山梨大学
UNIVERSITY OF YAMANASHI

予防医療分野における疫学データへの機械学習技術活用について ～スパースモデリングを活用した糖尿病発症予測と予測因子探索～

大岡 忠生[1)] 日野 英逸[2)] 横道 洋司[1)] 山縣 然太朗[1)]

[1)] 山梨大学大学院総合研究部社会医学講座

[2)] 統計数理研究所

自己紹介

大岡忠生

医師・産業医

専門：社会医学・公衆衛生

2015年4月～　山梨大学医学部附属病院　医師

2017年4月～　山梨大学大学院総合研究部医学域
社会医学講座　助教

その他：複数企業嘱託産業医

市区町村の健康診断業務など

社会医学(Social Medicine)とは何か

医学の三本柱

	人体の構造や仕組み	病気の診断や治療	社会における医療の在り方
	基礎医学	臨床医学	社会医学
	解剖学 生理学 生化学 病理学 薬理学 微生物学 …	内科 外科 整形外科 産婦人科 小児科 眼科 耳鼻科 …	公衆衛生学 産業医学 法医学 生命倫理学 病院管理法 保健所 地方自治体 …
医師の割合	1%	98%	1%

公衆衛生（Public Health）とは何か

WHO（世界保健機関）の定義：

「組織された地域社会の努力を通して、疾病を予防し、生命を延長し、身体的、精神的機能の増進をはかる科学であり技術である」

公衆衛生学の例

予防医学
例）健康診断による疾病発見、生活習慣病対策、母子保健など

産業医学
例）働く人のメンタルヘルス、安全衛生管理、など

伝染病（感染症）
例）インフルエンザ感染防止、BSE、SARS、コロナウィルス等の拡大防止、など

昨今の予防医療の重要性の向上

予防医療が解決しうる日本社会が抱える問題点

・医療費の高騰、健康寿命延伸の必要性

生活習慣病によるものが医療費の1/4、がんによるものを合わせると1/3程度になる
60歳以上が医療費の約7割を使用しており、そのうち半分はがんと循環器疾患によるもの。

・介護負担の増大/メンタルヘルスによる離職率の上昇

親の介護による成人人口の負担が増大、離職や家族問題につながっている。また、生産人口においてメンタル疾患が増加傾向にあり、多くの企業において生産性低下が発生。

・医師不足/医師の過重労働

国民への病気や健康への正しい理解の普及、未病段階での適切な介入が実質的な患者数や医療コストを減らす。病院に来る前の教育と費用対効果を見据えた適切な健康介入が必要。

現在自身で取り組んでいる研究

① 妊婦の喫煙行動の要因と子供への影響についての検討（公衆衛生）

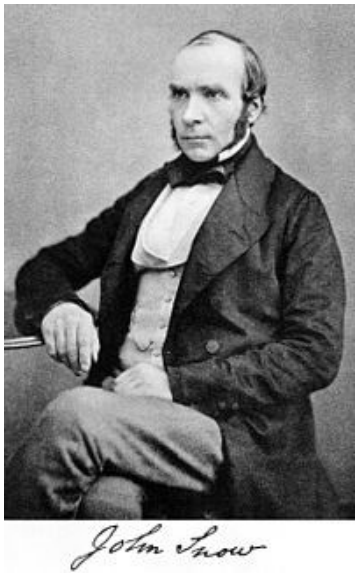
② 睡眠不足や睡眠障害が企業の従業員の生産性に与える影響（産業医学）

③ 疫学データへの機械学習技術の応用による予防医療の推進

疫学とは何か

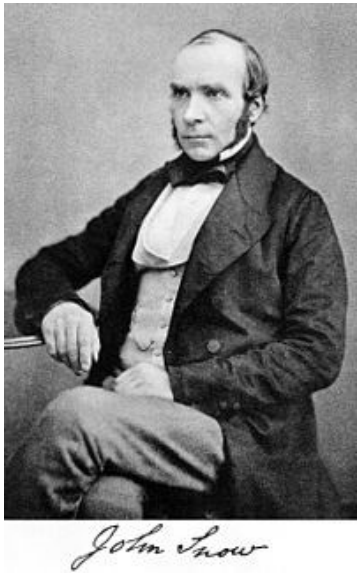
ロンドン1854年

・当時は悪臭説
・最初の3日間で死者126名
・9月末までに500名死亡



疫学の祖
ジョン・スノウ

メカニズムが分からなくても
病気は予防できる：疫学の考え方



1854年
スノウにより井戸は閉鎖され、
コレラの感染拡大は免れた。
原因は特定できなかった。

ロベルト・コッホ（1843-1910）

ドイツの微生物学者。
炭疽菌、結核菌、コレラ菌の発見者。
寒天培地やシャーレを考案するなど、
細菌培養法の基礎を確立した。

1884年
ロベルト・コッホにより
コレラ菌が発見された。

疫学研究に用いる疫学データとは？

主に公衆衛生分野（コホート研究や予防医学領域）で集められた疫学研究用データの事を指す

【疫学データの例】

① 大規模コホートデータ：数十、数百万人規模で数年～数十年にわたって集められた質問や検査

② 健康診断データ：毎年行っている血液検査、基本的な問診、画像検査等

③ 健康保険組合のデータ：病院への受診記録や請求書のデータ

④ 産業保健データ：会社の勤怠データ、ストレスチェックデータ

⑤ フィールド研究データ：住民に対して継続的に介入や調査を行う事で得たデータ

疫学データは、継時的に測定され、欠損を含むデータである事が多い
元々研究（予測）目的でとられていない、既存データを利用する事も多い

機械学習は疫学に対してフィットするか

疫学分野への機械学習手法適応例はまだ少ない

疾病の予測が中心テーマであるにも関わらず、疫学データに対して機械学習が活用されているケースは現状の研究であまり存在しない。機械学習が使われる例が少ない理由としては以下が考えられる。

