

明治大学総合数理学部

2023 年度

卒業研究

ヘルスケアデータを用いたポリファーマシーに関連する危険因子の調査

学位請求者 先端メディアサイエンス学科

石山 晴斗

目次

第 1 章	はじめに	3
第 2 章	先行研究	4
第 3 章	ポリファーマシーの評価	5
3.1	ポリファーマシーの定義	5
3.2	データセット	5
3.3	統計分析	5
3.4	基本統計量	6
3.5	ロジスティック回帰分析	6
3.6	考察	6
第 4 章	おわりに	16
	参考文献	17
付録 A	匿名化された健康診断と診療履歴の時系列データによる糖尿病罹患予測	19
A.1	はじめに	19
付録 B	データ	21
B.1	概要	21
B.2	健康診断データの前処理について	21
B.3	個人データを用いた特徴量変化の調査	22
B.4	時系列情報	23
付録 C	分析	25
C.1	分析方法	25
C.2	罹患予測モデル	26
C.3	一意率について	26
付録 D	分析結果	30
D.1	ロジスティック回帰	30
D.2	ニューラルネットワークによるモデル	31
D.3	サポートベクトルマシンによるモデル	31

D.4	診療履歴を考慮した一意率	31
付録 E	考察	35
E.1	診療履歴を考慮した特徴量変化について	35
E.2	ロジスティック回帰について	35
E.3	罹患予測モデルについて	36
E.4	診療履歴を考慮した一意率の推移について	36
付録 F	終わりに	37
	参考文献	38
付録 G	分担表	39

第1章

はじめに

日本の高齢者の人口増加に伴いポリファーマシーが増加している [1]. ポリファーマシーは、本来の疫病に必要以上の多くの薬を同時に飲んでいることにより引き起こされる有害事象のことである。服薬管理が複雑化し、薬物有害事象のリスクが高まることから大きな課題となっている。石崎らは、東京都の100万人以上の高齢者における薬物処方パターンとポリファーマシーに関連する因子について調査を行った [2]. しかしながら、ポリファーマシーと医薬品の関係のみに終始しており、医薬品の処方数に影響を及ぼす可能性がある健康診断や、傷病の種類については配慮がされていない。

そこで本研究の目的は、匿名加工されたヘルスケアデータを用いて、60歳以上の高齢外来患者における、ポリファーマシーと関連する因子を健康診断項目、傷病項目、医薬品項目の観点から明らかにすることである。匿名加工されたヘルスケアデータは、適用（被保険者台帳）データ、健康診断データ、傷病レセプトデータ、医薬品レセプトデータ、医薬品算定日レセプトデータから成っている。それぞれのデータに格納されている特徴量とポリファーマシーの関連を定量化する。

第 2 章

先行研究

石崎らは、東京都に在住する外来を受診したことある 75 歳以上で、2014 年 5 月から 8 月の間に 14 日以上慢性疾患の経口投与薬を定期的に処方されていた者を対象として、探索的因子分析を用いて 16 種類の主な薬剤の処方パターンを調査した [2]。多変量ロジスティック回帰モデルを用いて、5 種類以上の薬剤の併用と定義されるポリファーマシーの危険因子を調査した。

ポリファーマシーの有意な危険因子は、男性、80 歳以上の患者、受診医療機関数の多さ、医師による往診の利用の有無、期間中の入院などであった。そして、ポリファーマシーと最も強く関連した主な薬剤の種類は、鎮痛薬、利尿薬、抗糖尿病薬であった。

第3章

ポリファーマシーの評価

3.1 ポリファーマシーの定義

2023年現在、ポリファーマシーの定義について国際的な合意は得られていない。本稿では、石崎らの研究[2]に従い、ポリファーマシーを、個人における5種類以上の薬剤の併用処方と定義する。

3.2 データセット

本研究では、あるヘルスケア企業が取得して匿名加工した健康診断データ、適用（被保険者台帳）データ、傷病レセプトデータ、医薬品レセプトデータ、医薬品算定日レセプトデータ、医療機関レセプトデータを使用する。健康診断データは、BMIや中性脂肪などの身体的特徴12属性と問診結果1属性の計13属性の健康診断結果から成る。各レセプトデータは、各被験者が診断、処方された傷病（56種類）、医薬品（30種類）、医療機関の記録から成る。適用（被保険者台帳）データは被験者の生年月日や性別を示す。

60歳以上の患者を対象とし、2020年の健康診断データかつ、2020年5月1日から同年8月31日までのレセプトデータが1つでも存在する399,120人とした。各患者に処方された薬剤の種類数は、4ヶ月の調査期間中に処方された薬剤を数えることで定量化し、薬剤は14日以上処方されたものと定義する。医薬品、傷病の種類については、レセプトデータ内の頻出する事例から選定する。

3.3 統計分析

本研究では、ポリファーマシーに関連する因子を検討するため、多変量ロジスティック回帰分析を行う。目的変数 y をポリファーマシーの有無、説明変数 x を健康診断データ内13個の属性のみ、傷病レセプトデータ内59個の属性のみ、医薬品レセプトデータ内30個の属性とする。医薬品属性とポリファーマシーの関連が強すぎるため、説明変数を健康診断、傷病、医薬品の3つに分けて、3つのモデルで分析する。それぞれのデータセットについてロジスティックモデル

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

を用いて、交絡因子の影響を考慮した各説明変数 X_i のオッズ比 Odds Ratio(OR)、P値を求める。なお、年齢、性別、BMIの説明変数はそれぞれのデータセットに共通して用いる。

3.4 基本統計量

表 3.1 に適用, 健診, 医薬品データセットの統計量を示す. 患者の平均年齢は, 74.4 歳 (標準偏差 6.8) である. BMI の基準値 25 を超える割合は合計で約 4% を占めていた. 収縮期血圧の基準値 140 を下回る正常者の割合が全体の 68.7% なのに対して, 拡張期血圧の基準値 90 を下回る正常者の割合は 90.2% を占めている. got は基準値 10 から 40 までの間を占める割合は, 94.7% と正常者は多い傾向にあった.

処方された医薬品に関して, カルシウム拮抗剤やレニンアンジオテンシン系作用薬などが全体の 4 割近くの患者に処方されていた.

表 3.2 診断された傷病に関して, 6 割近くの患者が高血圧症と診断されており, 顕著な特徴を示した. ポリファーマシーの医薬品処方数が 5 つ以上の患者は全体の 34.1% を占めていた.

3.5 ロジスティック回帰分析

3.5.1 健康診断データ

説明変数を健康診断データの属性のみとしたロジスティック回帰分析の結果を表 3.3 に示す.

ポリファーマシーとの関連が特に強く, 有意であった属性は年齢, BMI, Hba1c, LDL, e-GFR, 処方日数平均であった. 全て $p < 0.05$ であった.

3.5.2 傷病データ

説明変数を傷病レセプトデータの属性のみとしたロジスティック回帰分析の結果を表 3.4 に示す. 傷病のうち, 高血圧のオッズ比が 1.521 とポリファーマシーに特に強い関連が見られた. 他にも, 不眠症 (OR = 1.332) や高コレステロール血症 (OR = 1.234) など, 生活習慣病を示す傷病とも有意な関連が見られた.

3.5.3 医薬品データ

説明変数を健康診断データの属性としたロジスティック回帰分析の結果を表 3.5 に示す.

健康診断データ, 傷病データに比べて, ポリファーマシーへの相関が強い. 特に, カルシウム拮抗剤, ビタミン剤, レニン-アンジオテンシン系作用剤はオッズ比が 2.5 前後とかなり高い関連を示している. しかしながら, 年齢と BMI は有意ではない ($p > 0.05$) ことを示めている.

3.6 考察

患者の 34.1% がポリファーマシーと定義されており, ポリファーマシーは例外ではなく, 標準的であることが示された. 先行研究 [2] では, 全体の 63.5% がポリファーマシーであった.

