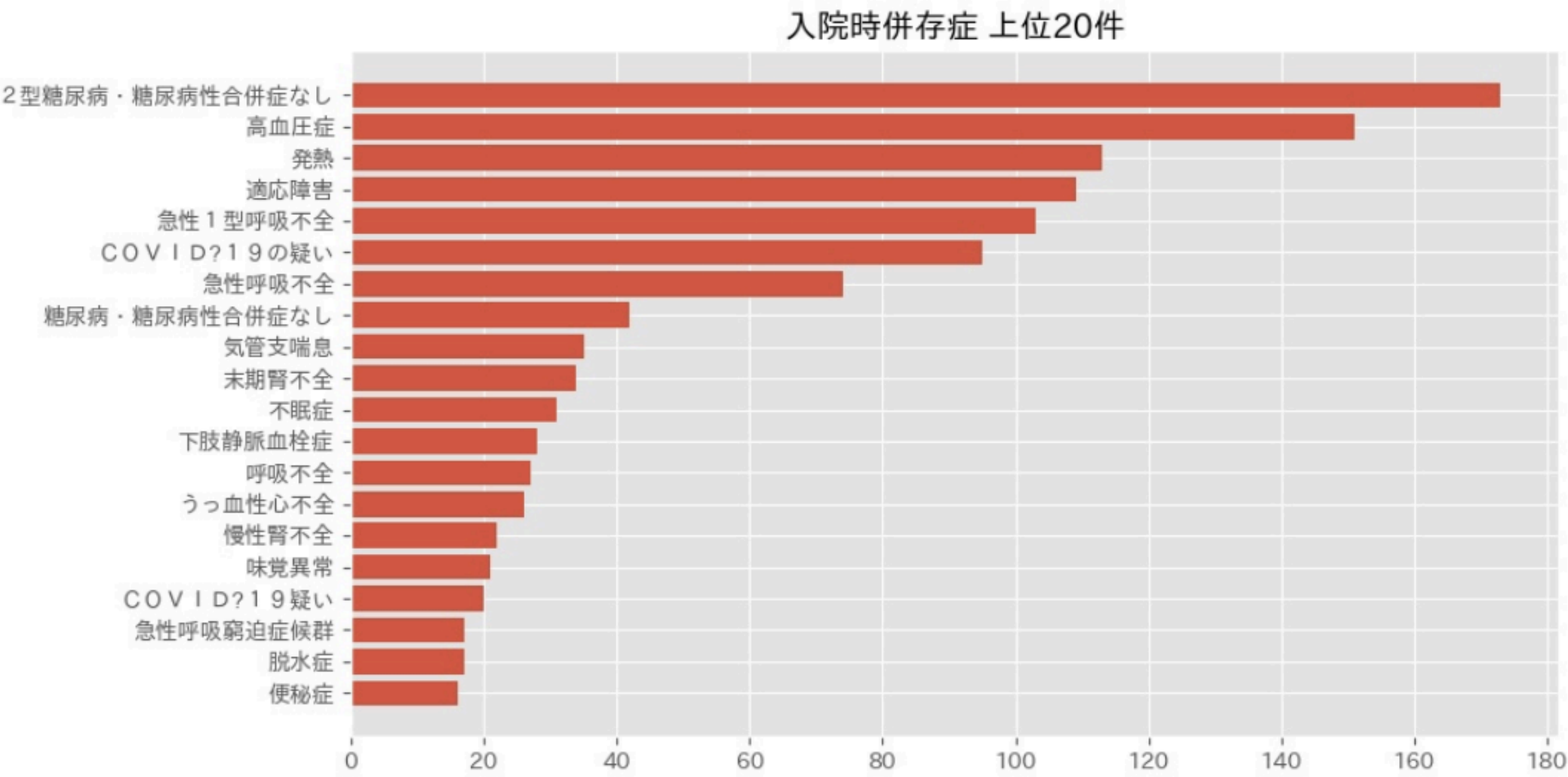
該当患者なし

基礎集計結果（重症度別の併存症の分布）



基礎集計結果（重症度別の併存症の分布）

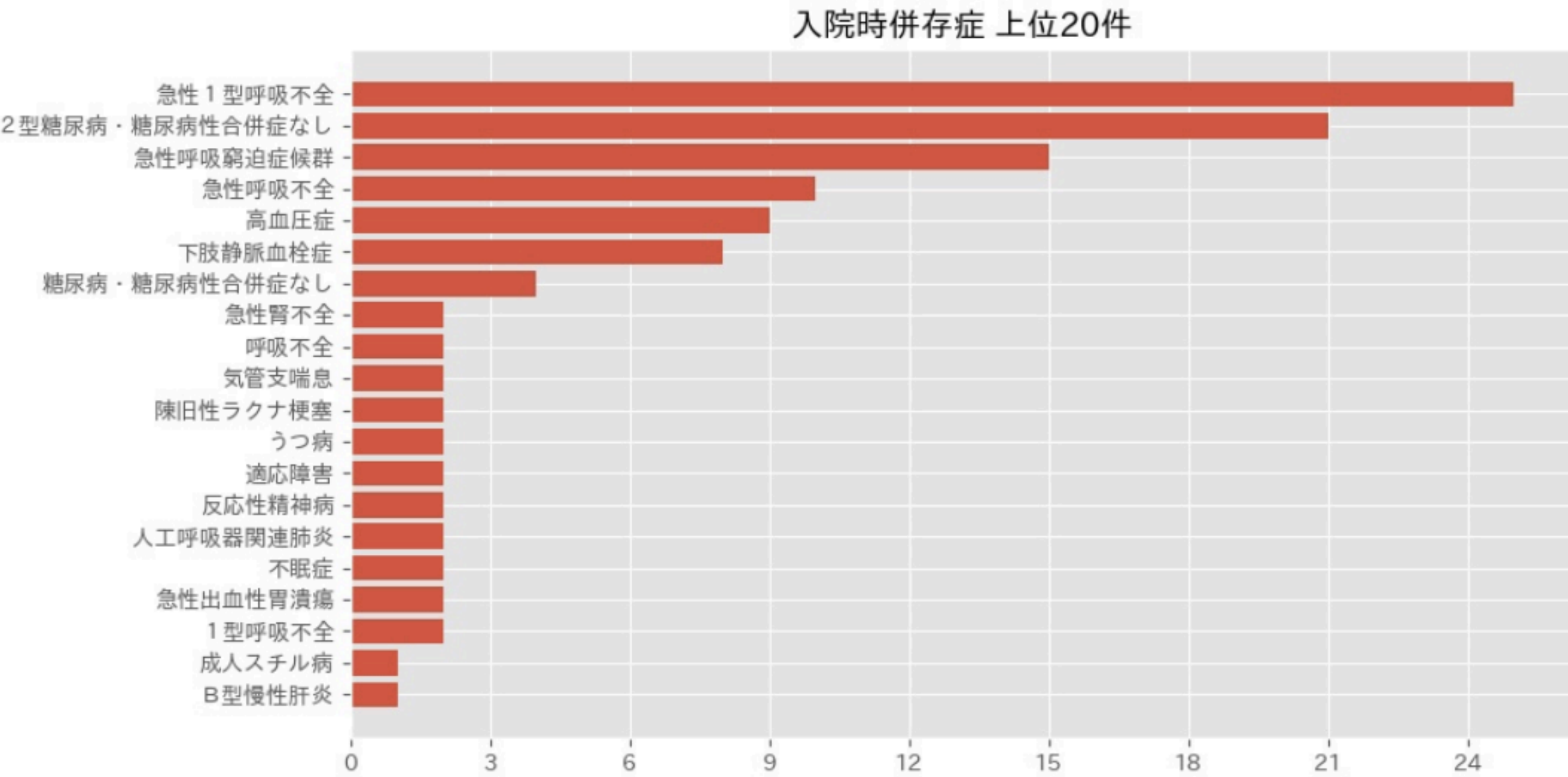
軽症（入院後の重症化あり）



2型糖尿病、高血圧、発熱、呼吸不全

基礎集計結果（重症度別の併存症の分布）

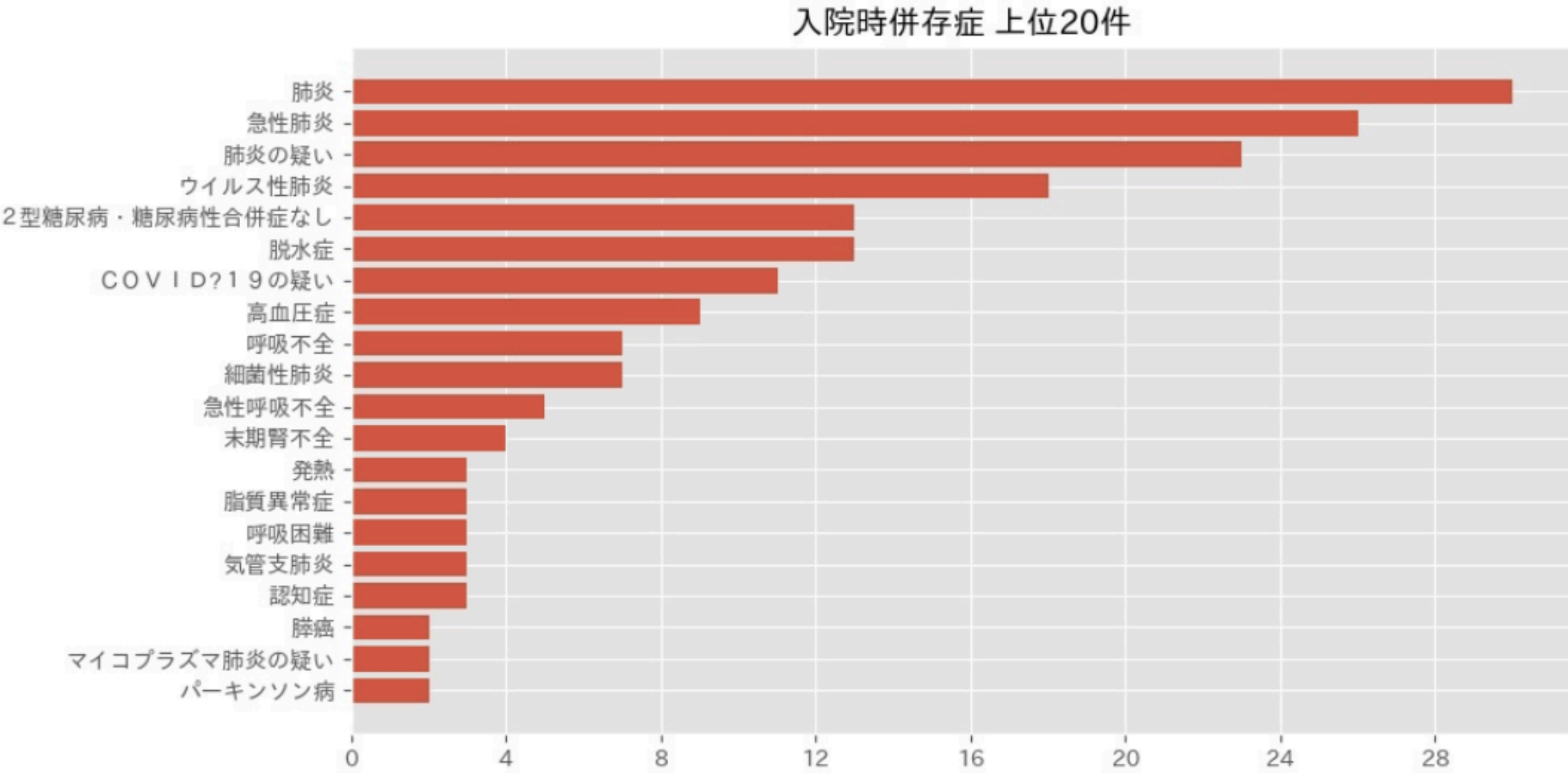
軽症（入院後の重症化なし）



2型糖尿病、高血圧、呼吸不全

基礎集計結果（重症度別の併存症の分布）

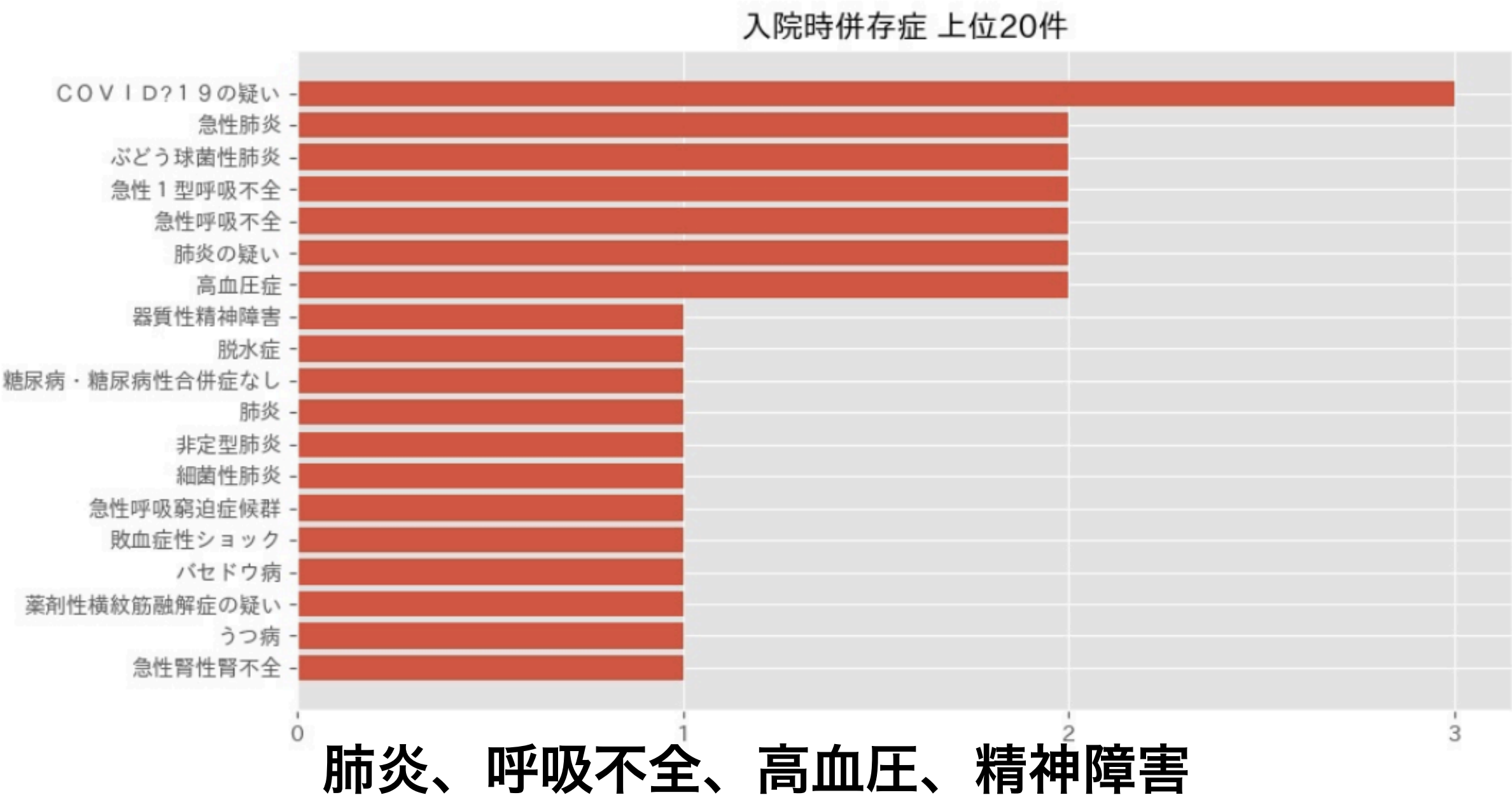
中等症Ⅰ（入院後の重症化あり）



肺炎、2型糖尿病

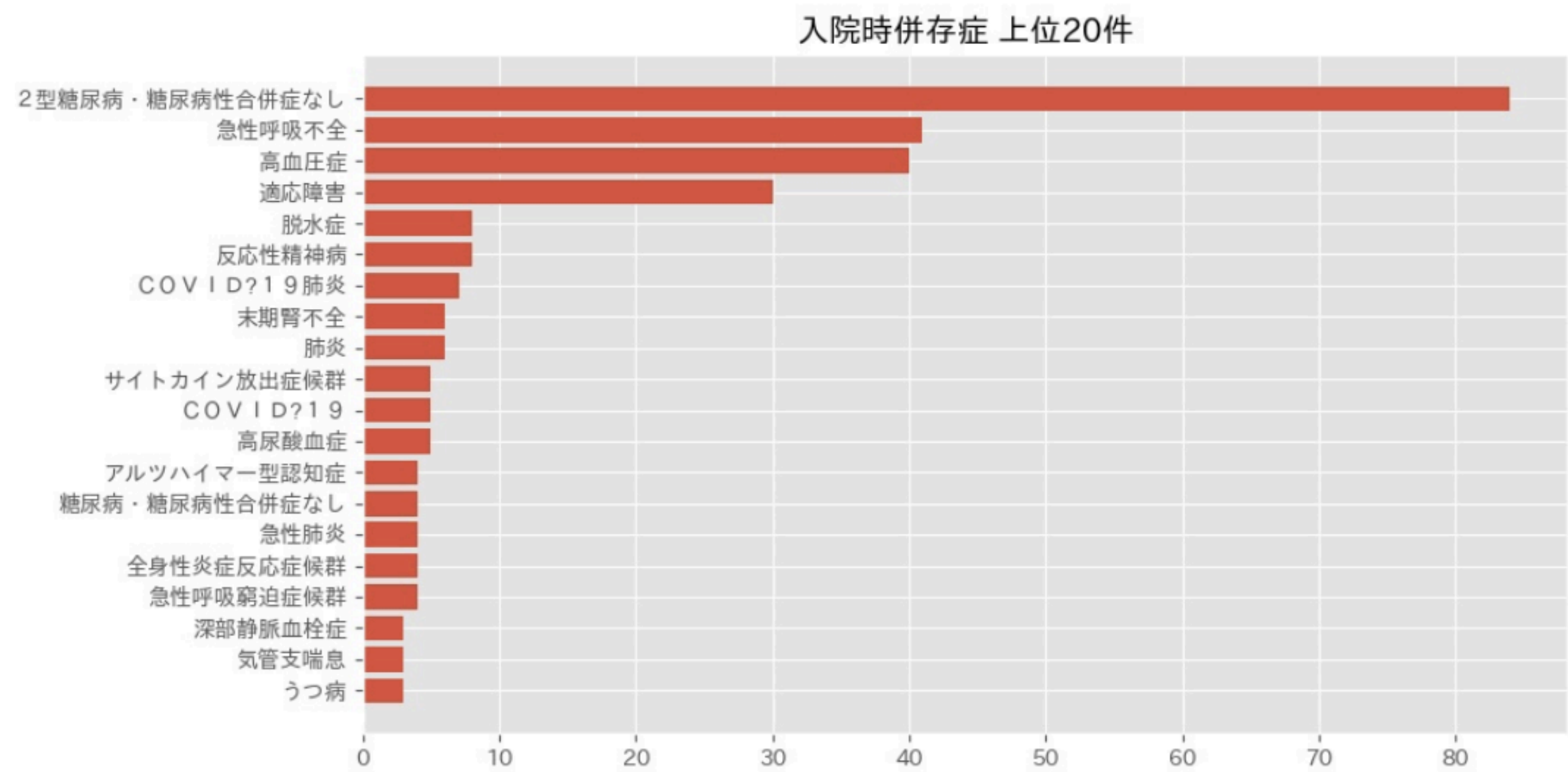