機械学習が使われていない大きな3つの原因＋α

①機械学習技術が疫学研究のどのような場面で有効であるか不明である

②機械学習技術の疫学データへの適用にそもそもメリットがあるか不明である

③研究者が新たに機械学習技術を習得・理解する難しさ、既存手法の伝統化

＋α　日本の医療者は機械学習や統計の専門家と接する機会が殆どなく、
機械学習の活用自体がそもそも選択肢に入ってこない。（特に中小規模の研究室）

統数研の共同利用に申請した理由

今回の研究に取り組むに際して、

① 機械学習の中身を理解し、

② 機械学習の解析を正しく行い、

③ 結果に対する正しい解釈が出来る

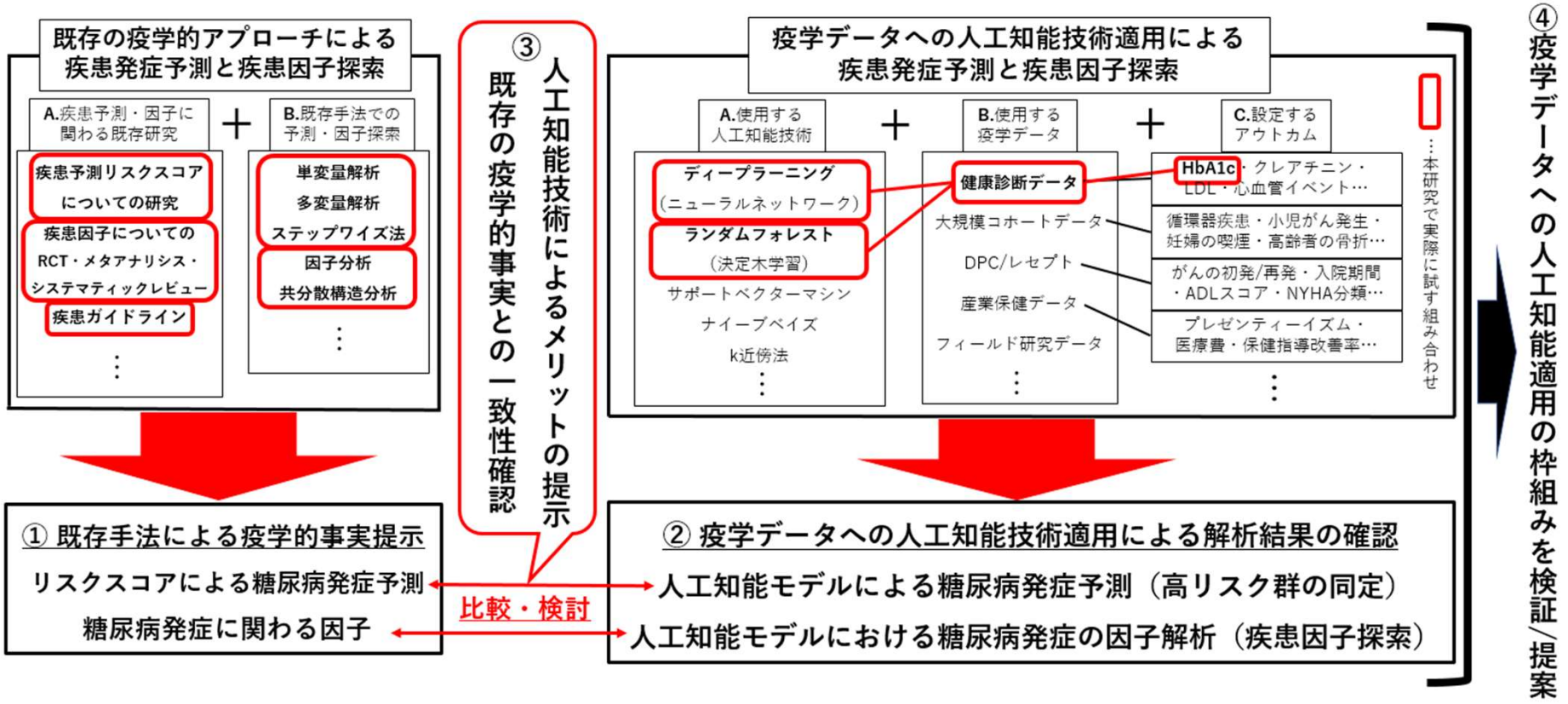
ようになるため。

2 共同利用の趣旨

研究所の共同利用は、大学等に所属する研究者が、これまで研究所が蓄積してきたさまざまな研究資源を活用して、統計に関する数理及びその応用の研究を行い、その成果をもって**学術研究の健全な発展に資することを目的**とするものです。研究所の物的資源としては、計算機設備や図書が利用できます。また、研究所が開発した一連のソフトウェア・パッケージや、その他のソフトウェアが利用できます。これらにも増して重要な資源は、研究所の有する統計科学全般に関するノウハウであり、人的資源です。研究所の公募型の共同利用はそうした研究所外の方々による研究所の様々な資源の利用を促進しその経費を助成するもので単なる助成研究とは異なります。研究所内外の研究者の交流の場を提供することを目的とし、統計科学の理論と応用における多面的な発展に寄与しています。

疫学・予防医学分野と統計数理・機械学習のコラボレート推進へ

我々が一連の研究で確認したいこと



（疫学研究における一般的な解析方法）
健康診断結果から糖尿病を予測したい場合

（基礎知識の説明）

糖尿病とは？

血糖を下げる物質であるインスリンの作用が十分でないため、血糖値が上昇して全身に様々な悪影響を及ぼす状態のこと。

糖尿病の診断に必要な血液検査項目
HbA1c（ヘモグロビンエーワンシー）

HbA1cとは簡単に言えば、２～３か月の血糖値の平均を数値化したものです。血糖値は食事や運動により大きく変化するため、糖尿病の診断には適していません。HbA1cは直前の食事や行動によって左右しないため、糖尿病の診断に有用であり、HbA1c:6.5以上の時に一般的に糖尿病と診断されます。

（疫学研究における一般的な解析方法）
健康診断結果から糖尿病を予測したい場合

①既存研究や日常の診療において関係のありそうな変数を選択
②糖尿病の発生の有無を目的変数としたロジスティック回帰分析を実施

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	性別	年齢	身長	体重	ＢＭＩ	体脂肪率	白血球数	赤血球数	ヘモグロビ	総コレステ	血小板数	総タンパク	アルブミン	血糖	ＨｂＡ１ｃ	収縮期	拡張期	尿酸	γ－ＧＴＰ
2	男	49	169.6	56.3	20	18	4.7	5.12	16.1	187	224	7.2	4.6	97	4.7	130	78	6.7	14
3	女	47	155.2	45.7	19	22	4.4	3.92	9.5	191	294	7.1	3.9	90	5.1	132	78	3.7	10
4	女	48	150.5	43.9	19	27	7.2	4.49	13.9	161	209	7.4	4.5	85	4.4	110	72	3.7	9
5	男	44	171.7	75.3	26	29	5.7	5.29	16.5	263	205	7	4.2	90	5.3	118	72	7.7	133
6	男	53	161.5	52.4	20	18	6.7	4.59	14.2	211	208	6.6	4.3	84	5.2	110	72	5.8	10
7	女	63	151	50.2	22	25	3.8	4.98	14.2	248	256	6.8	4.5	103	6.3	96	58	5.5	35
8	男	40	180	76.3	24	27	5.7	4.39	13.8	193	237	7	4.3	85	5.1	120	88	5.7	17
9	男	69	160.3	66.4	26	23	5.8	4.79	14.9	205	234	7.6	3.8	98	5.1	116	70	6.3	25
10	女	65	144.6	58.7	28	34	4.1	4.38	13	352	211	6.7	4.2	82	4.8	120	78	7.5	17
11	男	45	164.8	51.6	19	15	5.3	5.17	14.9	208	205	7.6	4.9	81	4.6	114	70	5.7	6
12	男	53	175.7	82.2	27	26	3.9	5.23	16.5	166	219	6.9	4	114	5.1	130	86	6	23
13	女	47	152.2	59.4	26	36	5.3	4.45	13.1	178	299	6.8	4	94	4.8	120	74	3.8	12
14	男	67	158	57	23	26	4.3	4.88	15.6	164	218	7.3	4.2	103	5.2	150	80	5.8	41
15	男	69	158.1	66.5	27	28	6.6	5.06	16.5	243	229	7	4.3	93	5	108	74	6.2	27
16	女	53	147	42.6	20	24	3.6	4.11	12.5	212	178	7.3	4.2	86	4.4	120	80	4.5	15
17	女	44	160.4	56.4	22	29	7.9	4.77	14.8	209	167	6.4	4	94	4.8	120	72	4.2	7
18	女	45	153.3	44.1	19	19	4.8	3.49	8.4	202	272	6.4	3.9	86	5.1	104	68	2.6	5
19	女	44	155.6	55.3	23	27	4.2	3.85	10	176	108	7.2	4	92	5.6	102	60	4	6
20	女	41	159.6	61.2	24	33	5	4.18	7.7	186	416	7.2	4.3	90	5.4	122	62	2.9	19

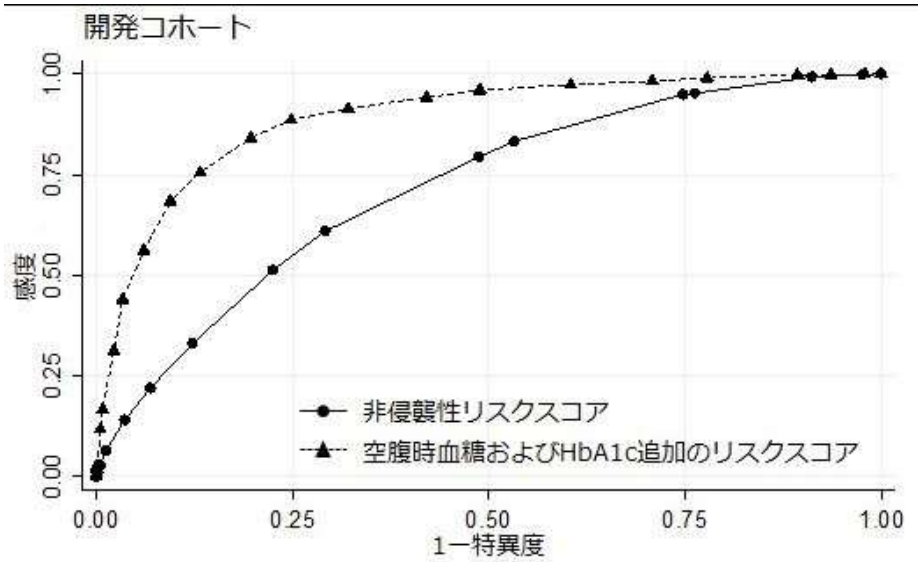
…

疫学における一般的な予測手法）
リスクモデルの作成とROC曲線の描画

項目	カテゴリ	点数
① 性別	女性	0
	男性	1
② 年齢 (歳)	30~39	0
	40以上	2
③ BMI (kg/m²) 体重(kg)÷身長(m)÷身長(m)	23.0未満	0
	23.0~24.9	1
	25.0以上	2
④ 腹部肥満 腹囲：男性90cm以上、女性80cm以上	なし	0
	あり	1
⑤ 喫煙	吸わない	0
	吸う	2
⑥ 高血圧 収縮期血圧140mmHg以上または 拡張期血圧90mmHg以上または 血圧を下げる薬の服用	なし	0
	あり	2
⑦ 空腹時血糖 (mg/dl)	100未満	0
	100~109	3
	110~125	5
⑧ HbA1c (%) 国際基準値	5.6未満	0
	5.6~5.9	3
	6.0~6.4	5