3.6.1 健康診断データ

ロジスティック回帰分析の結果に関して, 性別は先行研究同様, 男性の方が, ポリファーマシーのリスクが高いことを示している. 表 3.6 に男女別の傷病罹患率を示す. 傷病数は種類が多いため一部抜粋し, 男性 (%) - 女

表 3.1 健診と医薬品の統計量

	All subjects (n = 399,120)				Polypharmacy subjects (n = 136,225)			
	N	%	mean	std	N	%	mean	std
sex	399,120							
men	177,879	44.6			64,828	47.6		
women	221,241	55.4			71,397	52.4		
age	399,120		74.371	6.766	136,225		76.372	7.378
60代	84,889	21.3			21,217	15.6		
70代	233,412	58.5			72,180	53.0		
80代	70,431	17.6			35,944	26.4		
90代	10,124	2.5			6,714	4.9		
100歳以上	254	0.1			170	0.1		
bmi	399,120		23.398	3.451	136,225		23.860	3.656
bmi < 18.5	23,838	6.0			7,411	5.4		
bmi 18.5 - 25	258,468	64.8			81,463	59.8		
bmi 25 - 30	101,073	25.3			39,740	29.2		
bmi ≥ 30	15,741	3.9			7,611	5.6		
収縮期血圧	399,120		132.893	16.608	136,225		132.870	16.484
収縮期血圧 140	274,219	68.7			93,621	68.7		
収縮期血圧 ≥ 140	124,901	31.3			42,604	31.3		
拡張期血圧	399,120		75.784	10.615	136,225		74.559	10.617
拡張期血圧 < 90	360,079	90.2			125,286	92.0		
拡張期血圧 ≥ 90	39,041	9.8			10,939	8.0		
中性脂肪	399,120		119.986	73.097	136,225		123.085	75.616
中性脂肪 < 30	523	0.1			170	0.1		
中性脂肪 30 - 150	308,205	77.2			103,329	75.9		
中性脂肪 ≥ 150	90,392	22.6			32,726	24.0		
hdl	399,120		63.797	16.807	136,225		61.800	16.619
hdl < 40	18,172	4.6			8,349	6.1		
hdl 40 - 70	249,993	62.6			88,967	65.3		
hdl ≥ 70	130,955	32.8			38,909	28.6		
ldl	399,120		118.122	29.192	136,225		111.773	29.111
ldl < 70	14,297	3.6			8,104	5.9		
ldl 70 - 140	297,681	74.6			105,845	77.7		
ldl ≥ 140	87,142	21.8			22,276	16.4		
got	399,120		24.725	10.331	136,225		25.242	11.224
got < 10	17	0.0			14	0.0		
got 10 - 40	378,141	94.7			127,568	93.6		
got ≥ 40	20,962	5.3			8,643	6.3		
hba1c	399,120		5.864	0.671	136,225		5.997	0.786
hba1c < 5.5	93,254	23.4			28,098	20.6		
hba1c 5.5 - 6.5	253,376	63.5			81,157	59.6		
hba1c ≥ 6.5	52,490	13.2			26,970	19.8		
e_gfr	399,120		66.435	14.243	136,225		64.072	15.668
e_gfr < 59	114,679	28.7			49,408	36.3		
e_gfr 59 - 90	263,380	66.0			79,984	58.7		
e_gfr ≥ 90	21,061	5.3			6,833	5.0		
喫煙者	35,211	8.8			11,492	8.4		
医薬品処方日数平均			41.222	19.711			7.367	2.718
医薬品処方数 ≥ 5 (polypharmacy)	136,225	34.1			-	-		
その他のすべての治療用製剤	42,706	10.7			27,955	20.5		
その他の中枢神経作用薬	18,360	4.6			12,950	9.5		
カルシウム拮抗剤	134,503	33.7			62,970	46.2		
ビタミン剤	66,653	16.7			39,680	29.1		
ベータ遮断剤	31,131	7.8			20,967	15.4		
レニンアンジオテンシン系作用薬	140,490	35.2			66,034	48.5		
便秘薬および腸管洗浄剤	42,307	10.6			29,589	21.7		
全身作用の性ホルモン剤および同効薬	11,175	2.8			5,820	4.3		
全身用コルチコステロイド剤	6,785	1.7			4,616	3.4		
全身用抗ヒスタミン剤	42,706	10.7			22,899	16.8		
利尿剤	17,162	4.3			12,485	9.2		
利胆剤および肝保護剤	8,382	2.1			5,211	3.8		
制酸剤、抗鼓腸剤および抗潰瘍剤	142,885	35.8			89,261	65.5		
咳および感冒用製剤	15,566	3.9			10,625	7.8		
心臓治療薬	11,974	3			9,179	6.7		
抗てんかん剤	18,759	4.7			13,566	10.0		
抗喘息および慢性閉塞性肺疾患用製剤	12,772	3.2			8,163	6.0		
抗炎症および抗リウマチ薬	36,320	9.1			24,103	17.7		
抗痛風製剤	39,114	9.8			20,862	15.3		
抗肥満製剤を除く精神刺激剤	16,364	4.1			11,857	8.7		
抗血栓剤	67,850	17			47,048	34.5		
機能的胃腸障害用薬	16,763	4.2			12,508	9.2		
泌尿器科用剤	35,921	9			22,564	16.6		
甲状腺療法	11,175	2.8			5,845	4.3		
精神抑制薬	63,460	15.9			40,523	29.7		
糖尿病用剤	53,083	13.3			31,740	23.3		
脂質調整/抗アテローム製剤	177,209	44.4			79,383	58.3		
脳および末梢血管用剤	6,386	1.6			4,391	3.2		
腸疾患用製剤	16,763	4.2			11,906	8.7		
鎮痛剤	14,767	3.7			10,975	8.1		

表 3.2 傷病の統計量

	All subjects (n = 399,120)				Polypharmacy subjects (n = 136,225)			
	N	%	mean	std	N	%	mean	std
高血圧症	242,266	60.7			102,034	74.9		
高脂血症	108,162	27.1			47,641	35.0		
高コレステロール血症	110,556	27.7			46,253	34.0		
不眠症	82,219	20.6			45,879	33.7		
慢性胃炎	86,609	21.7			44,566	32.7		
腰痛症	79,425	19.9			40,154	29.5		
アレルギー性鼻炎	85,412	21.4			39,229	28.8		
便秘症	80,622	20.2			46,760	34.3		
骨粗鬆症	71,842	18			36,182	26.6		
逆流性食道炎	63,061	15.8			33,851	24.8		
歯周炎	65,855	16.5			25,858	19.0		
脂質異常症	57,473	14.4			25,143	18.5		
維持療法の必要な難治性逆流性食道炎	45,101	11.3			27,133	19.9		
胃潰瘍	41,908	10.5			22,566	16.6		
2型糖尿病	44,701	11.2			23,216	17.0		
胃炎	37,517	9.4			19,363	14.2		
腰部脊柱管狭窄症	32,728	8.2			20,369	15.0		
狭心症	41,109	10.3			24,894	18.3		
白内障	50,688	12.7			20,446	15.0		
眼内レンズ挿入眼	41,109	10.3			19,193	14.1		
前立腺肥大症	35,921	9			19,511	14.3		
近視性乱視	55,079	13.8			21,435	15.7		
末梢神経障害	31,930	8			18,915	13.9		
過活動膀胱	24,745	6.2			15,423	11.3		
難治性逆流性食道炎	23,149	5.8			14,313	10.5		
脂質欠乏症	21,153	5.3			10,893	8.0		
甲状腺機能低下症	20,754	5.2			10,924	8.0		
慢性心不全	17,960	4.5			11,470	8.4		
鉄欠乏性貧血	20,355	5.1			11,416	8.4		
めまい症	16,763	4.2			9,087	6.7		
慢性気管支炎	17,162	4.3			10,247	7.5		
うつ病	13,969	3.5			9,511	7.0		
関節リウマチ	13,969	3.5			7,431	5.5		
脂質欠乏性湿疹	13,969	3.5			6,925	5.1		
前立腺癌	22,351	5.6			10,282	7.5		
変形性頸椎症	11,574	2.9			6,614	4.9		
内頸動脈狭窄症	11,175	2.8			5,973	4.4		
痛風	12,373	3.1			5,443	4.0		
神経症	9,978	2.5			6,115	4.5		
うつ血性心不全	10,377	2.6			6,665	4.9		
糖尿病網膜症	13,171	3.3			8,019	5.9		
変形性股関節症	7,982	2			4,116	3.0		
口内炎	9,579	2.4			5,413	4.0		
心室期外収縮	7,583	1.9			3,860	2.8		
末梢神経炎	7,184	1.8			4,486	3.3		
糖尿病性腎症	11,175	2.8			6,322	4.6		
ヘルコバクター・ピロリ感染症	9,978	2.5			4,125	3.0		
慢性副鼻腔炎	10,776	2.7			5,846	4.3		
皮膚そう痒症	9,180	2.3			4,984	3.7		
爪白癬	8,382	2.1			3,916	2.9		
本態性高血圧症	10,776	2.7			4,934	3.6		
混合乱視	12,772	3.2			5,751	4.2		
メニエール病	8,382	2.1			4,568	3.4		
腰椎圧迫骨折	4,789	1.2			2,764	2.0		
食欲不振	5,588	1.4			3,664	2.7		
胃癌	15,167	3.8			7,174	5.3		
アトピー性皮膚炎	6,785	1.7			3,231	2.4		
慢性疼痛	5,987	1.5			4,160	3.1		
浮腫	4,390	1.1			2,924	2.1		