基礎集計結果（重症度別の併存症の分布）

中等症Ⅰ（入院後の重症化なし）



基礎集計結果（重症度別の併存症の分布）

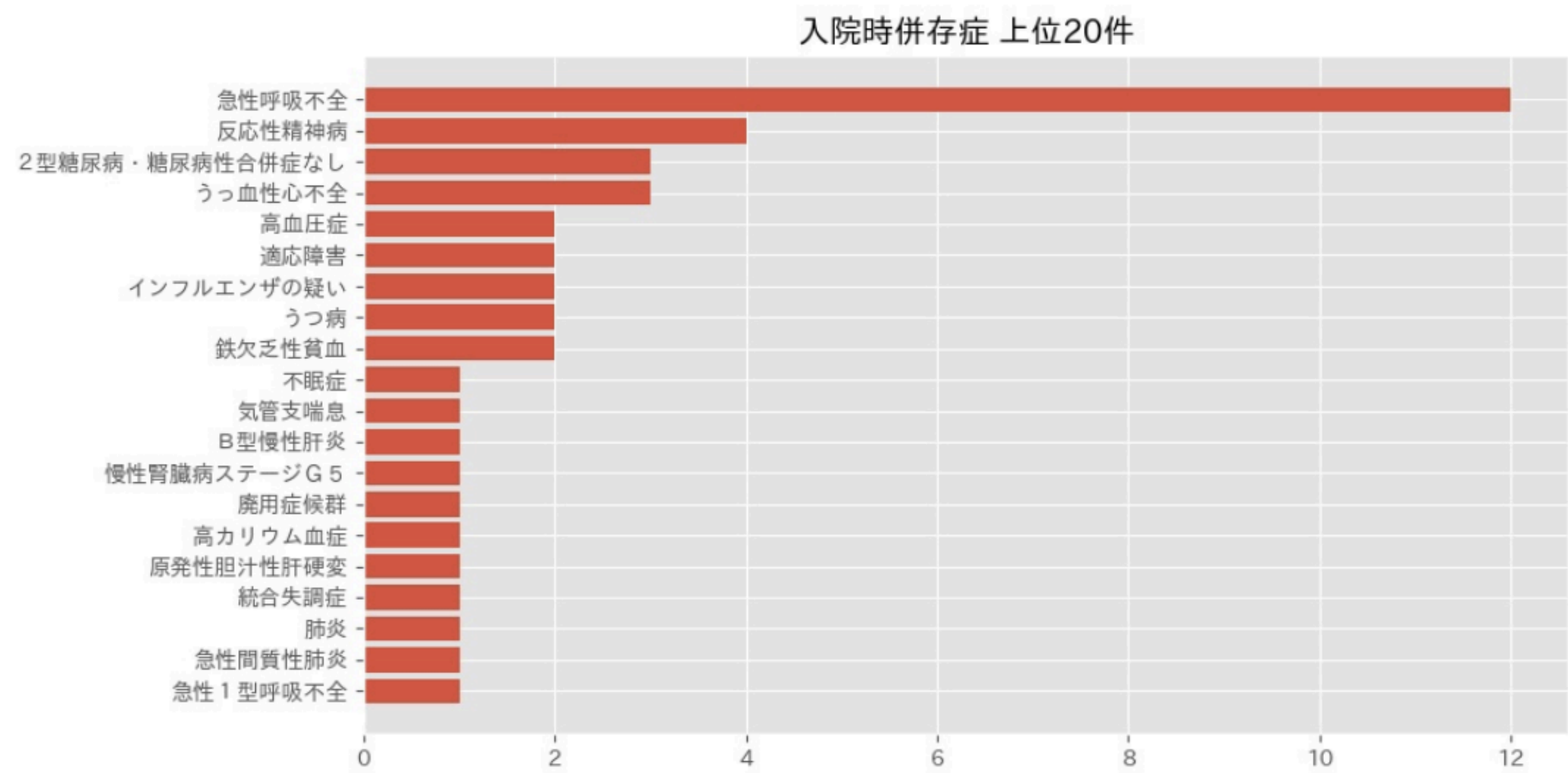
中等症Ⅱ（入院後の重症化あり）



2型糖尿病、呼吸不全、高血圧、精神疾患

基礎集計結果（重症度別の併存症の分布）

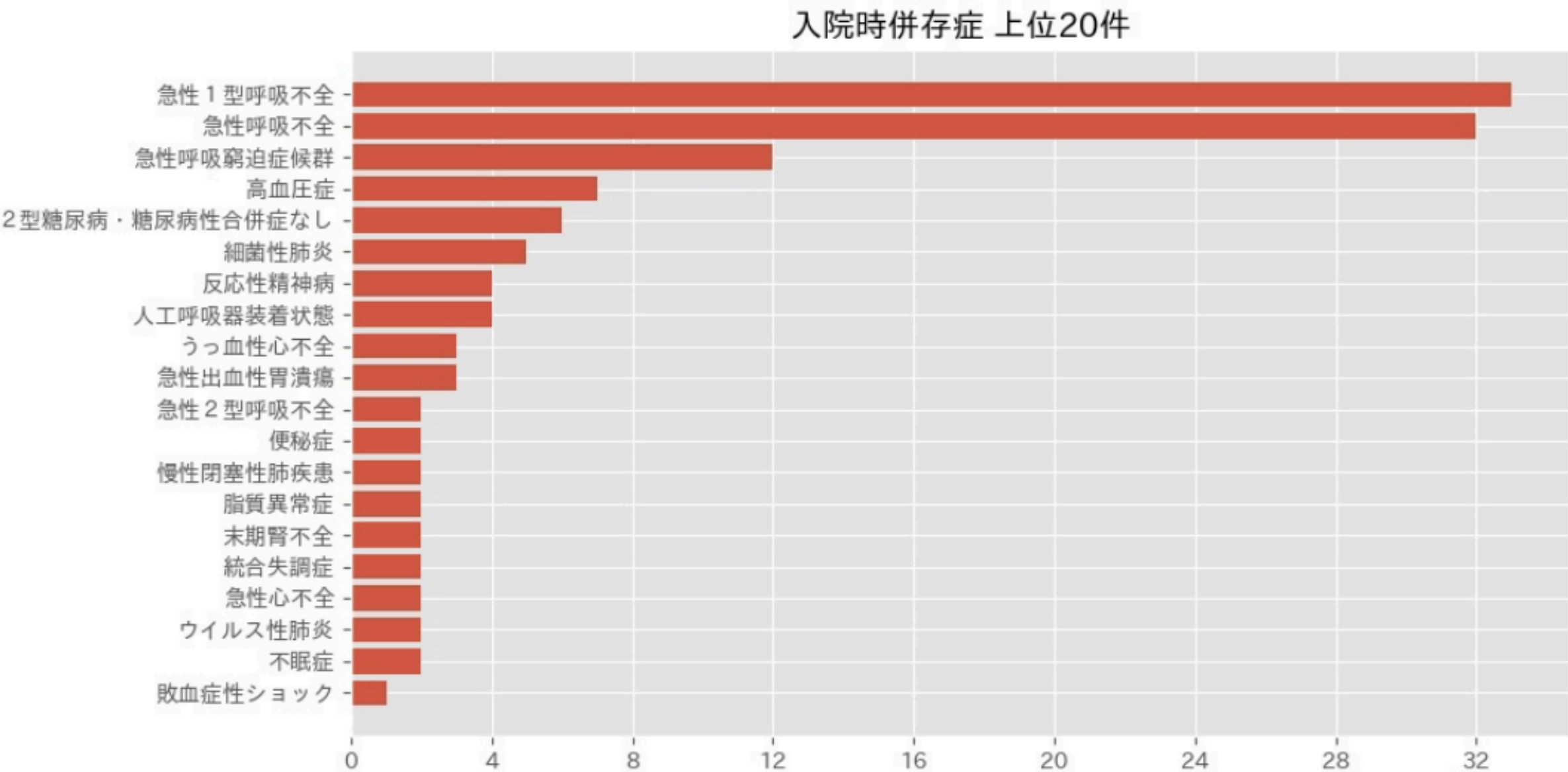
中等症Ⅱ（入院後の重症化なし）



呼吸不全、精神疾患、2型糖尿病、心不全

基礎集計結果（重症度別の併存症の分布）

重症（入院後の重症化あり）



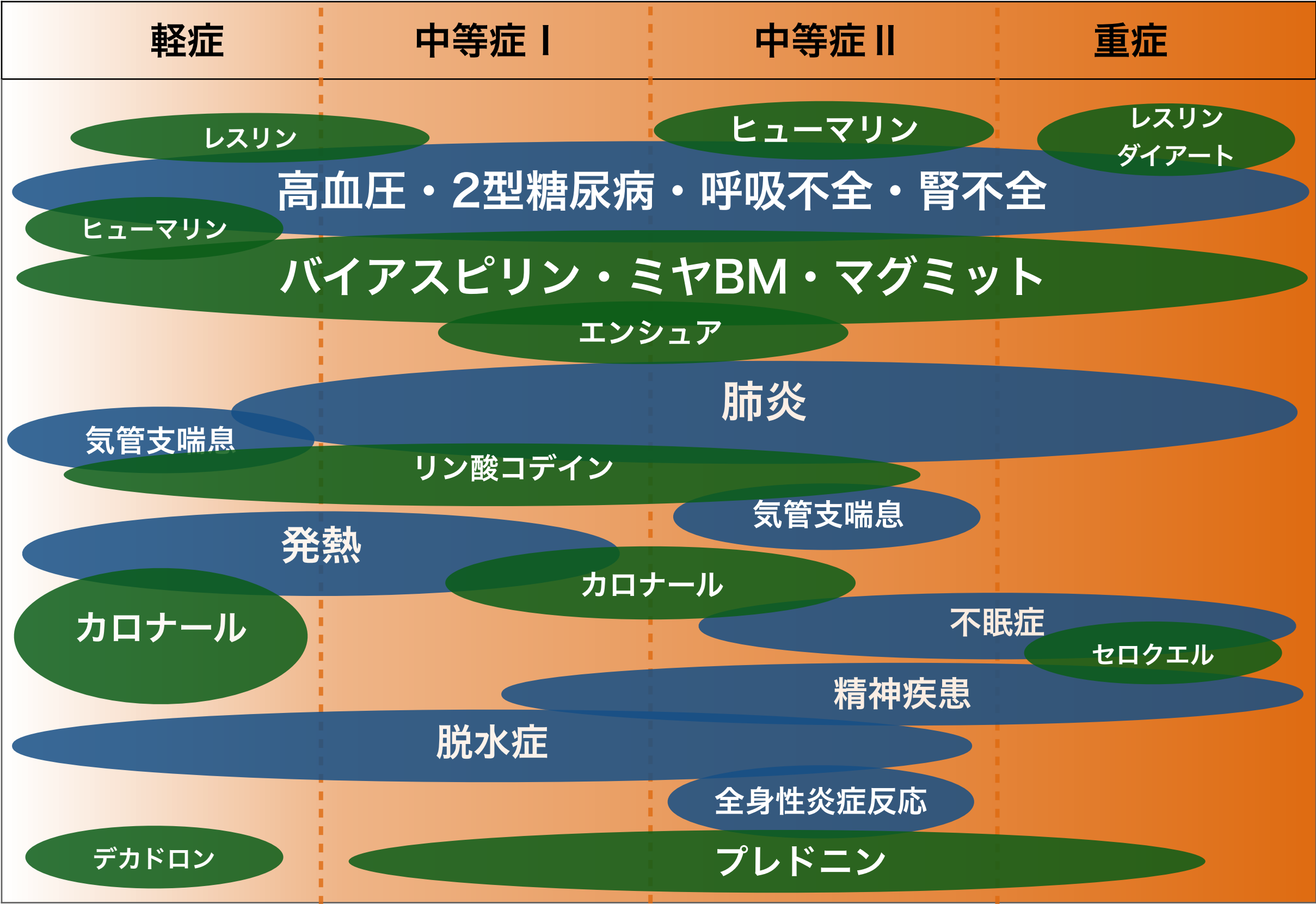
呼吸不全、高血圧、2型糖尿病、肺炎、精神疾患

基礎集計結果（重症度別の併存症の分布）