①~⑧の合計点数	発症確率
0~6	<1%
7~9	1~3%
10	4%
11~12	6~9%
13~14	13~18%
15~16	25~34%
17~18	44~55%
19~20	>65%

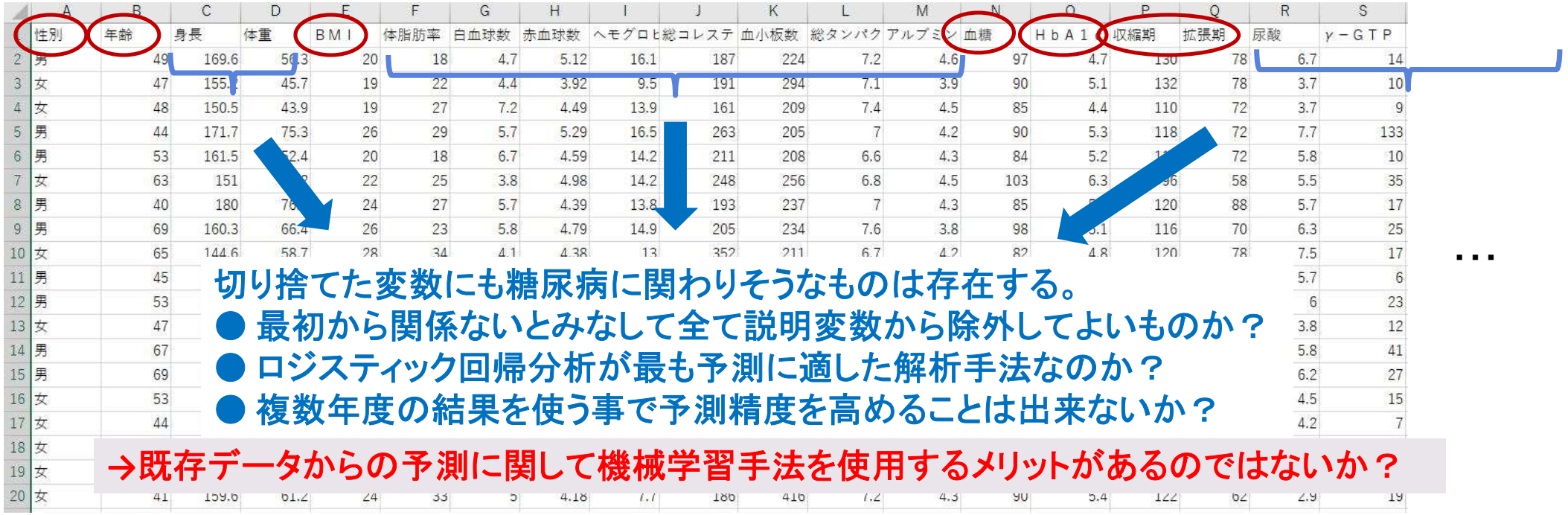
図3　糖尿病発症を予測する空腹時血糖およびHbA1cを追加したリスクスコア



Nanri A, Nakagawa T, Kuwahara K, Yamamoto S, Honda T, et al. (2018)Correction: Development of Risk Score for Predicting 3-Year Incidence of Type 2 Diabetes: Japan Epidemiology Collaboration on Occupational Health Study. PLOS ONE 13(6): e0199075

ここで出てくる疑問

①既存研究や日常の診療において関係のありそうな変数を選択
②糖尿病の発生の有無を目的変数としたロジスティック回帰分析を実施



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	性別	年齢	身長	体重	BMI	体脂肪率	白血球数	赤血球数	ヘモグロビ	総コレステ	血小板数	総タンパク	アルブミン	血糖	HbA1c	収縮期	拡張期	尿酸	γ－GTP
2	男	49	169.6	56.3	20	18	4.7	5.12	16.1	187	224	7.2	4.6	97	4.7	130	78	6.7	14
3	女	47	155.2	45.7	19	22	4.4	3.92	9.5	191	294	7.1	3.9	90	5.1	132	78	3.7	10
4	女	48	150.5	43.9	19	27	7.2	4.49	13.9	161	209	7.4	4.5	85	4.4	110	72	3.7	9
5	男	44	171.7	75.3	26	29	5.7	5.29	16.5	263	205	7	4.2	90	5.3	118	72	7.7	133
6	男	53	161.5	52.4	20	18	6.7	4.59	14.2	211	208	6.6	4.3	84	5.2	112	72	5.8	10
7	女	63	151	52	22	25	3.8	4.98	14.2	248	256	6.8	4.5	103	6.3	136	58	5.5	35
8	男	40	180	76	24	27	5.7	4.39	13.8	193	237	7	4.3	85	5.2	120	88	5.7	17
9	男	69	160.3	66.4	26	23	5.8	4.79	14.9	205	234	7.6	3.8	98	5.1	116	70	6.3	25
10	女	65	144.6	58.7	28	34	4.1	4.38	13	352	211	6.7	4.2	82	4.8	120	78	7.5	17
11	男	45																5.7	6
12	男	53																6	23
13	女	47																3.8	12
14	男	67																5.8	41
15	男	69																6.2	27
16	女	53																4.5	15
17	女	44																4.2	7
18	女																		
19	女																		
20	女	41	159.6	61.2	24	33	5	4.18	7.7	186	416	7.2	4.3	90	5.4	122	62	2.9	19

…

切り捨てた変数にも糖尿病に関わりそうなものは存在する。

- 最初から関係ないとみなして全て説明変数から除外してよいものか？
- ロジスティック回帰分析が最も予測に適した解析手法なのか？
- 複数年度の結果を使う事で予測精度を高めることは出来ないか？

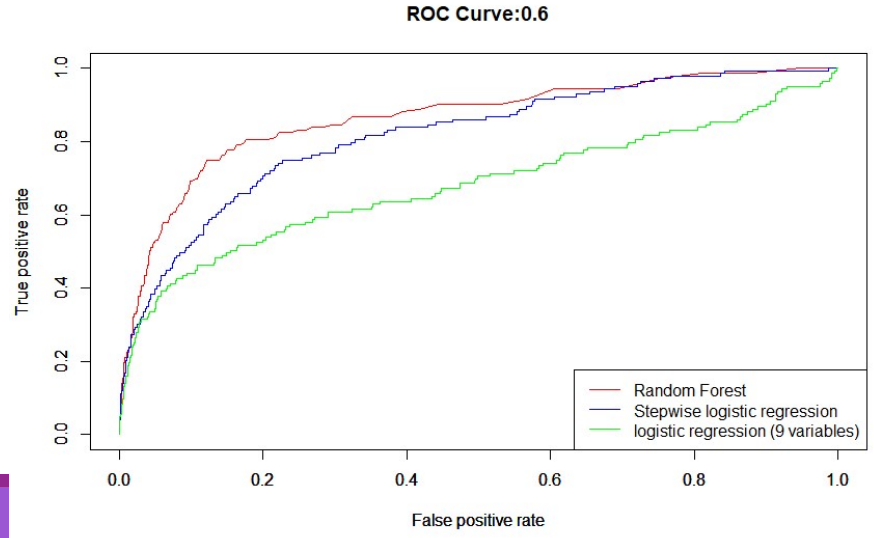
→既存データからの予測に関して機械学習手法を使用するメリットがあるのではないか？