性 (%) の数値が正だと男性の割合が多いことを示し、負だと女性の割合が多いことを示す。表 3.6 から、男性の割合が多い傷病として、高血圧症、2 型糖尿病、痛風があげられる。このうち高血圧症は、表 3.4 より、傷病の中で最もポリファーマシーとの相関が高い。このことが、男性が女性よりポリファーマシーのリスクが高い理由として考えられる。しかしながら、高コレステロール血症や不眠症、骨粗鬆症などの傷病は女性の割合が多く、ポリファーマシーとの相関も高いため、高血圧症のみで結論づけることは難しい。

BMI については、BMI の数値が高いことで生活習慣病のリスクが高まり、高血圧、糖尿病、高尿酸血症など複数の傷病の罹患する可能性がある。それに伴い医薬品処方数も高くなったことが考えられる。

表 3.3 ロジスティック回帰分析 (健康診断)

変数	OR	p 値
性別コード (ref : men)	0.968	0.000
年齢	1.482	0.000
BMI	1.211	0.000
収縮期血圧	0.977	0.003
拡張期血圧	0.902	0.000
中性脂肪	1.009	0.018
HDL	0.959	0.000
LDL	0.726	0.000
GOT	1.032	0.000
Hba1c	1.306	0.000
e-GFR	0.850	0.000
喫煙コード (ref : no)	1.009	0.017
処方日数平均	0.692	0.000

Hba1c については、Hba1c の数値が高いと糖尿病予備軍、糖尿病と診断される。糖尿病は、神経障害、腎症などの合併症に罹患するリスクがあることから、医薬品処方数が多くなる傾向が高いと言える。e-GFR については、数値が低いことで腎不全などの傷病に罹患するリスクが高まるので、ポリファーマシーと負の相関があることは正しい。

処方日数平均とポリファーマシーには負の相関がある。医薬品の処方が短期間である理由として、急性な症状への対応、検査や再検討が必要な場合、副作用や耐性のリスクが主にあげられる。まず、一時的な症状や急性の状態に対して、医師は短期間の薬物処方を行うことがある。例えば、感染症や急性の疼痛に対する治療がその一例と考えられる。また、医師は患者の状態を定期的に評価し、必要に応じて治療計画を見直す。検査結果や症状の変化に基づいて、医師は処方を調整することがある。副作用や耐性のリスクに対して、長期間使用することで副作用や薬物耐性のリスクがある場合がある。医師はこれらのリスクを考慮して、適切な期間での使用を検討する。これらの理由が処方日数平均とポリファーマシーの関係に影響を与えている可能性がある。

3.6.2 傷病データ

表 3.4 から、高血圧症、高脂血症、不眠症などの生活習慣病はほとんどのオッズ比が 1.2 以上と高く、ポリファーマシーと有意な関連が見られている。これは、先行研究 [2] と同様の結果を示している。特に高血圧症 ($OR = 1.521$) に関して、利尿薬やカルシウム拮抗剤など多くの医薬品が処方されるため、高血圧症のみを罹患している患者でもポリファーマシーである可能性は高いと考えることができる。

3.6.3 医薬品データ

年齢、BMI は p 値が 0.05 を超えているため、ポリファーマシーと有意な関連はないと言える。これは、医薬品変数が全体的にポリファーマシーと強く関連していることが影響を与えていると考えられる。医薬品属性に関して、カルシウム拮抗剤や、糖尿病用剤など生活習慣病に関連する医薬品は特にオッズ比が高いが、そのほか

の医薬品についてもポリファーマシーとの有意な関連が強く見られる。目的変数がポリファーマシー (医薬品処方数 ≥ 5) であるので、直接影響を受けていることが理由に考えられる。

先行研究 [2] と本研究を比較した結果を表 3.7 に、医薬品対応表を表 3.8 に示す。比較した結果、数値の大きさは多少異なるが、ほとんどの属性は先行研究と同様な結果を示している。糖尿病用剤、精神抑制剤については両者ともに他の属性よりも高い数値を示している。鎮痛剤、利尿薬は対照的に、先行研究ではそれぞれ、6.62, 5.78 と他の属性よりも高い数値を示していたのに対して、本研究では、1.401, 1.6 であり、比較的低い。これらの原因として以下のことが考えられる。先行研究 [2] では東京都に在住する人々のデータのみを扱っていたが、本研究では日本全国を対象としてデータセットを作成した。地域によって医療機関数、生活習慣、医療に関する政策などが異なるため、地域によって罹りやすい傷病の違いがある。このことが本研究の結果に影響を与えたのではないかと考える。匿名加工データからは患者の住所、来院した医療機関に関する情報を得ることはできないため詳しい調査を行うことは制度上できない。

表 3.4 ロジスティック回帰分析 (傷病)

変数	OR	p 値
性別コード (ref: men)	0.845	0.000
年齢	1.159	0.000
BMI	1.197	0.000
高血圧症 (ref: no)	1.521	0.000
高脂血症 (ref: no)	1.211	0.000
高コレステロール血症 (ref: no)	1.234	0.000
不眠症 (ref: no)	1.332	0.000
慢性胃炎 (ref: no)	1.18	0.000
腰痛症 (ref: no)	1.09	0.000
アレルギー性鼻炎 (ref: no)	1.127	0.000
便秘症 (ref: no)	1.327	0.000
骨粗鬆症 (ref: no)	1.284	0.000
逆流性食道炎 (ref: no)	1.234	0.000
歯周炎 (ref: no)	0.978	0.000
脂質異常症 (ref: no)	1.156	0.000
難治性逆流性食道炎 (ref: no)	1.301	0.000
胃潰瘍 (ref: no)	1.099	0.000
2型糖尿病 (ref: no)	1.173	0.000
胃炎 (ref: no)	1.146	0.000
腰部脊柱管狭窄症 (ref: no)	1.225	0.000
狭心症 (ref: no)	1.186	0.000
眼内レンズ挿入眼 (ref: no)	1.021	0.000
前立腺肥大症 (ref: no)	1.19	0.000
末梢神経障害 (ref: no)	1.175	0.000
過活動膀胱 (ref: no)	1.169	0.000
難治性逆流性食道炎 (ref: no)	1.187	0.000
皮脂欠乏症 (ref: no)	1.059	0.000
甲状腺機能低下症 (ref: no)	1.076	0.000
慢性心不全 (ref: no)	1.135	0.000
鉄欠乏性貧血 (ref: no)	1.057	0.000
めまい症 (ref: no)	1.06	0.000
慢性気管支炎 (ref: no)	1.107	0.000
うつ病 (ref: no)	1.24	0.000
関節リウマチ (ref: no)	1.083	0.000
皮脂欠乏性湿疹 (ref: no)	1.051	0.000
前立腺癌 (ref: no)	0.968	0.000
変形性頸椎症 (ref: no)	1.065	0.000
内頸動脈狭窄症 (ref: no)	1.043	0.000
痛風 (ref: no)	1.063	0.000
神経症 (ref: no)	1.112	0.000
うっ血性心不全 (ref: no)	1.076	0.000
糖尿病網膜症 (ref: no)	1.183	0.000
変形性股関節症 (ref: no)	1.016	0.000
口内炎 (ref: no)	1.045	0.000
心室期外収縮 (ref: no)	1.013	0.001
末梢神経炎 (ref: no)	1.074	0.000
糖尿病性腎症 (ref: no)	1.105	0.000
ヘリコバクター・ピロリ感染症 (ref: no)	0.951	0.000
慢性副鼻腔炎 (ref: no)	1.097	0.000
皮膚そう痒症 (ref: no)	1.043	0.000
本態性高血圧症 (ref: no)	1.129	0.000
混合乱視 (ref: no)	1.017	0.000
メニエール病 (ref: no)	1.049	0.000
腰椎圧迫骨折 (ref: no)	1.029	0.000
食欲不振 (ref: no)	1.077	0.000
胃癌 (ref: no)	0.986	0.000
アトピー性皮膚炎 (ref: no)	1.054	0.000
慢性疼痛 (ref: no)	1.088	0.000
浮腫 (ref: no)	1.068	0.000