重症（入院後の重症化なし）

該当患者なし

結果（重症度別の投与薬剤・併存症の分布）



考察（重症度別の投与薬剤・併存症の分布）



集計結果の医学的妥当性

- ・ 集計結果における併存症と投与薬剤が対応している
- ・ 出現頻度の高い併存症は、いずれも報告されているリスク因子（基礎疾患・腎不全等）と一致している



重症化リスク因子

高血圧・糖尿病等の基礎疾患、及び肺炎



重症化患者における精神疾患について

- ・ 未知の疾病へ罹患したことに対する精神的負荷の増加
- ・ 統合失調・うつ等による重症化リスクの増加が報告されている

（Lamiece Hassan et al, Disparities in COVID-19 infection, hospitalisation and death in people with schizophrenia, bipolar disorder, and major depressive disorder: a cohort study of the UK Biobank., **Molecular psychiatry** 2021 Dec 07 )

COVID-19患者の死亡予測モデルの開発

患者予後（死亡/生存）の予測



トリアージや医師の負担軽減



システム概要（対象データ）

医療機関	共同研究機関26施設
入院期間	2019年4月1日～2021年9月30日
患者数	COVID-19入院患者 延べ8,658人（死亡患者数592人）
データ種類	入院時併存症・年齢・性別・BMI・喫煙指数 等
使用特徴量	59個

システム概要（前処理）

共通

- 次の5項目をキーとしてDPCデータ（FF1, EF）を統合；
- ①施設ID ②患者ID ③入院年月日 ④診療年月日 ⑤データ取込み年月日
- ※ 複数インデックスによる検索効率化、また患者データの重複の防止
- 前処理①：薬剤データ区分：内服, 頓服, 皮下筋肉内, 静脈内に限定
- 前処理②：レセプト電算コード：薬剤に該当するコードに限定
- 前処理③：欠損値削除
- 前処理④：カテゴリデータのフラグ化

前処理による データ数増減

8,676件→ 6,742件（重複している患者レコード18件、性別・年齢・喫煙指数・BMIに欠損値を含むレコード1,934件を削除）

システム概要（予測モデル）

複数の機械学習モデルを利用し比較

決定木	Decision Tree Classifier (DT)
決定木 (アンサンブル学習)	Random Forests (RF)