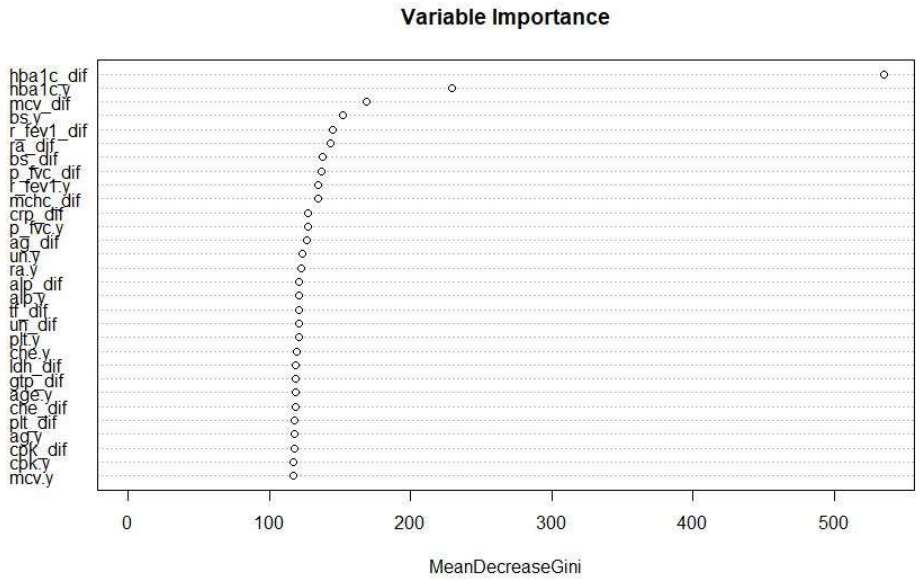
昨年の発表で取り組んだこと

ランダムフォレストを用いた糖尿病発症リスクの予測

① 2年分の健康診断結果から1年後の糖尿病発症リスクを検討（左図）

② 変数重要度(Variable Importance)を確認し、予測因子を探索（右図）





昨年の発表での課題・頂いたコメント

昨年の発表での課題

1. 連続3年以上の健康診断結果を生かして解析をすることは出来ないか？

2. 複数年度の推移も考慮して、適切な変数を選んでモデルを作ることは出来ないか？

↓

3年以上の健康診断結果の差分や交互作用を全て導入した上で、スパースロジスティック回帰分析を行ってみてはどうか。

今回の発表に際して取り組んだこと

スパースモデリングを用いた**糖尿病発症リスク**の予測

・3年分の健康診断結果の**差分**と**交互作用項**を考慮に入れてスパースロジスティック回帰を行い、糖尿病の発症予測と構造確認を行う。

・作成したモデルの現場での実用性を考え、1年後ではなく
「**5年以内の糖尿病発症の有無**」を目的変数として設定した。

・他手法を適用したモデルと比較する事で、スパースロジスティック回帰を用いるメリットを明らかにする。

スパースモデリングの特徴と利点

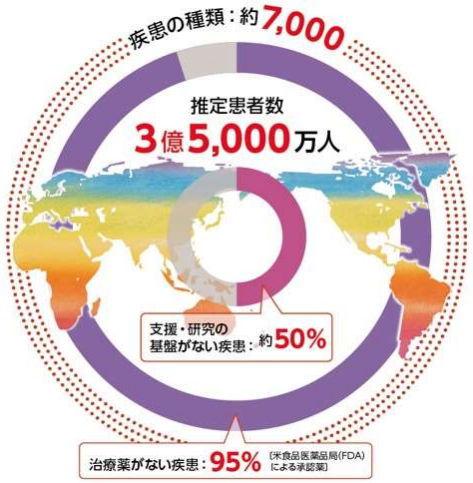
オッカムの剃刀 （William of Ockham、1285-1347年）

「ある事項を説明するためには、必要以上に過程を多くするべきではない。」

↓

同じ説明力を持つならば、特徴量（変数）の数は少ない方がいい

↓



希少疾患データや遺伝子データ等、

医療界には**N<p**となるデータも多い。

$lasso$

Tibshirani(1996)はlassoと呼ばれる次の制約付き最小化問題を提案した。

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 \right\} \quad subject\ to \quad ||\boldsymbol{\beta}||_1 \leq s$$

上式はいくつかのパラメータ推定値を0にする性質（スパース性）を持つ。

これに対して、ラグランジュの未定乗数法を適用する事により、

$$S_\lambda(\boldsymbol{\beta}) = \frac{1}{2n} ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$$

をパラメータ$\boldsymbol{\beta}$に関して最小化することにより得られる解と同値になる。

一般化線形モデル(GLM)への発展

★本研究では、5年以内の糖尿病発症の有無を判別（2クラス判別）するため、ロジスティック回帰分析を用いた。

スパース推定を行うため、ロジスティック回帰分析の対数尤度関数にL1正則化をかけて、これを最大にするベクトル$\boldsymbol{\beta}$を求める。

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-log\{1+\exp\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\}\right]-\lambda||\boldsymbol{\beta}||_1$$

glmnetによるスパースモデリング実装

Jerome Friedmanらにより実装されたスパースモデリングにおける標準的ツール

R,Python,Matlabでのインターフェースがあるが、本研究ではRのglmnetを使用。

正則化項は $\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1$ で定められる。(α =1でlasso、α =0でridge)

CVによって正則化パラメータλの数値を最適化する事が出来る。

健康診断データへの適用

・厚生労働省が定める定期健康診断の項目は約25項目であり、毎年1500万人が受診をしている。

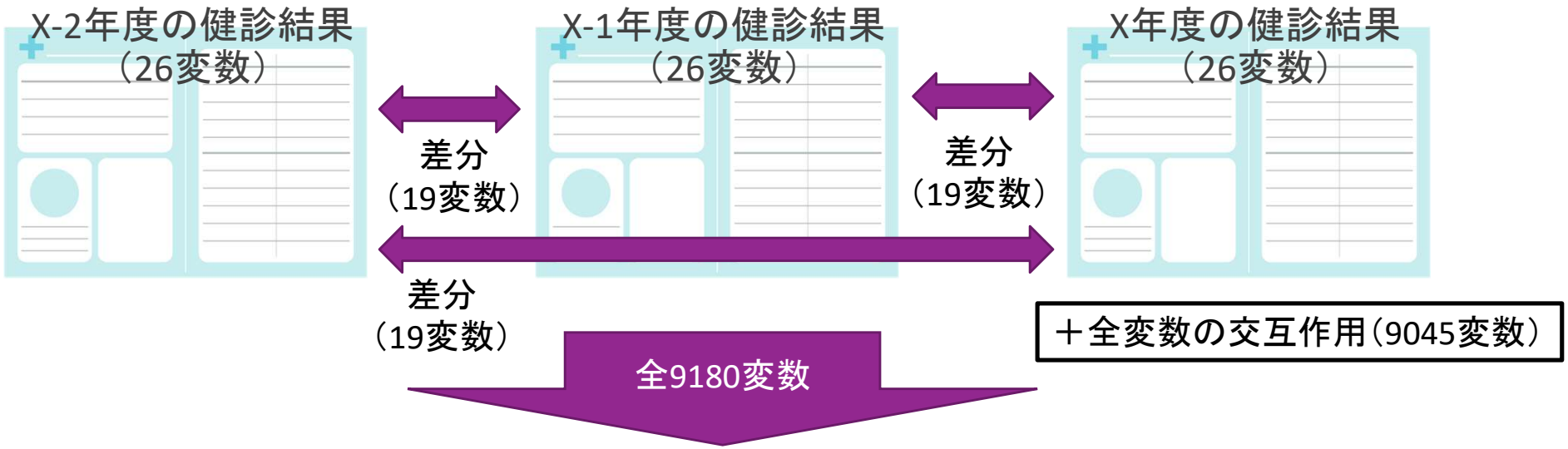
・単年度の結果であればN<pとなることはないが、
複数（n）年度の結果を生かして解析を行う場合、例えば
① 各年度の結果 (25項目 $\cdot\, n$年度 $= 25n$変数)
② 各年度同士の差分 $\left(25\text{項目} \cdot {}_n C_2 = \frac{25}{2} n(n-1)\text{変数}\right)$
③ ①②に関する交互作用項（${}_{\frac{25}{2}n(n+1)} C_2$変数）
の合計数を使う事で、変数はかなり増える。
(例：n=3の場合11175変数、n=10の場合：946000変数）

検査項目			健診コース			特定健診項目（40歳以上必須）
			生活習慣病健診	主婦健診	人間ドック	
身体計測		問診	○	○	○	
		身長	○	○	○	●
		体重	○	○	○	●
		BMI	○	○	○	●
		腹囲	○	○	○	●
		視力	○	○	○	
		聴力	○	○	○	
		血圧	○	○	○	●
尿検査		尿蛋白	○	○	○	●
		尿糖	○	○	○	●
		尿潜血	○	○	○	
血液検査	貧血	赤血球数	○	○	○	
		ヘマトクリット	○	○	○	
		ヘモグロビン	○	○	○	
	肝機能	GOT	○	○	○	●
		GPT	○	○	○	●
		γ-GTP	○	○	○	●
		ALP	○	○	○	
	脂質代謝	総コレステロール	○	○	○	
		中性脂肪	○	○	○	●
		HDLコレステロール	○	○	○	●
		LDLコレステロール	○	○	○	●
	糖代謝	空腹時血糖又はHbA1c	○	○	○	●
	腎機能	クレアチニン	○	○	○	●
	痛風検査	尿酸	○	○	○	
	血清検査	CRP	○	○	○	

厚生労働省が定める企業用の定期健康診断項目

今回行った解析について

説明変数



目的変数

X+1年度～X+5年度（5年以内）に糖尿病を発症したか否か(1/0)

今回利用した健康診断データセット

山梨県内の健康診断施設から2つの異なるデータセットを入手。

① 1999年度～2009年度（10年分）のべ**189,412**人分（8年連続受診者7800人）

② 2011年度～2018年度（8年分）のべ**203,379**人分(8年連続受診者7617人)