表 3.5 ロジスティック回帰分析 (医薬品)

変数	OR	p 値
性別コード (ref : men)	0.92	0.000
年齢	1.012	0.080
BMI	0.998	0.788
その他すべての治療用製剤 (ref : no)	2.141	0.000
その他の中枢神経作用薬剤 (ref : no)	1.675	0.000
カルシウム拮抗剤 (ref : no)	2.423	0.000
ビタミン剤 (ref : no)	2.64	0.000
ベータ遮断剤 (ref : no)	1.741	0.000
レニン-アンジオテンシン系作用薬 (ref : no)	2.494	0.000
便秘薬および腸管洗浄剤 (ref : no)	1.469	0.000
全身作用の性ホルモン剤および同効薬 (ref : no)	1.36	0.000
全身用コルチコステロイド剤 (ref : no)	1.377	0.000
全身用抗ヒスタミン剤 (ref : no)	1.968	0.000
利尿剤 (ref : no)	1.6	0.000
利胆剤および肝保護剤 (ref : no)	1.361	0.000
制酸剤, 抗鼓腸剤および抗潰瘍剤 (ref : no)	3.058	0.000
咳および感冒用製剤 (ref : no)	1.676	0.000
心臓治療薬 (ref : no)	1.435	0.000
抗てんかん剤 (ref : no)	1.485	0.000
抗喘息および慢性閉塞性肺疾患用製剤 (ref : no)	1.441	0.000
抗炎症および抗リウマチ薬 (ref : no)	2.018	0.000
抗痛風製剤 (ref : no)	1.74	0.000
抗肥満製剤を除く精神刺激剤 (ref : no)	1.596	0.000
抗血栓剤 (ref : no)	2.225	0.000
機能的胃腸障害用薬 (ref : no)	1.682	0.000
泌尿器科用剤 (ref : no)	2.149	0.000
甲状腺療法 (ref : no)	1.455	0.000
精神抑制薬 (ref : no)	2.564	0.000
糖尿病用剤 (ref : no)	3.246	0.000
脂質調整/抗アテローム製剤 (ref : no)	2.644	0.000
脳および末梢血管用剤 (ref : no)	1.318	0.000
腸疾患用製剤 (ref : no)	1.61	0.000
鎮痛剤 (ref : no)	1.401	0.000

表 3.6 男女別の傷病罹患率 (一部抜粋)

	男性 (N=177,879)	女性 (N=221,241)	男性 (%)	女性 (%)	男性 (%) - 女性 (%)
高血圧症	118299	124055	66.5	56.1	10.4
高脂血症	47129	61141	26.5	27.6	-1.1
高コレステロール血症	42561	67938	23.9	30.7	-6.8
不眠症	28890	53332	16.2	24.1	-7.9
慢性胃炎	35432	51193	19.9	23.1	-3.2
腰痛症	32822	46783	18.5	21.1	-2.7
アレルギー性鼻炎	35390	50017	19.9	22.6	-2.7
便秘症	32460	47982	18.2	21.7	-3.4
骨粗鬆症	7013	64863	3.9	29.3	-25.4
2型糖尿病	25274	19621	14.2	8.9	5.3
胃炎	14765	22676	8.3	10.2	-1.9
腰部脊柱管狭窄症	13306	19612	7.5	8.9	-1.4
狭心症	21402	19718	12.0	8.9	3.1
白内障	19313	31374	10.9	14.2	-3.3
眼内レンズ挿入眼	16129	24965	9.1	11.3	-2.2
痛風	10932	1313	6.1	0.6	5.5

表 3.7 先行研究との比較

		先行研究 [2]		本研究	
		OR	p 値	OR	p 値
Sex (ref: men)	Women	0.79	<0.001	0.92	0.000
Age group (ref: 75-79 years)	80-89 years	1.21	<0.001		
	90-99 years	1.20	<0.001		
	>100 years	0.99	0.869		
医療機関数	2 institutions	1.38	<0.001		
	3 institutions	1.87	<0.001		
	4 institutions	2.48	<0.001		
	>5 institutions	3.67	<0.001		
医師の家庭訪問 (ref: no)	Yes	1.37	<0.001		
入院治療期間 (ref: none)	1 month	1.39	<0.001		
	2 months	1.88	<0.001		
	3 months	2.50	<0.001		
	4 months	3.71	<0.001		
自己負担率 (ref: 10%)	30%	0.91	<0.001		
抗高血圧薬 (ref: no)	Yes	4.05	<0.001	1.6 - 2.494	0.000
抗うつ薬 (ref: no)	Yes	4.56	<0.001		
抗凝固薬 (ref: no)	Yes	3.86	<0.001	2.225	0.000
抗血小板薬 (ref: no)	Yes	4.16	<0.001	2.225 or 2.018	0.000
抗精神病薬 (ref: no)	Yes	4.45	<0.001	2.564	0.000
抗糖尿病薬 (ref: no)	Yes	5.70	<0.001	3.246	0.000
痛風抑制薬 (ref: no)	Yes	3.67	<0.001	1.74	0.000
抗認知症薬 (ref: no)	Yes	2.91	<0.001		
抗骨粗鬆薬 (ref: no)	Yes	5.48	<0.001		
抗高脂血症薬 (ref: no)	Yes	2.70	<0.001	2.644	0.000
鎮痛剤 (ref: no)	Yes	6.62	<0.001	1.401	0.000
抗不安薬 or 催眠薬 (ref: no)	Yes	4.90	<0.001		
胃酸抑制薬 (ref: no)	Yes	5.36	<0.001	3.058	0.000
利尿薬 (ref: no)	Yes	5.78	<0.001	1.6	0.000
浸透圧作用下剤 (ref: no)	Yes	4.75	<0.001	1.6	0.000
抗悪性腫瘍薬 (ref: no)	Yes	4.68	<0.001		

表 3.8 医薬品対応表

先行研究 [2]	本研究
抗高血圧薬	カルシウム拮抗剤, ベータ遮断剤 レニン-アンジオテンシン系作用薬, 利尿剤
抗うつ薬	対応なし
抗凝固薬	抗血栓剤
抗血小板薬	抗血栓剤, 抗炎症およびリウマチ薬
抗精神病薬	精神抑制剤
抗糖尿病薬	糖尿病用剤
痛風抑制薬	抗痛風製剤
抗認知症薬	対応なし
抗骨粗鬆薬	対応なし
抗高脂血症薬	脂質調整/抗アテローム製剤
鎮痛剤	鎮痛剤
抗不安薬 or 催眠薬	対応なし
胃酸抑制薬	制酸剤, 制酸剤・抗鼓腸および抗潰瘍剤
利尿薬	利尿剤
浸透圧作用下剤	利尿剤
抗悪性腫瘍薬	対応なし

第4章

おわりに

本研究では、健康診断データ、傷病データ、医薬品データにおいて、60歳以上のポリファーマシーに関連する危険因子の調査を行った。結論として、60歳以上の高齢者患者に対する、ポリファーマシーに関連した危険因子をそれぞれのデータセットから明らかにすることができた。しかしながら、先行研究との比較を詳細に分析することができなかった。今後は、地域別でポリファーマシーについて調査をすることを目的とし、研究を進めていく。

参考文献

- [1] Mitsutake S, Ishizaki T, Teramoto C, Shimizu S, Ito H. "Patterns of Co-Occurrence of Chronic Disease Among Older Adults in Tokyo", Japan. *Prev Chronic Dis* 2019;16:180170.
- [2] Tatsuro Ishizaki, Seigo Mitsutake, Shota Hamada, Chie Teramoto, Sayuri Shimizu, Masahiro Akishita, Hideki Ito, "Drug prescription patterns and factors associated with polypharmacy in 1 million older adults in Tokyo", *Geriatr Gerontol Int* 2020, 20:304-311.