システム評価（予測モデルの患者群ごとの精度評価）

患者群Ⅰ：65歳以上、BMI30以上

患者群Ⅱ：65歳以上、BMI30未満

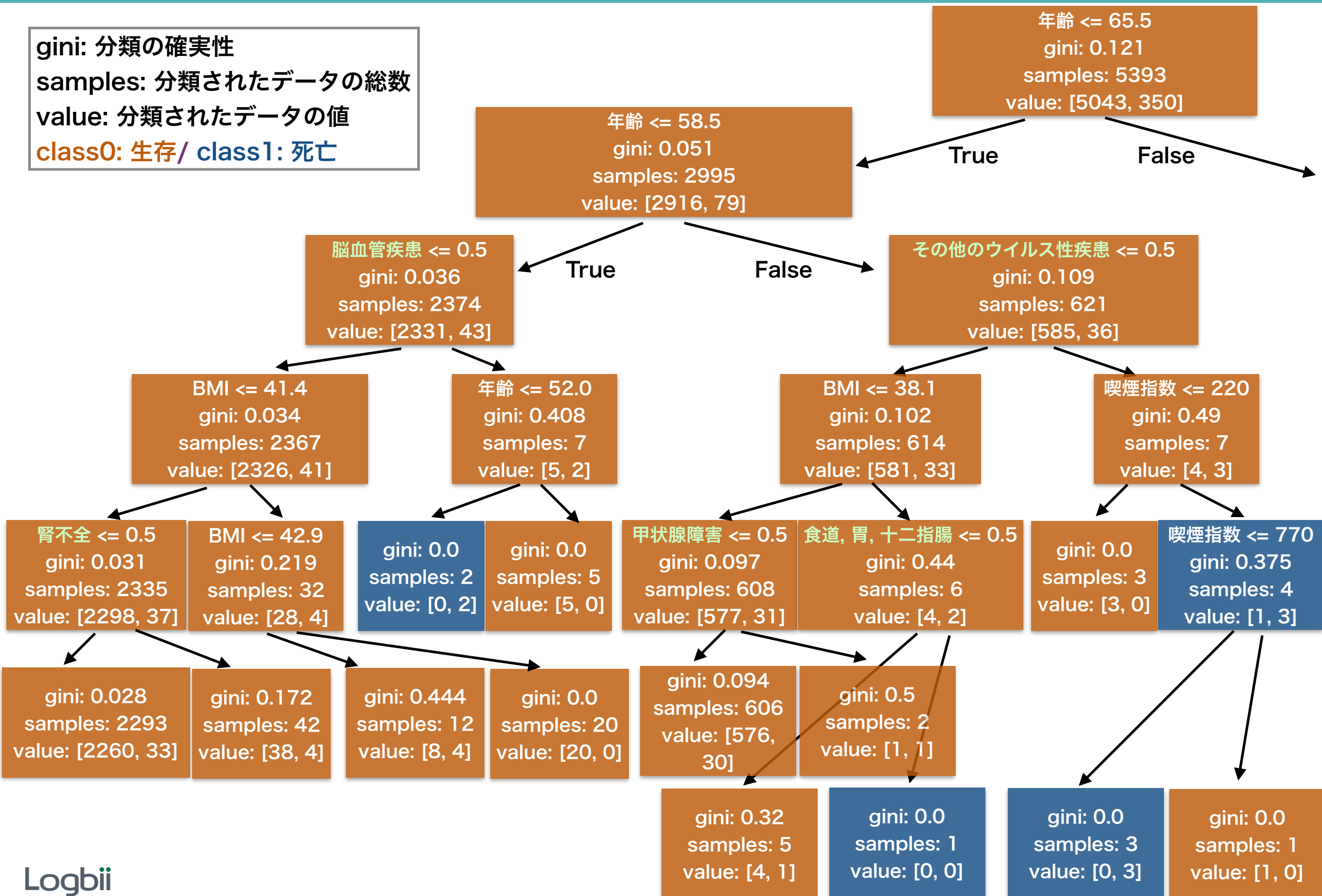
患者群Ⅲ：65歳未満、BMI30以上

患者群Ⅳ：65歳未満、BMI30未満

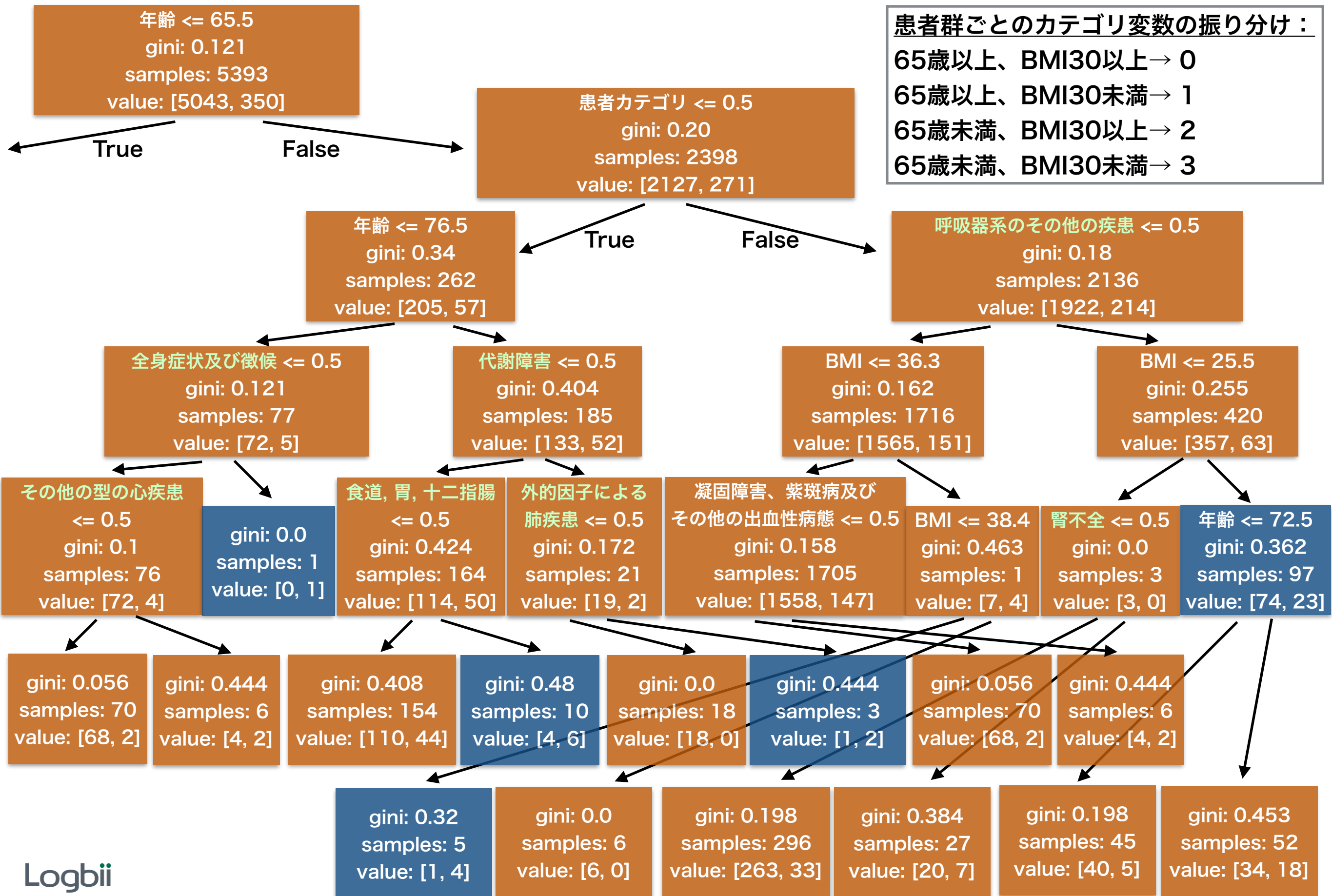
死亡		AUC	F値	適合率	再現率
DT	全患者	0.58	0.23	0.26	0.20
	患者群Ⅰ	0.56	0.22	0.50	0.14
	患者群Ⅱ	0.56	0.20	0.67	0.12
	患者群Ⅲ	0.58	0.20	0.20	0.20
	患者群Ⅳ	0.60	0.25	0.33	0.20
RF	全患者	0.53	0.2	0.60	0.068
	患者群Ⅰ	0.56	0.22	0.50	0.14
	患者群Ⅱ	0.53	0.11	0.60	0.060
	患者群Ⅲ	0.50	0.0	0.0	0.0
	患者群Ⅳ	0.55	0.14	0.25	0.10