※①内、②内では各年度での同一患者をマージすることが出来るが、
①②の間では患者を相互にマージできない仕様で保存されていた。

時間の前後も考慮に入れ、異なるデータセット間での予測能を確認するため、

①の全データを訓練データ、②の全データをテストデータとして解析を行った。

解析に用いたデータの詳細

●健康診断検査項目のうち、厚生労働省が指定する項目26変数を使用

目的変数

X+1年度～X+5年度（5年以内）に糖尿病を発症したか否か(1/0)

説明変数（特徴量）

性別、年齢、身長、体重、BMI、白血球数、赤血球数、ヘモグロビン、ヘマトクリット、γ-GTP、AST、ALT、尿素窒素、クレアチニン、尿酸、中性脂肪、HDLコレステロール、LDLコレステロール、空腹時血糖、HbA1c、収縮期血圧、拡張期血圧、飲酒習慣、喫煙習慣(現在/過去/非喫煙)　（**全26変数×連続3年分**）

※青色の項目の差分は意味をもたないため、差分をとる場合にはそれ以外の項目のみとした。

使用したデータセットイメージ

※以下はダミーデータです

26変数（特徴量）×3年度分　＋　1変数（目的変数：以後5年分を使用）

gender	age	height	weight	bmi	fat	wbc	rbc	hb	ht	mcv	mch	mchc	plt	tp	alb	ag	che	t
1	49	169.6	56.3	20	18	4.7	5.12	16.1	46.1	90	31.4	34.9	224	7.2	4.6	1.77	450	
0	47	155.2	45.7	19	22	4.4	3.92	9.5	30.8	78.6	24.2	30.8	294	7.1	3.9	1.22	294	
0	48	150.5	43.9	19	27	7.2	4.49	13.9	40.8	90.9	31	34.1	209	7.4	4.5	1.55	331	
1	44	171.7	75.3	26	29	5.7	5.29	16.5	47.2	89.2	31.2	35	205	7	4.2	1.5	420	
1	53	161.5	52.4	20	18	6.7	4.59	14.2	42.6	92.8	30.9	33.3	208	6.6	4.3	1.87	372	
0	63	151	50.2	22	25	3.8	4.98	14.2	43.2	86.7	28.5	32.9	256	6.8	4.5	1.96	598	
1	40	180	76.3	24	27	5.7	4.39	13.8	40.2	91.6	31.4	34.3	237	7	4.3	1.59	465	
1	69	160.3	66.4	26	23	5.8	4.79	14.9	45	93.9	31.1	33.1	234	7.6	3.8	1	335	
0	65	144.6	58.7	28	34	4.1	4.38	13	39.5	90.2	29.7	32.9	211	6.7	4.2	1.68	535	
1	45	164.8	51.6	19	15	5.3	5.17	14.9	43.1	83.4	28.8	34.6	205	7.6	4.9	1.81	424	
1	53	175.7	82.2	27	26	3.9	5.23	16.5	47.1	90.1	31.5	35	219	6.9	4	1.38	324	
0	47	152.2	59.4	26	36	5.3	4.45	13.1	38.5	86.5	29.4	34	299	6.8	4	1.43	250	
1	67	158	57	23	26	4.3	4.88	15.6	47.1	96.5	32	33.1	218	7.3	4.2	1.35	394	
1	69	158.1	66.5	27	28	6.6	5.06	16.5	46.9	92.7	32.6	35.2	229	7	4.3	1.59	369	
0	53	147	42.6	20	24	3.6	4.11	12.5	36.8	89.5	30.4	34	178	7.3	4.2	1.35	409	
0	44	160.4	56.4	22	29	7.9	4.77	14.8	43.4	91	31	34.1	167	6.4	4	1.67	369	
0	45	153.3	44.1	19	19	4.8	3.49	8.4	28.4	81.4	24.1	29.6	272	6.4	3.9	1.56	294	

4種類のモデルを作成

モデル	説明変数	変数の数
Model1	連続3年分のうち、最新1年分の検査項目のみ	26
Model2	連続3年分の検査項目のみ	78
Model3	連続3年分の検査項目＋各年度の差分	135
Model4	連続3年分の検査項目＋各年度の差分 ＋検査項目と差分の全交互作用項	9180

3種類の手法を適用

① スパースロジスティック回帰分析

② ランダムフォレスト

③ ロジスティック回帰分析＋ステップワイズ法（変数増減法）

前述の4種類の説明変数セットを用いて、

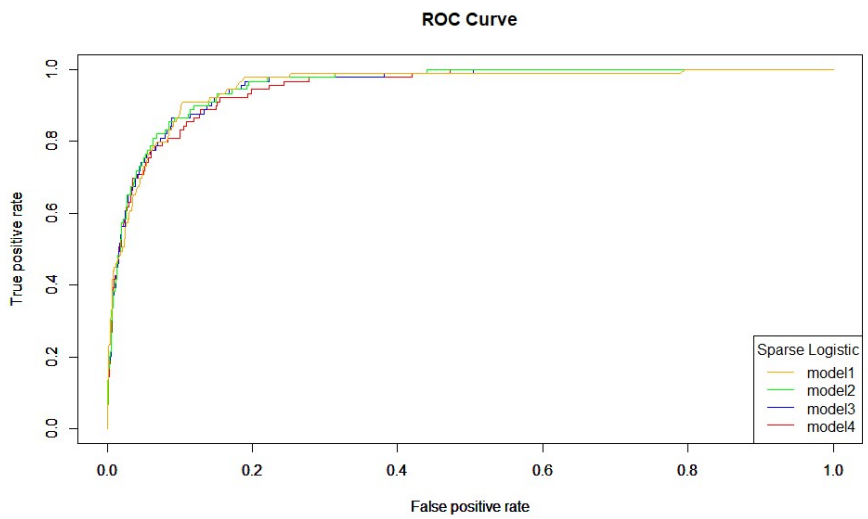
5年以内の糖尿病発症の有無を予測するモデルを構築

＋　テストデータを用いてROC曲線・AUCを算出

解析結果①（スパースロジスティック回帰）

●スパースロジスティック回帰分析を適用した場合

※使用したRコード （glmnetパッケージ：R ver3.6.1を使用）
model_sl<- cv.glmnet(x_train, y_train , nfolds = 5, family="binomial", standardize=TRUE)
（説明変数：x_train，目的変数:y_train，CVの分割数:nflods，family：目的変数の種類, standardize:標準化）

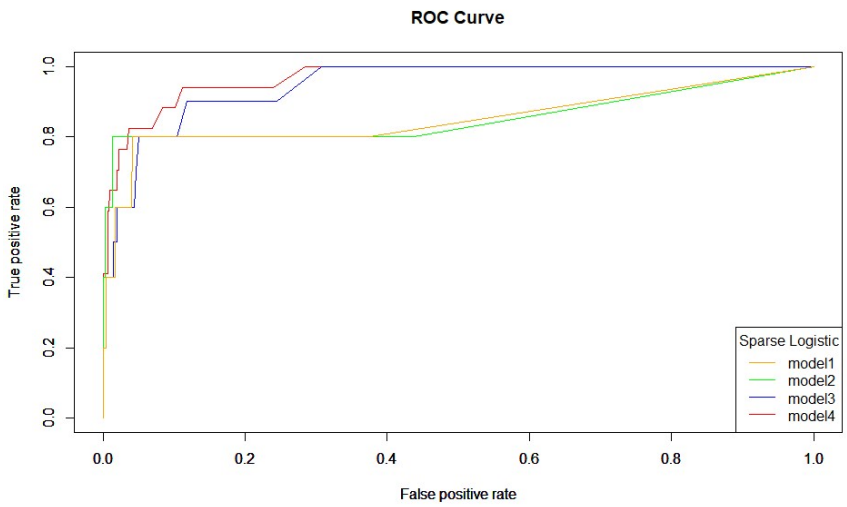


モデル	説明	AUC
Model1	単年度 項目のみ	0.952
Model2	連続3年度 項目のみ	0.955
Model3	連続3年度 項目＋差分	0.953
Model4	連続3年度 項目＋差分＋交互作用項	0.948

解析結果②（ランダムフォレスト）

●ランダムフォレストを適用した場合

※使用したRコード （randomForestパッケージ：R ver3.6.1を使用）
model_rf <- randomForest(dm_5year ~., data= dataset_all, ntree=500)
（目的変数:dm_5year，訓練データセット:dataset_all，決定木の本数：ntree）