謝辞

本論文の執筆にあたり、多くの人に支えられ、助言を頂きました。心から感謝の意を表明いたします。特に、指導教員である菊池教授には、専念したご指導を頂いたことに深く感謝申し上げます。また、研究において貴重な意見や示唆を頂いた進藤先輩方にも感謝の意を捧げます。皆さまの協力があったこそ、充実した研究が行えたことを心から嬉しく思っています。心より感謝申し上げます。

付録 A

匿名化された健康診断と診療履歴の時系列データによる糖尿病罹患予測

A.1 はじめに

近年、機械学習や AI の発展によりビッグデータの利活用が様々な場面で盛んになっている。なかでも、健康診断データは病気の罹患を予測する有効な情報と考えられる。診断結果と傷病の相関を見るために、従来は、コホート研究が主流である。例えば、野田ら [1] は 10 万人について 8 年間の追跡調査を行い、住民検診の検査結果とその後の脳卒中等による死亡の関係を明らかにした。対象者の健康状態を追跡し、詳細に分析することは生活改善や健康施設作りに、有益な知見を得ることにつながると考える。しかし、従来のコホート研究は、考えられる要因を持つ集団（曝露群）と持たない集団（非曝露群）を追跡し、両群の疾病の罹患率または死亡率を比較する方法であり、追跡時の時間的な推移を考慮することができていないという課題があった。また、個人情報を含んだデータにはプライバシー保護のための匿名加工技術についても利用・研究されている。池上ら [3] は 20 万人分の健康診断データと 28 万人分のレセプトデータ、32 万人分の適用データを使用し、従来のコホート研究と比較することによって匿名加工情報の有用性を示した。また、伊藤ら [2] はレセプトデータや健康診断データから得られる個人の身体的特徴や問診表への回答がどの程度一意であるのか調査し、データから個人が識別されるリスクを評価した。しかし、これらの研究では、各個人の診療履歴を考慮していない。かつ、身体的特徴量の時系列変化は個人を識別する際に、有効な情報であると考えられる。そこで本研究では、あるヘルスケア企業が取得し、匿名加工情報とした約 230 万人分の健康診断データと約 580 万人分の傷病レセプトデータ（レセプトデータ、約 900 万人分の基本データを使用し、診療履歴を考慮していないデータと診療履歴を考慮したデータを比較し、変化を確認した。また、診療履歴を考慮した際の一意率への影響を考察する。本研究の概要を図 A.1 に示す。該当のデータセットについて、データの特徴量、分布などを調べ機械学習アルゴリズムを用いて、ある個人が 3 年後までに糖尿病に罹患しているかどうかを予測するモデルを作成する。このオリジナルのデータから作成したモデルをモデル M1 と表す。オリジナルデータに対して時間的な変化を調査しその推移をオリジナルデータに付加したものに対してモデルを作成する。このモデルをモデル M2 と表す。その後、M1 と M2 を評価し、比較する。時間的な推移の付加方法とモデルの作成、評価方法はそれぞれ節 2.4 と節 3.3 に示す。また、オリジナルデータに対してレセプトデータを突合し、ロジスティック回帰を行うことで健康診断データの種類の特徴量から糖尿病に関する統計的に有意な因子を抽出する。

（結果）

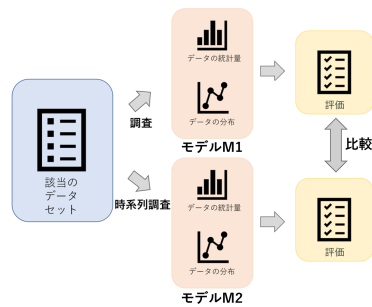


図 A.1 研究概要

- ・特徴量の変化の推移については、10名中7名のデータから、中性脂肪の増加がみられ、腹囲実測の値についても10名中7名の増加がみられた。
- ・p値に関しては、有意か判断する基準値の0.5を下回る特徴量は、両者全体的に少ないが、比較的付加した方が少し多いことが分かった。
- ・腹囲実測のオッズ比の値が元データが1.037なのに対し、付加後の数値は1.106と増加している
- ・罹患予測モデルの精度にてオリジナルデータと比較して診療履歴を考慮したデータの方が平均0.229大きくなった。
- ・診療履歴が大きくなればなるほど一意率は大きくなっていった。

本研究の新規性を次に示す。

- (1) 従来の匿名加工健康診断データと時間変化を考慮したデータとの違いを分析したこと。
- (2) 健康診断データに対して時間変化を考慮した特徴量の計算方法を提案したこと。
- (3) 診療履歴を考慮した一意率を確認したこと。

付録 B

データ

B.1 概要

本研究ではあるヘルスケア企業が取得して匿名加工した健康診断データ、基本データ、傷病レセプトデータを使用する。健康診断データは、被験者 2,345,128 名の体重や身長などの身体的特徴 53 属性と問診結果 50 属性の計 103 属性の 2014 年 4 月から 2021 年 9 月までの健康診断結果から成る。傷病レセプトデータは、各被験者が診断された傷病の記録である。また、基本データには被験者の生年月日や性別を示す。

B.2 健康診断データの前処理について

健康診断データには、分析の障害になる欠損値を含むレコードや相関が高い冗長なデータが含まれている。そこで、次の前処理を行なった。ここで、問診結果は十分な睡眠が取れているか否かなど、生活習慣に関する質問に対する患者の結果の事を差す。

1. 複数のデータの取得方法がある特徴量（Ldl 可視, Ldl 紫外など）を平均値で統合（27 特徴量）。
2. 欠損値レコードの多い 39 特徴量を削除。
3. 多重共線性をなくすために、相関係数が 0.7 以上ある 2 変数の一方を削除（5 特徴量）。
4. 欠損値を含むレコード（行）の削除。
5. カテゴリカル変数をダミー変数に変更。

処理前後の健康診断データの統計量を表 B.1 に示す。

表 B.1 健康診断データの詳細

	対象年数	被験者数	身体的特徴量数	問診結果数	特徴量数	レコード数	欠損セル数
処理前	7	2,345,128	55	50	105	7,028,931	439,127,300
処理後	7	172,819	11	17	28	1,858,163	0

本分析では、表 B.1 に示すデータの 2014 年から 2020 年の健康診断データを使用し、2017 年のデータを基準とし、2014 年から 2017 年のデータを診療履歴とする。2018 年から 2020 年の間のレセプトデータを用いて、糖尿病に罹患しているか否かを用いて 3 年後までの罹患を予測するモデルを作成する。また、健康診断データに含まれる被験者 id と 受信日、保健指導レベルコードは削除する。

B.3 個人データを用いた特徴量変化の調査

B.1 に可視化したグラフの1つの例を示す。10名の各特徴量の推移を調査した。すると、hba1cの値が6.2以下から6.5以上に変わった年に様々な特徴量の変化が見られた。hba1cは、糖尿病に罹患しているかしていないかを判定するひとつの特徴量であり、規準となる平常値は6.2%以下で、罹患と判断する基準値は6.5%以上である。まず量的変数について述べていく。例えば、10名中7名のデータから、中性脂肪の値が1年で最大90 mg/dlの増加がみられ、他の6名も大きな増加がみられた。また、bmiは、10名中3名の増加が確認され、うち一人は、標準体重の規準範囲を超え、肥満と分類される25以上に増加しているのを確認した。他にも血清クレアチニンの値が10名中6名、約1.0 mg/dlの減少がみられ、腹囲実測値に関しては、10名中7名が5から10 cmほど増加している。hdl, ldl, gptについてもいくつか変化が大きいデータがみられた。質的変数に関しては、10名中5名が歩行 or 身体活動コードか、30分以上運動コードのどちらかについて、日常の運動の頻度が少なくなるという変化がみられた。さらに、生活習慣の改善の意識については、10名中4名がちょうど糖尿病に罹患した年から意識が低くなっていることが確認された。ある個人に関しては糖尿病に罹患したとに限り、改善するつもりはないと示し、それまでの年では、既に改善に取り組んでいるという回答をしていた。他にも歩行速度が遅くなるというデータが確認された。

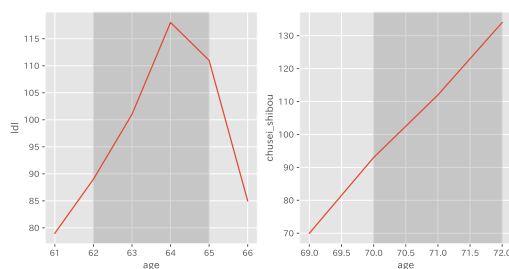


図 B.1 例

B.4 時系列情報

本節では、2.2 節の健康診断データに対して時系列情報の付加をする。ある個人の 2017 年のデータを取り出し、その個人の診療履歴を集計し、以下の 2 つの方法を用いて時系列情報をデータに付加する。