システム評価（患者の死亡に起因した併存症調査）

gini: 分類の確実性
samples: 分類されたデータの総数
value: 分類されたデータの値
class0: 生存 / class1: 死亡



システム評価（患者の死亡に起因した併存症調査）



考察（決定木図およびモデル評価）

集計結果の医学的妥当性

各ノードの特徴量は、いずれも報告されているリスク因子（基礎疾患・腎不全・高齢等）と一致している

決定木のノードに基礎疾患が含まれていない点について

高血圧・糖尿病等の基礎疾患は、相関の高い他の特徴量（年齢・BMI等）によって代表されている可能性がある

予測モデルの精度評価値が低い点について

説明変数が限られている（併存症・年齢・性別・BMI・喫煙指数）ため、今後、検歴データの利用により更なる精度向上が見込まれる

COVID-19患者の重症度別の薬剤・入院時併存症の集計

重症度別の併存症と投与薬剤を明らかにし、
概ね臨床的な解釈をあてはめることができた

COVID-19患者の死亡予測モデルの開発

利用できる説明変数が限られていることにより、
現状十分な予測精度ではないが、
検歴データ等の利用により今後の改善が期待される

今後の計画

- ☑ 投与薬剤・併存症の集計について、患者の年齢や性別を考慮したクロス集計を行う
- ☑ 決定木に現れなかった特徴量を含めた特徴量同士の相関などを明らかにし、真に患者の死亡に寄与した併存症の特定を目指す
- ☑ 目的変数に患者の重症度を利用し、マルチラベル分類モデルの構築を行う
- ☑ 説明変数に検歴データを追加し、予測モデルの精度向上を目指す

次世代医療基盤法に基づく匿名加工医療 情報を利用した入院後敗血症の予測モデ ルの開発における課題と対策の検討

株式会社ログビー	<u>松田敦義</u> 、池田遼太郎
宮崎大学医学部附属病院	荒木賢二
一般社団法人ライフデータ イニシアティブ	松尾亮輔

次世代医療基盤法とは



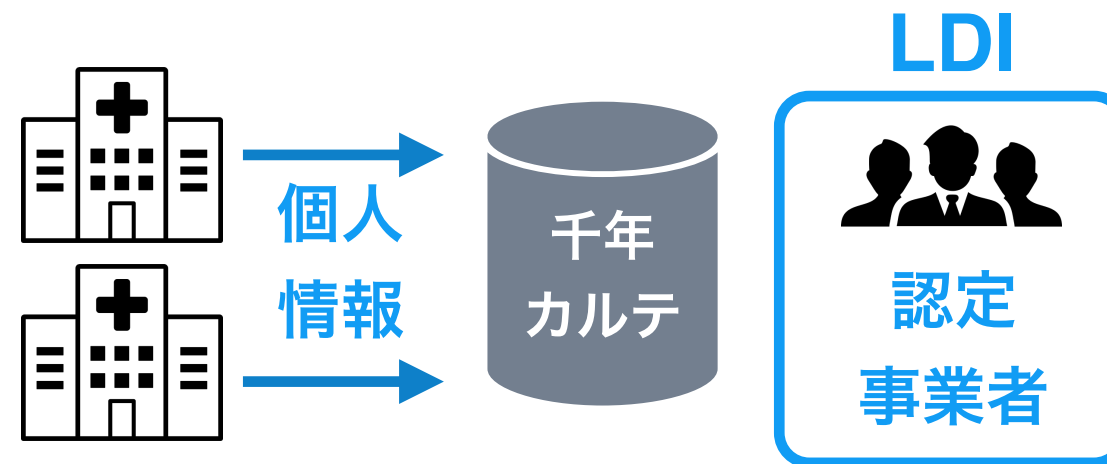
効果

- ・新薬開発
- ・臨床研究
- ・個別化医療

匿名加工医療情報

- ・個人が識別できない
- ・個人情報を復元できない

千年カルテとLDI



千年カルテ

- ・ 多数の医療機関のデータベース

ライフデータイニシアティブ (LDI)

- ・ 認定事業者
- ・ 千年カルテのデータ活用

課題

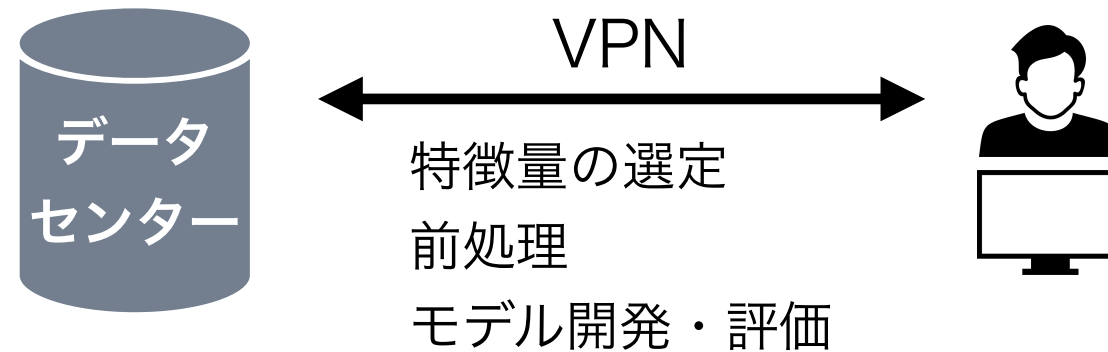
- ・ 匿名加工医療情報の前例ない
- ・ AI活用の知見が少ない
- ・ 安全性の知見が少ない