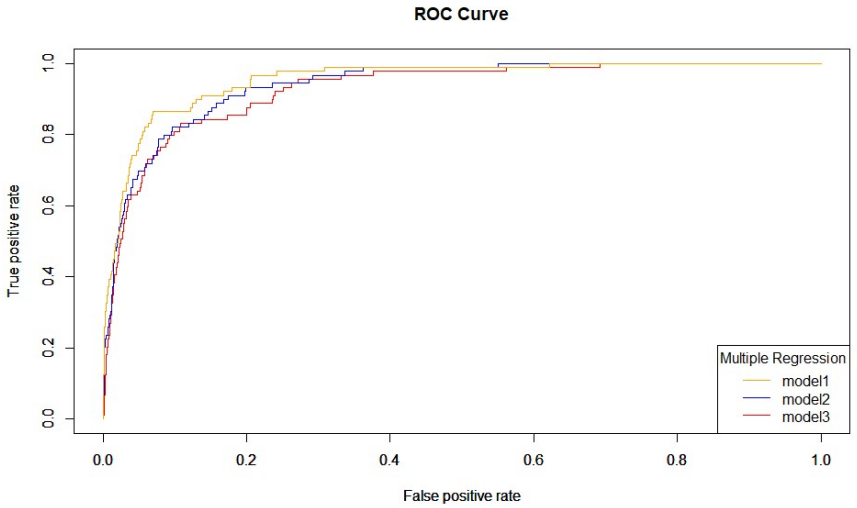


モデル	説明	AUC
Model1	単年度 項目のみ	0.928
Model2	連続3年度 項目のみ	0.949
Model3	連続3年度 項目＋差分	0.951
Model4	連続3年度 項目＋差分＋交互作用項	0.947

解析結果③（ロジスティック回帰＋ステップワイズ）

●ロジスティック回帰＋ステップワイズ法を適用した場合

※使用したRコード （glm,statsパッケージ：R ver3.6.1を使用）
model_mr <- glm(dm_5year ~., family=binomial, data=train_dataset)
step.result <- stats::step(model_mr)（目的変数:dm_5year，stats::step:ステップワイズ法（変数減少法））



モデル	説明	AUC
Model1	単年度 項目のみ	0.953
Model2	連続3年度 項目のみ	0.939
Model3	連続3年度 項目＋差分	0.927
Model4	連続3年度 項目＋差分＋交互作用項	解析不能

解析結果まとめ

モデル	説明	変数の数	SparseLogistic	RandomForest	StepwiseLogistic
			AUC	AUC	AUC
Model1	単年度 項目のみ	26	0.952	0.928	0.953
Model2	連続3年度 項目のみ	78	0.955	0.949	0.939
Model3	連続3年度 項目＋差分	135	0.953	0.951	0.927
Model4	連続3年度 項目＋差分＋交互作用項	9180	0.948	0.947	解析不能

解析結果からの考察

- スパースロジスティック回帰では、どのモデルでも精度があまり上下しなかった。

→ 複数年度のデータや差分・交互作用は予測にあまり寄与していない可能性。

- ランダムフォレストでは特徴量が少ないと精度が低くなる傾向にあった
- ステップワイズ＋ロジスティックは特徴量が多いと精度が低くなる傾向にあった。
- オッカムの剃刀の考えを土台にすると、Model1が最も良く使えそう。

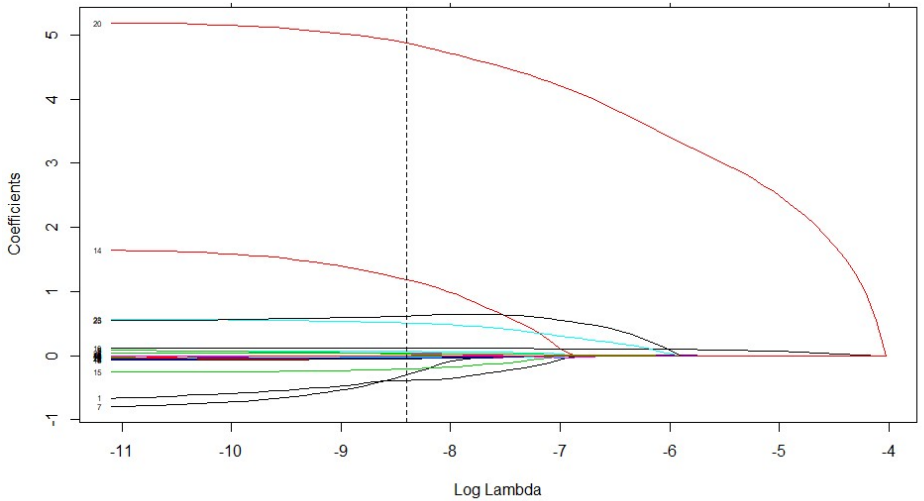
因子探索 model1:単年度の結果

●スパースロジスティック回帰分析

※使用したRコード （glmnetパッケージ：R ver3.6.1を使用）

```
> plot_glmnet(model_sl_one,label=20)
> plot(model_sl_one_1, xvar="lambda", label=20)
> abline(v=log(model_sl_one$lambda.min),lty=2)
```

↓点線はCVにより得られたλの最適値

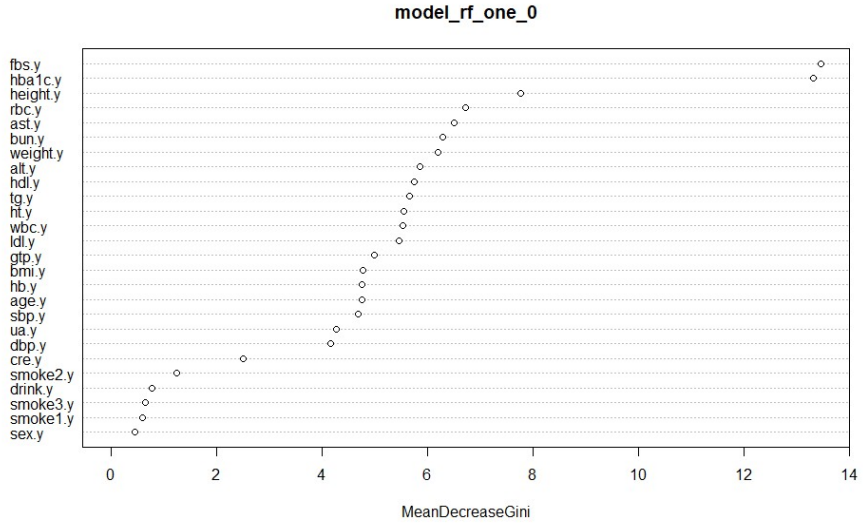


変数	標準化偏回帰係数
Hba1c	1.493
Fbs	1.007
Sbp	0.573
Smoke	0.273
Drink	0.254
Cre	0.191
Bmi	0.184
…	…
Ldl	-0.241
Ua	-0.273
Dbp	-0.375
Age	-0.379
Hdl	-0.386

因子探索 model1:単年度の結果

●ランダムフォレスト（Variable Importance）

※使用したRコード （randomForestパッケージ）
> model_rf_one <- randomForest(dm_5year ~., data= val_dataset_one, ntree=500)
> varImpPlot(model_rf_one,type=2)



変数	ジニ係数
Fbs	13.46
Hba1c	13.32
Height	7.76
Rbc	6.72
Ast	6.50
Bun	6.28
Weight	6.20
Alt	5.85
Hdl	5.75
Tg	5.65
…	…

因子探索 model1:単年度の結果

●ロジスティック回帰＋ステップワイズ法

※使用したRコード （glm,statsパッケージ）

```
> model_mr_one_0 <- glm(dm_5year ~., family=binomial,
data=val_dataset_one)
step.result_one_0 <- stats::step(model_mr_one_0)
```

変数	標準化偏回帰係数
Hba1c	5.038
Cre	1.539
Smoke	0.865
Drink	0.718
Fbs	0.123
Bmi	0.076
Sbp	0.043
Ldl	-0.009
Alt	-0.014
Hdl	-0.031

因子探索 model1:まとめ

Sparse Logistic(AUC:0.952)		Random Forest(ROC:0.928)		Stepwise Logistic(AUC:0.953)	
変数	標準偏回帰係数	変数	ジニ係数	変数	標準偏回帰係数
Hba1c	1.493	Fbs	13.46	Hba1c	5.038
Fbs	1.007	Hba1c	13.32	Cre	1.539
Sbp	0.573	Height	7.76	Smoke	0.865
Hdl	-0.386	Rbc	6.72	Drink	0.718
Age	-0.379	Ast	6.50	Fbs	0.123
Dbp	-0.375	Bun	6.28	Bmi	0.076
Smoke	0.273	Weight	6.20	Sbp	0.043
Ua	-0.273	Alt	5.85	Hdl	-0.031
Drink	0.254	Hdl	5.75	Alt	-0.014
Ldl	-0.241	Tg	5.65	Ldl	-0.009