1.tilt アルゴリズム

平均変化率はある x 年の i 番目の個人の健康診断データのある特徴量 $y_i(x)$ とその直後の $x + 1$ の健康診断データについて特徴量 $y_i(x + 1)$ の勾配 $\delta y_i = y_i(x + 1) - y_i(x)$ の n 人の平均である。平均変化率積は同様にして積をとったものである。図 B.2 に tilt アルゴリズムの概略を示す。

2.Linear アルゴリズム

個人 i の特徴量 y に関して、説明変数を x 年、目的変数を y として最小二乗法を用いて線形回帰を行い、回帰係数 a を時系列情報とする。基準年のデータに付加する。図 B.3 に Linear アルゴリズムの概略を示す。

特徴量 y の線形単回帰モデルを

$$y = ax + c$$

とする。ここで説明変数 (x) 年, a は回帰係数, c は定数項である。本稿では、この 2 つの方法で定めた特徴量を追加して、傷病予測モデルを作成し、オリジナルデータとの比較をする。

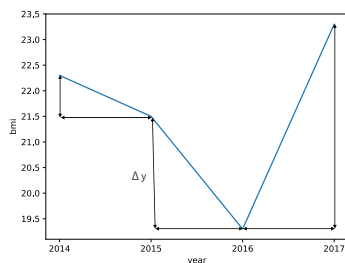


図 B.2 tilt アルゴリズムの概要

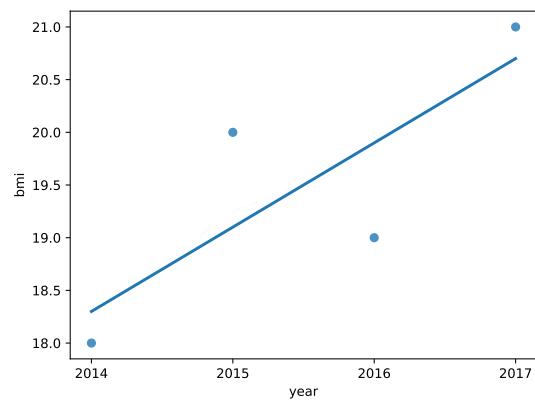


図 B.3 Linear アルゴリズムの概要

付録 C

分析

C.1 分析方法

本稿では、次の分析を行う。

1. ロジスティック回帰をし、時系列情報を考慮したデータの有無が診療結果の統計的有用性にどのように影響するかを調査する。
2. 3年後までに糖尿病罹患を予測するモデルを作成し、その精度を用いて評価する。
3. 診療履歴を考慮した一意率の調査する。

C.1.1 健康診断と傷病の関係

3年後までの罹患を目的変数、健康診断結果を説明変数として、ロジスティック回帰を用いて分析する。ある被験者 i の3年以内の傷病罹患確率 p_i を両者ともに

$$p_i = \frac{1}{1 + e^{-Z_i}}$$

で表す。ここで、 Z_i は健康診断データから得られる $M = 29(39)$ 種類の説明変数

$$Z_i = \alpha + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_M z_M$$

で定められる。

ある特徴量 z_1 について、他の変数の影響を調整したオッズ比 (adjusted Odds Ratio) は、

$$OR_1 = e^{\beta_1}$$

で与えられる。この分析では、罹患数が十分に少ない時、オッズ比と相対リスク (Relative Risk) が等しいことを利用して、説明変数 z_1 による罹患影響をオッズ比 $OR = Pr(\text{罹患} | z_1 = 1) / Pr(\text{非罹患} | z_1 = 1)$ で与える。また、ロジスティック回帰の結果から得られる各説明変数の偏回帰係数に対する標準誤差、Z 値、p 値を分析に用いる。ここで、ある説明変数を m とすると、z 値は、

$$z \text{ 値}_m = \frac{\text{偏回帰変数}_m}{\text{標準誤差}_m}$$

で表す。また、p 値は有意差 5%未満であれば、その説明変数 m は目的に対して影響を与えていると判断する。

C.2 罹患予測モデル

健康診断データから3年後までの糖尿病の罹患予測モデルを作成する。学習時には罹患者と同数の非罹患者レコードをランダムサンプリングして用いる。学習アルゴリズムには Multi-layer Perceptron classifier(MLPclassifier), Support Vector Machine(SVM) を使用する。各モデルの評価は、再現率と適合率の調和平均である F 値を利用する。モデルは python の scikit-learn を用いて実装し、MLPclassifier のハイパーパラメータにはデフォルト、SVM のハイパーパラメータについてはグリッドサーチを使用する。2017 年を基準の年として、2020 年までに罹患しているかどうか予測する。その際のデータセットを表 C.1 に示す。罹患者と非罹患者について被験者数が大きく異なるので非罹患者から同じ数の非罹患者をサンプリングしてデータの予測を行う。また、サンプリングによる分布の違いを考慮するために、モデル作成を 1000 回行い統計量や分布を比較検証する。表 C.2 にオリジナルデータ、表 C.3 に tilt アルゴリズムを適用したデータ、表 C.4 に Linear アルゴリズムを適用したデータの一例を示す。

表 C.1 罹患予測するのに用いるデータセット

	被験者数	属性数	問診結果数
3年以内に糖尿病に罹患する	209	13	17
3年以内に糖尿病に罹患しない	209	13	17

C.3 一意率について

診療履歴の数に対する安全性への影響を [2] における一意率を用いて観察する。一意率は、すべての個人の集合における、基準年から length 年間における一意な特徴量ベクトルを持つ個人の割合で定める。特徴量ベクトルとは、ある個人 i に対してある特徴量ベクトル x を長さ n の診療履歴のある特徴量 $x_1, x_2, x_3, \dots, x_n$ を用いて次のように定義する。

$$\mathbf{x} = (x_1, x_2, x_3, \dots, x_n)$$

表 C.2 オリジナルデータの一例

属性	値
sex_code	1
bmi	21.5
fukui	78.7
systolic_blood_pressure	119
diastolic_blood_pressure	56
chusei_shibou	90
hdl	70
ldl	142
got	20
gpt	22
gamma_gt	18
kessei_cr	0.75
taijuu_henka_20sai_code	2
undou_shuukan_30pun_code	2
hokou_or_shintai_katsudou_code	2
hokou_sokudo_code	2
tabekata1_hayagui_code	2
tabekata2_shuushinmae_code	1
shokushuukan_code	2
inshu_code	1
suimin_code	1
kitsuen_code	2
seikatsu_shuukan_kaizen_code	3
fukuyaku1_ketsuatsu_code	2
fukuyaku3_shishitsu_code	2
kiourek1_noukekkan_code	2
kiourek2_shinkekkan_code	2
kiourek3_zinfuzen_code	2
hinketsu_code	2
old	36

表 C.3 tilt アルゴリズムを適用したデータの一例

属性	値
sex_code	1
bmi	21.5
fukui	78.7
systolic_blood_pressure	119
diastolic_blood_pressure	56
chusei_shibou	90
hdl	70
ldl	142
got	20
gpt	22
gamma_gt	18
kessei_cr	0.75
taijuu_henka_20sai_code	2
undou_shuukan_30pun_code	2
hokou_or_shintai_katsudou_code	2
hokou_sokudo_code	2
tabekata1_hayagui_code	2
tabekata2_shuushinmae_code	1
shokushuukan_code	2
inshu_code	1
suimin_code	1
kitsuen_code	2
seikatsu_shuukan_kaizen_code	3
fukuyaku1_ketsuatsu_code	2
fukuyaku3_shishitsu_code	2
kioueki1_noukekkan_code	2
kioueki2_shinkekkan_code	2
kioueki3_zinfuzen_code	2
hinketsu_code	2
old	36
bmi_tilt_ave	0.0049
bmi_tilt_product	0.0049
chusei_shibou_tilt_ave	0.087
chusei_shibou_tilt_product	0.087
diastolic_blood_pressure_tilt_ave	-0.021
diastolic_blood_pressure_tilt_product	-0.021
systolic_blood_pressure_tilt_ave	0.03
systolic_blood_pressure_tilt_product	0.03
hdl_tilt_ave	0.055
hdl_tilt_product	0.055
ldl_tilt_ave	0.15
ldl_tilt_product	0.15
got_tilt_ave	-0.0055
got_tilt_product	-0.0055
gpt_tilt_ave	0.0055
gpt_tilt_product	0.0055
gamma_gt_tilt_ave	0.011
gamma_gt_tilt_product	0.011
kessei_cr_tilt_ave	-0.00014
kessei_cr_tilt_product	-0.00014