目的

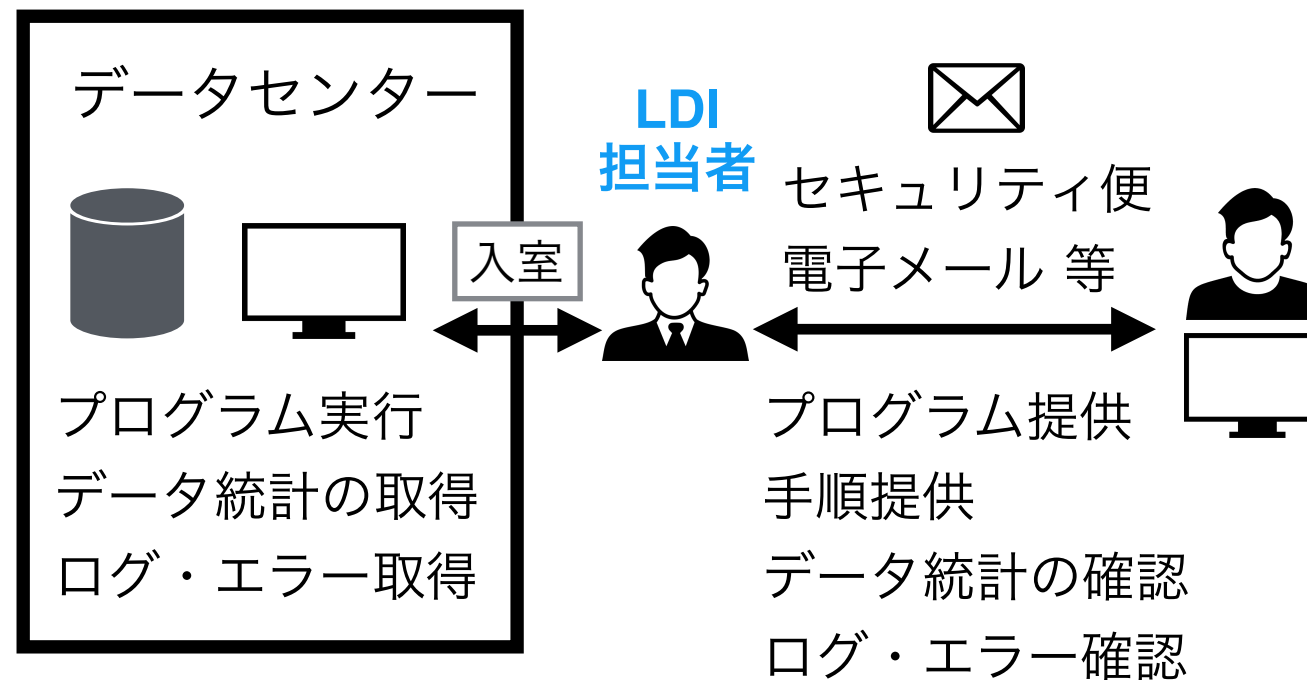
- ・ 匿名加工医療情報を用いた機械学習モデルの開発
- ・ セキュリティ要件の担保
- ・ プロセスの課題と対策の検討

3.1 開発プロセス

①宮崎大学での試作

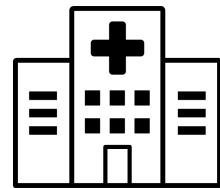


②千年カルテでの開発



3.2 予測モデル開発

データ 2015～2021年度 DPC



× 20

- ・ 患者基本情報
- ・ 入院時疾患
- ・ 前処理後 68,593件
- ・ 特徴量 166個

予測モデル

- ・ がん患者の入院後敗血症を予測
- ・ Random Forest
- ・ 特徴量重要度を抽出

前処理

- ・ 入院単位のデータ
- ・ フラグ化、欠損削除

4.1 結果: 開発プロセス

時間がかかったプロセス

	プロセス
1	エラーの確認と対応
2	データ授受等の進め方の確認
3	テーブル確認、対象テーブルの修正
4	データベース確認、接続情報の修正
5	インストールライブラリの確認と承認
6	取得するデータ種類の確認と承認

- ・ LDIとの電子メールのやり取り
は100回近く

4.2 結果: 精度評価

2020年度と本研究の比較

研究	AUC	F値	適合率	再現率
2020年度	0.727	0.424	0.374	0.490
本研究	0.769	0.528	0.536	0.523

2020年度: 宮崎大学のみ

- ・ データ数 13,527件（前処理後）
- ・ 特徴量 80個（バイタル、検体検査、実施オーダーも利用）

敗血症の発症別（本研究）

敗血症	F値	適合率	再現率
発症なし	0.997	0.996	0.998
発症あり	0.058	0.075	0.048

4.3 結果: 特徴量重要度

本研究の特徴量重要度

	特徴量	重要度
1	体重	0.150
2	身長	0.130
3	年齢	0.077
4	併存症_真菌症	0.041
5	併存症_原虫疾患	0.036
6	併存症_悪性新生物<腫瘍>	0.028
7	併存症_高血圧性疾患	0.024
8	併存症_食道/胃/十二指腸の疾患	0.023
9	併存症_凝固障害/紫斑病及びその他の出血性病態	0.023
10	性別	0.022

- ・ バイタル、検体検査、実施オーダーも利用すると、重要な特徴量が変化するのではないか

開発プロセスの課題

- ・ 個票データの確認ができない
(ノイズや表記ゆれなど)
- ・ 結果はLDIを通して確認が必要
- ・ EDAや特徴量選定に時間がかかる

課題の対策 (LDIの施策案)

- ・ 検証環境を用意し、予め指針をたてられるようにする
- ・ データ集計や機械学習モデルのAPI等テンプレートを提供し、簡単に利用できる形にする

匿名加工医療情報の活用

- ・ 多施設データ利用で精度向上に期待
- ・ 安全性とAI開発は少なからずトレードオフ
- ・ 研究プロセス改善の検討が望まれる

今後

- ・ 引き続き多施設の医療情報の活用
- ・ COVID19の予測モデル開発

千年カルテの多施設のデータを活用するメリット

- ・ 多施設のデータを使うことで、データの母数を増やし、データの偏りを減らして集計、分析ができる
- ・ データの母数を増やすことで、機械学習モデルの精度向上が期待できる

匿名加工医療情報を活用した開発プロセスの課題

- ・ 直接プログラムの実行結果やログが確認できないため、開発に時間がかかるので、何かしらの開発プロセスを効率化する施策が望まれる