因子探索からの考察

・収縮期血圧が高いことが、5年以内の糖尿病発症と関連しているかもしれない。

・HDL(善玉）コレステロールが低いことが、5年以内の糖尿病発症と関連しているかもしれない。

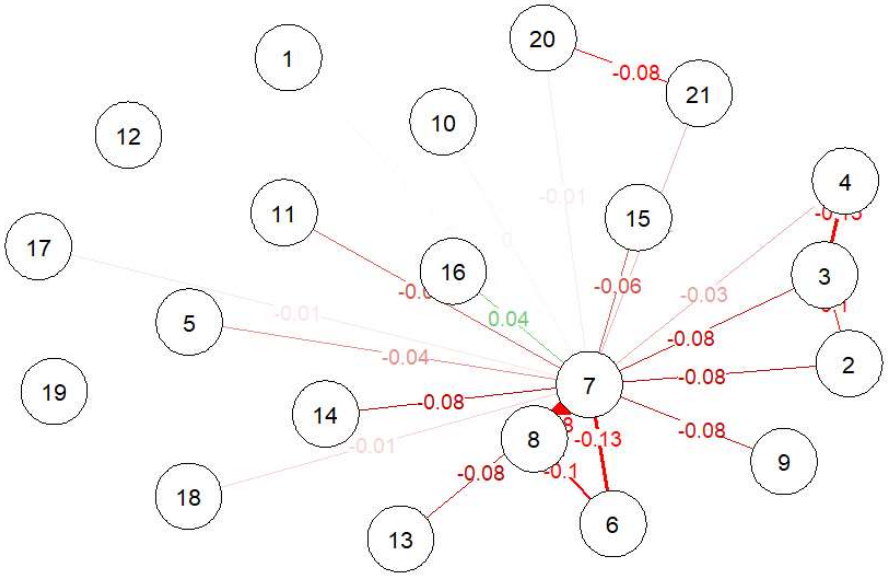
※ただし、スパースモデリングは単に「**関係の有無**」を発見する道具にすぎないため、これだけで精密な議論をすることは出来ない。

構造探索$Graphical\ lasso(glasso)$

$Graphical\ lasso$とは、ガウシアングラフィカルモデル$N(\mu, \Omega)$に従う確率変数ベクトルがあった時、変数間の関係を指定し、グラフ化する手法の事。

精度行列の各成分にL_1正則化を課して、最尤推定を行う。

$$\max_{\Lambda} log|\Lambda| - Tr\hat{\Sigma}\Lambda - \rho||\Lambda||_1$$



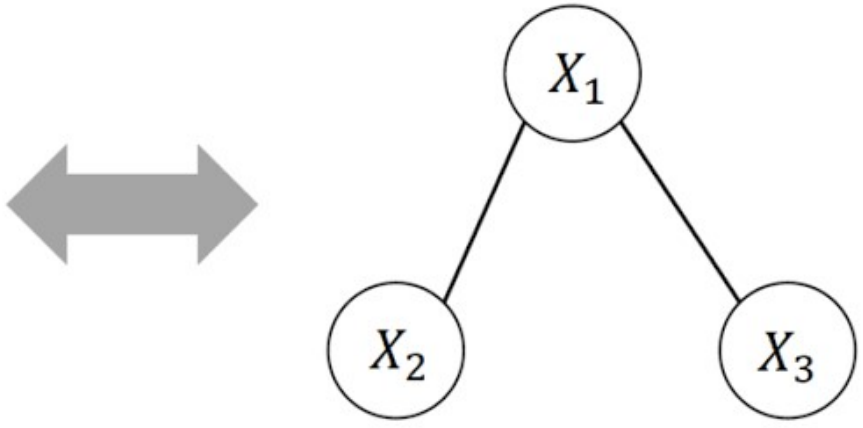
ガウシアングラフィカルモデルとは

ガウシアングラフィカルモデル(GGM: Gaussian graphical model)

多変量正規分布$N(\boldsymbol{\mu}, \Sigma)$に従う確率変数ベクトル$\boldsymbol{X} = (X_1, \ldots, X_p)^T$があった時に、変数$X_j$と$X_k (j \neq k)$の関係性を推定し、それをグラフ化する手法のこと

Ωを

$$\Omega = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} * & * & * \\ * & * & 0 \\ * & 0 & * \end{bmatrix} \end{array}$$



し、

上式におい

独立となる。

共分散

応する。

*Graphical lasso*の実装（下準備）

```
# 陽性例の群と陰性例の群に場合分けしておく
xpos_one <- g_val_dataset_one %>% filter(dm_5year == 1) %>% select(-dm_5year) %>% scale()
xneg_one <- g_val_dataset_one %>% filter(dm_5year == 0) %>% select(-dm_5year) %>% scale()

# Box-cox変換（上で作った二群を正規分布に近づける）※xneg_oneは省略
x <- xpos_one
for (I in 1:dim(x)[2]) {
  tmp <- powerTransform(x[, i] –min(x[,i]) + 0.1)
  xpos_one[, i] <- bcPower(x[, i] –min(x[,i])+ 0.1, tmp$roundlam)
}
```

*Graphical lasso*の実装

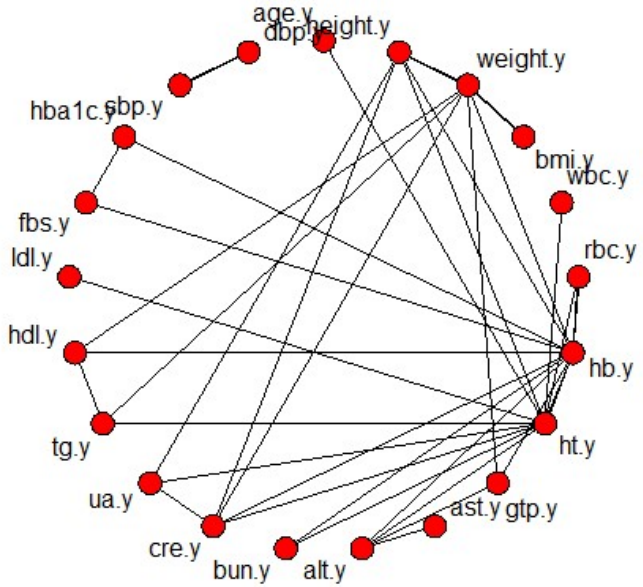
```
# Graphical lasso Plot

ret_one <- glasso(cov(xpos_one),rho=.3)
ret2_one <- glasso(cov(xneg_one),rho=.3)

gplot(abs(ret_one$wi), gmode="graph", jitter=FALSE, displaylabels=TRUE,
edge.lwd=abs(ret_one$wi),label=colnames(x),mode="circle")

gplot(abs(ret2_one$wi), gmode="graph", jitter=FALSE, displaylabels=TRUE,
edge.lwd=abs(ret2_one$wi),label=colnames(x),mode="circle")
```

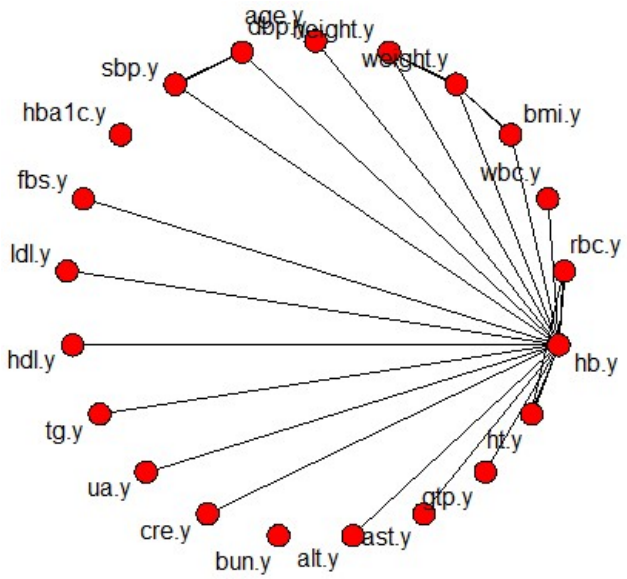
単年度健診結果　*Graphical lasso*



精度行列の値

row	col	value
hb.y	rbc.y	-0.352
gtp.y	ast.y	-0.279
weight.y	height.y	-0.249
wbc.y	rbc.y	-0.228
age.y	height.y	-0.219
height.y	ht.y	-0.137
cre.y	bun.y	-0.115
hba1c.y	sbp.y	-0.110
ua.y	tg.y	0.100
ldl.y	fbs.y	-0.093

5年以内糖尿病発症有り群



精度行列の値

row	col	Value
hb.y	rbc.y	-0.382
weight.y	height.y	-0.131
wbc.y	rbc.y	-0.125
wbc.y	hb.y	-0.101
age.y	height.y	-0.098
hba1c.y	sbp.y	-0.083
cre.y	rbc.y	-0.082
bun.y	rbc.y	-0.082
rbc.y	age.y	-0.080
height.y	rbc.y	-0.078