表 C.4 Linear アルゴリズムを適用したデータの一例

属性	値
sex_code	1
bmi	21.5
fukui	78.7
systolic_blood_pressure	119
diastolic_blood_pressure	56
chusei_shibou	90
hdl	70
ldl	142
got	20
gpt	22
gamma_gt	18
kessei_cr	0.75
taijuu_henka_20sai_code	2
undou_shuukan_30pun_code	2
hokou_or_shintai_katsudou_code	2
hokou_sokudo_code	2
tabekata1_hayagui_code	2
tabekata2_shuushinmae_code	1
shokushuukan_code	2
inshu_code	1
suimin_code	1
kitsuen_code	2
seikatsu_shuukan_kaizen_code	3
fukuyaku1_ketsuatsu_code	2
fukuyaku3_shishitsu_code	2
kioueki1_noukekkan_code	2
kioueki2_shinkekkan_code	2
kioueki3_zinfuzen_code	2
hinketsu_code	2
old	36
bmi_coef	1.8
chusei_shibou_coef	32
diastolic_blood_pressure_coef	-8
systolic_blood_pressure_coef	11
hdl_coef	20
ldl_coef	54
got_coef	-2
gpt_coef	2
gamma_gt_coef	4
kessei_cr_coef	-0.05

付録 D

分析結果

D.1 ロジスティック回帰

表 D.1 にロジスティック回帰を 1000 回行った結果の平均を示す。両者の結果に関して、全体的にあまり顕著な変化は起きなかった。ただ、腹囲実測のオッズ比の値が元データが 1.037 なのに対し、付加後の数値は 1.106 と増加している。また、拡張期血圧の付加変数のオッズ比が 0.957 と負の傾向を示している。他の付加変数のオッズ比に関しては、1.0 付近の数値を示している。他には、服薬 1 血圧コード、服薬 3 脂質コード、年齢（両者ともに、正または負に大きな傾向を示している。p 値に関しては、有意か判断する基準値の 0.5 を下回る特徴量は、両者全体的に少ないが、比較的付加した方が少し多いといえる。さらに、先ほどオッズ比が大きいと示した特徴量のうち、服薬 1 血圧コード、年齢の二つに関しては、p 値が限りなく低いため、これらの特徴量は有意であることがわかる。

D.2 ニューラルネットワークによるモデル

表 D.2 に 1000 回, 各アルゴリズムを用いて付加した特徴量を加えて作成した際の F 値の統計量を示す. 図 D.1 には F 値の分布を示す. 最も F 値の平均値, 中央値, 最小値, 最大値が高いのは Linear アルゴリズムを適用したデータであり, その次に, tilt アルゴリズムを適用したデータである. それに加え, 標準偏差も最も小さいのは Linear アルゴリズムを適用したデータであり, その次に, tilt アルゴリズムを適用したデータである.

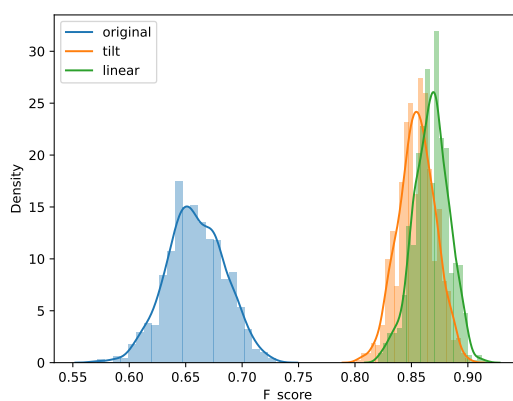


図 D.1 ニューラルネットワークにおける F 値の分布

D.3 サポートベクトルマシンによるモデル

表 D.3 に 1000 回, 各アルゴリズムを用いて付加した特徴を加えて作成した際の F 値の統計量を示す. 図 D.2 には F 値の分布を示す. 最も F 値の平均値, 中央値, 最小値, 最大値が高いのは Linear アルゴリズムを適用したデータであり, その次に, tilt アルゴリズムを適用したデータである. それに加え, 標準偏差も最も小さいのは Linear アルゴリズムを適用したデータであり, その次に, tilt アルゴリズムを適用したデータである.

D.4 診療履歴を考慮した一意率

表 D.4 に履歴期間 length に対応した一意率の平均を示す. 各属性は診療履歴の個数を表している. 診療履歴が長くなると, 一意率が単調に増加していく. また, bmi や血圧などの身体的特徴量に関しては length が 5 以上になると全ての身体的特徴量で 1.0 となり, 問診票などのカテゴリカル変数に関しては一意率が 1.0 となるようなものではなく, length が 2 のときは全てのカテゴリカル特徴量で一意率が 0 となった.

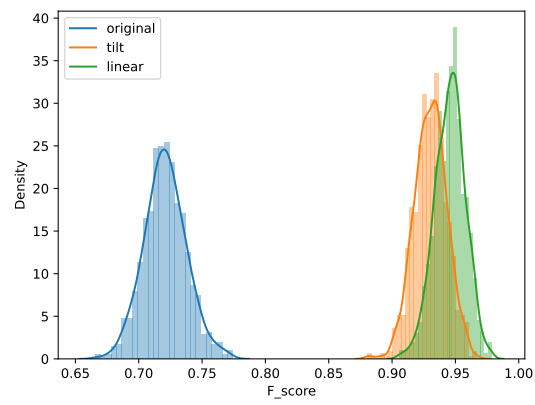


図 D.2 SVM における F 値の分布

表 D.1 ロジスティック回帰結果

	odds_ratio		std_err		zvalue		pvalue	
	original	changed	original	changed	original	changed	original	changed
sex_code	1.003	1.006	0.043	0.047	0.058	0.125	0.746	0.708
bmi	1.006	0.962	0.089	0.100	0.062	-0.385	0.711	0.643
fukui	1.037	1.106	0.091	0.100	0.393	0.988	0.645	0.362
diastolic_blood_pressure	1.040	1.056	0.060	0.076	0.658	0.719	0.526	0.478
systolic_blood_pressure	0.998	0.965	0.060	0.074	-0.047	-0.498	0.728	0.560
chusei_shibou	1.033	1.025	0.041	0.055	0.773	0.449	0.465	0.614
hdl	0.973	0.975	0.044	0.051	-0.616	-0.493	0.548	0.574
ldl	0.986	0.993	0.037	0.042	-0.395	-0.185	0.639	0.676
got	1.008	1.009	0.039	0.051	0.195	0.179	0.747	0.697
gamma_gt	1.033	1.024	0.042	0.059	0.764	0.395	0.471	0.633
kessei_cr	1.004	0.989	0.043	0.051	0.083	-0.213	0.788	0.732
taijuu_henka_20sai_code	0.981	0.996	0.041	0.046	-0.463	-0.099	0.616	0.698
undou_shuukan_30pun_code	1.012	1.006	0.039	0.043	0.312	0.131	0.664	0.723
hokou_or_shintai_katsudou_code	1.006	0.994	0.038	0.041	0.144	-0.139	0.720	0.713
hokou_sokudo_code	1.017	1.013	0.035	0.039	0.476	0.342	0.611	0.701
tabekata1_hayagui_code	1.008	1.004	0.035	0.039	0.226	0.094	0.693	0.710
tabekata2_shuushinmae_code	1.022	1.019	0.037	0.040	0.589	0.472	0.563	0.59
shokushuukan_code	0.984	0.998	0.036	0.039	-0.454	-0.053	0.621	0.77
inshu_code	1.034	1.037	0.038	0.042	0.865	0.862	0.422	0.453
suimin_code	0.984	0.976	0.035	0.038	-0.455	-0.659	0.625	0.539
kitsuen_code	1.002	0.995	0.037	0.040	0.041	-0.142	0.726	0.688
seikatsu_shuukan_kaizen_code	1.009	1.002	0.037	0.041	0.242	0.053	0.680	0.741
fukuyaku1_ketsuatsu_code	0.899	0.906	0.040	0.045	-2.636	-2.193	0.017	0.056
fukuyaku3_shishitsu_code	0.966	0.970	0.038	0.044	-0.932	-0.697	0.393	0.517
kioueki1_noukekkan_code	0.993	0.984	0.035	0.039	-0.215	-0.419	0.695	0.636
kioueki2_shinkekkan_code	0.998	0.994	0.035	0.039	-0.054	-0.155	0.688	0.662
kioueki3_zinfuzen_code	1.016	1.013	0.021	0.025	0.386	0.339	0.528	0.524
hinketsu_code	1.000	0.987	0.037	0.040	-0.016	-0.345	0.737	0.642
age	1.116	1.139	0.039	0.043	2.787	3.036	0.011	0.005
bmi_coef		0.980		0.042		-0.492		0.621
fukui_coef		0.992		0.060		-0.129		0.698
chusei_shibou_coef		1.027		0.050		0.544		0.540
diastolic_blood_pressure_coef		0.957		0.057		-0.774		0.482
systolic_blood_pressure_coef		1.018		0.056		0.295		0.662
hdl_coef		1.026		0.043		0.586		0.566
ldl_coef		0.983		0.041		-0.421		0.631
got_coef		0.996		0.050		-0.084		0.711
gamma_gt_coef		1.006		0.054		0.127		0.655
kessei_cr_coef		1.015		0.047		0.336		0.707