5年以内糖尿病発症無し群

構造探索からの考察

・糖尿病を発生しやすい人は、各因子とヘモグロビンとの関連性が崩れている可能性がある。

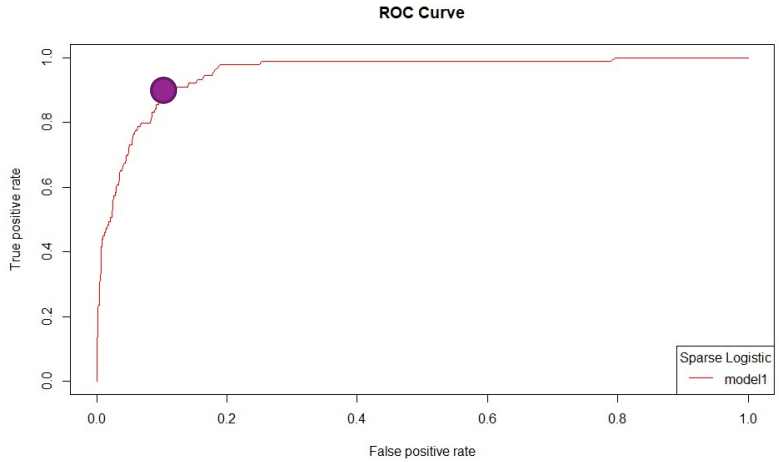
・糖尿病を発生しやすい人は、γGTPとASTが逆相関している（どちらかだけ高い）可能性がある。

※ただし、スパースモデリングは単に「**関係の有無**」を発見する道具にすぎないため、これだけで精密な議論をすることは出来ない。

（さいごに）
作成したモデルは実用性があるのか？

単年度の結果のみを用いたModel1の実用最適モデル

予測 | 正答	0	1	
0	6741	8	感度91.0%
1	787	81	特異度89.4%



- ● 感度・特異度共に9割近く、5年以内の糖尿病発症を予測できている。
- ● このモデルを用いることで、実際の保健指導の対象者を適切に絞ることが出来る可能性は高い。
- ● 異なる時期のテストデータを用いて検証したため、時代変化によるバイアスにも耐えられそう
- ▲ 陽性的中率は9.3%しかないので、対象者になっても危機感を抱いてくれないかもしれない。

本研究の強みと限界点、良かったこと

■ 5年以内の糖尿病発症を十分な精度で予測するモデルを作成することが出来た。

■ 予測に関係のない差分などの変数が投入されても精度が保てていることは、スパースモデリングの強みである。

▲ スパースロジスティックにおける因子探索については、回帰係数にバイアスが含まれているため、精密な議論はできない。

● スパースモデリングを医療者として実際に理解し、疫学データへ実装する事が出来た。（ LaTeX も少し書けるようになった。）

今後の展望

・より発展したスパースモデリングや他手法の組み合わせの検討

→　より陽性的中率の高いモデルを作成することは出来るのか

・深層学習等による予測率との比較、他施設のデータセットでのテスト

→　深層学習等の他手法や他施設データセットでも、精度を確認する

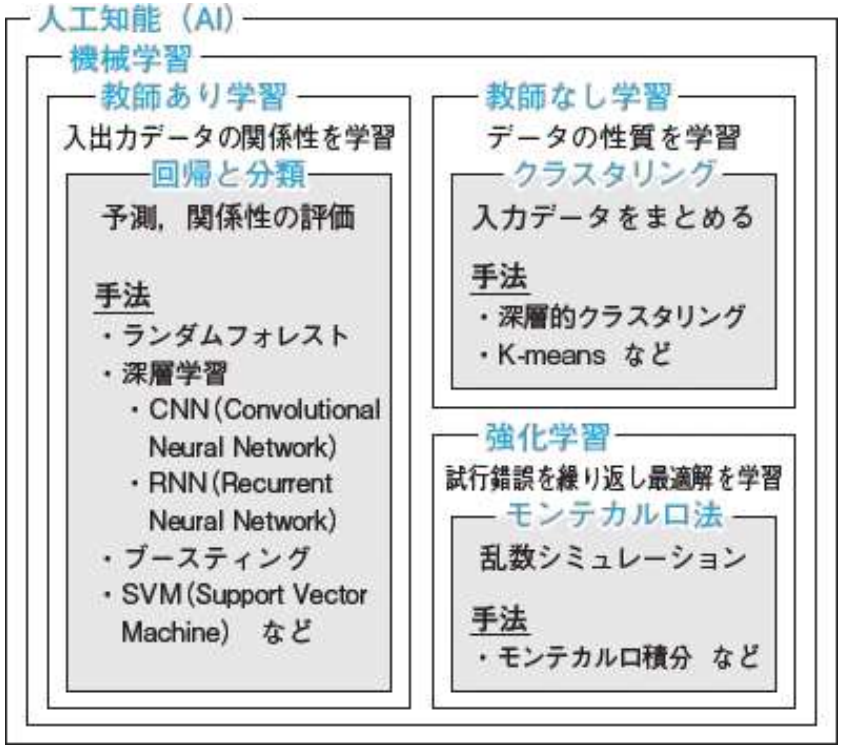
・各々の統計数理手法がどの医療領域でメリットを生むのかについての検討

→　ランダムフォレストやスパースモデリングを他医療領域の課題にも適用

・医療界における既存の疾患Gradeやフェノタイプを疑ってみる

→　「ディープフェノタイプ研究」の推進

各々の統計数理手法がどの医療領域でメリットを生むのかについての検討・実装



総合診療内科	脳神経内科	呼吸器内科
循環器内科	腎臓・高血圧内科	糖尿病内科
消化器内科	外科・消化器外科	乳腺科
整形外科	眼　科	耳鼻咽喉科
皮膚科	泌尿器科	もの忘れ外来
心療内科	鍼灸外来	

ディープラーニング一辺倒ではなく、様々な統計数理手法が医療のどの分野で活用できそうなのか検討し、実装してみる

ディープフェノタイプ研究の推進

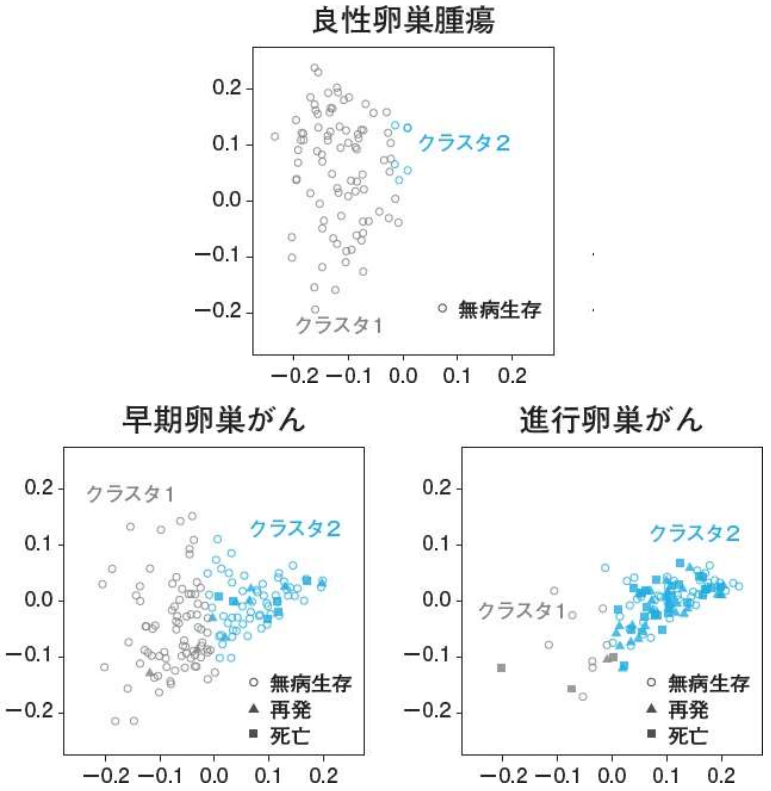
ディープフェノタイプ研究

まだ見つけられていない疾患フェノタイプを
機械学習等の活用によって発見し、
精密な発症予測や治療選択につなげる

● 従来から定められてきた疾患分類は本当に正しいのだろうか？

● 病気を発症しやすいグループを早期に同定し、適切な事前介入を実施できるのではないか？

● 同一疾患同一治療ではなく、一人一人の治療感受性による最適投薬ができるのでは？



ご指南いただきたいこと

・今回の手法や解析全般に対する忌憚なきご意見やご指南

・本研究のデータを用いて、より精度を上げると思われる機械学習手法、または複数手法の組み合わせにより精度を上げる方法はあるか

・医療分野・疾患発症予測に適用できそうな統計数理の手法はあるか

・その他、もし疫学・医療分野でお手伝いできることがあれば。

謝辞

今回、このような貴重な場を提供していただいた
福水先生はじめ統計数理研究所の皆様に心より感謝申し上げます。