表 D.2 ニューラルネットワークを用いた時の F 値の統計量

	平均	中央値	標準偏差	最小値	最大値
original	0.659	0.658	0.0256	0.572	0.730
tilt	0.855	0.854	0.0165	0.801	0.902
linear	0.867	0.868	0.0156	0.821	0.916

表 D.3 SVM を用いた時の F 値の統計量

	平均	中央値	標準偏差	最小値	最大値
original	0.721	0.721	0.0168	0.666	0.774
tilt	0.930	0.931	0.0128	0.881	0.967
linear	0.950	0.950	0.0120	0.906	0.979

表 D.4 履歴期間 length における一意率の推移

length	2	3	4	5	6	7
unique rate	0.262	0.424	0.437	0.687	0.783	0.795

付録 E

考察

E.1 診療履歴を考慮した特徴量変化について

特徴量変化の調査について、本調査では、様々な特徴量の変化が確認された。特に、中性脂肪や bmi、腹囲実測値などの肥満に関する特徴量が多く変化していることが確認された。この結果の原因として、[4]によると、肥満は、インスリンの働きを鈍くし、すると膵臓の β 細胞などがインスリンを多く作ろうと働き続け、体内のインスリン量も増加する。そして、膵臓が疲れ、機能が低下し、血糖値が上がり hba1c の数値が上がるということがこの結果の原因である。なので、肥満に関わる特徴量の変化すると糖尿病の有無にかかわるということである。また、歩行 or 身体コードや、30 分以上運動コード、生活習慣改善コードは、運動不足に関わり、肥満の原因となるため、間接的に bmi などの特徴量と同じ原因であるといえる。hdl コレステロール、ldl コレステロール、gpt に関しては、脂肪肝 (中性脂肪が肝臓に蓄積する病気) の原因となる特徴量のため [5]、肥満と同じと考える。さらに、血清クレアチニンは [6] によると、血清クレアチニンの低値と 2 型糖尿病は関連があるとされている。そのため、多くのデータから血清クレアチニンの変化が確認されたと考える。最後に歩行速度が遅くなっている件については、高血糖により歩幅が減少することが歩行速度の低下に繋がるので変化がみられたのではないかと考える。[7]

E.2 ロジスティック回帰について

ロジスティック回帰について、本研究では、ロジスティック回帰から糖尿病と各特徴量の相関を調べ、時系列の考慮の有無を比較した。まず、腹囲実測値のオッズ比の値に差があることに関しては、付加後の p 値が 0.362 なので有意な結果を示している。ただし、腹囲実測値の回帰係数のオッズ比は 0.992、p 値は 0.698 なので偶然このような結果になったのではないかと考える。腹囲実測の測定は腸管内ガスの影響を受けやすいため、数値の変動が大きいというのも考えられる。次に、拡張期血圧のオッズ比が低いことに関して、糖尿病にかかりやすい高齢者は、収縮期血圧が高くなりやすいのに対して拡張期血圧は低くなりやすいというのがオッズ比の値が低くなる理由だと考える。服薬コードに関しては、血圧、脂質の両方がいずれも糖尿病に関わる要素だといえるため大きな値が得られたのだと考える。最後に、年齢に関しては、年齢を重ねるにつれ糖尿病になりやすくなるという事を示している。

E.3 罹患予測モデルについて

罹患予測モデルについて、本稿では診療履歴を tilt アルゴリズムと Linear アルゴリズムを用いて考慮したデータセットを作成し、オリジナルデータと比較することで診療履歴を考慮したデータセットの有用性を評価した。ニューラルネットワークと SVM どちらのアルゴリズムを用いたとしても診療履歴を考慮したデータセットの方が高い精度を出すことがわかった。さらに、診療履歴を考慮したデータセットのほうが標準偏差も小さく、安定性が高かった。これは時系列を考慮したデータが特微量として有用であり、モデルが識別しやすい特微量であるということを示していると考える。次に tilt アルゴリズムと Linear アルゴリズムの比較をすると、Linear アルゴリズムの方が全ての代表値で優秀な結果を示した。これは tilt アルゴリズムがかなり不安定なアルゴリズムであるからだと考える。特に、tilt アルゴリズムの平均変化率積に関しては、特微量の変化量が 0 の履歴期間があると、それ以外の診療履歴の値に関わらず、平均変化率積の値が 0 になってしまい特微量としてはかなり不安定である。最後に、そのような不安定な tilt アルゴリズムがオリジナルデータよりも高い精度を出しているのかについて考察する。不安定とはいえど、オリジナルデータよりも詳細な情報を持っているため、高くなるのは不思議ではないと考える。また、糖尿病罹患者と糖尿非罹患者はどちらか一方は平均変化率が変化しないことが多くあり、今回の健康診断データにて平均変化率積が有効な値になったと考える。

E.4 診療履歴を考慮した一意率の推移について

一意率について、本稿では診療履歴の長さによってどのように特微量の一意率が変化するかを特微量ベクトルを定義して調査した。やはり、診療期間が大きくなればなるほど一意率は上昇していった。これは特微量ベクトルの長さが大きくなることによって特微量ベクトルが一致する確率が小さくなったからだと考える。また、身体的特微量とカテゴリカル変数を比較したときに身体的特微量の方が一意率が高くなる傾向があった。身体的特微量は連続値であり、カテゴリカル変数は離散値であり値の種類が高々 3 つしかないことからこのような傾向があったと考える。

付録 F

終わりに

本研究からは、様々な結果が得られた。まず、特徴量変化の調査については、中性脂肪や腹囲実測値など、主に肥満の原因となる特徴量の変化が多く確認された。次に、ロジスティック回帰の結果から得られる健康診断と傷病の関係については、顕著な結果はでなかったが、付加変数を加えたデータセットのほうが元データセットよりも全体的にやや有意な統計値を得ることがわかった。

作成した罹患予測モデルでは、ニューラルネットワークにてオリジナルデータと tilt アルゴリズムを適用したデータでは平均 F 値の差が 0.196 であった。またオリジナルデータと Linear アルゴリズムを適用したデータでは平均 F 値の差が 0.208 であった。また、サポートベクトルマシンにて同様に比較すると、平均 F 値の差がそれぞれ 0.209, 0.229 であった。それぞれのモデルで診療履歴を考慮したデータの方が F 値は大きくなった。

一意率については、履歴期間が大きくなると一意率が増加していくことがわかった。

今後の課題として、tilt アルゴリズムと Linear アルゴリズムは不安定なアルゴリズムであるため、より安定した診療履歴を考慮するアルゴリズムの開発と診療履歴の履歴期間が大きくなることで一意率が上昇し、匿名性が失われてしまうという問題点を解消することが挙げられる。

参考文献

- [1] 野田博之, 磯博康 西連地利己, 入江ふじこ 深澤伸子, 烏山佳則, 大田仁史, 能勢忠男, 住民健診 (基本健康診査) の結果に基づいた脳卒中・虚血性心疾患・全循環器疾患・がん・総死亡の予測, 日本公衛誌 53: 265277, 2006.
- [2] 伊藤聡志, 池上和輝, 菊池浩明匿名加工情報の応用 (1): 健康診断データとレセプトデータの分析とプライバシーリスク評価
- [3] 伊藤聡志, 池上和輝, 菊池浩明匿名加工情報の応用 (2): 各種傷病を予測する健康診断 モデル 2020
- [4] 門脇孝, 肥満症と糖尿病, 日本内科学会雑誌第 100 巻第 4 号, pp.939-944, 2011.
- [5] 奥田昌恵, et al. , 2 型糖尿病における脂肪肝-その頻度及び臨床的特徴-, 糖尿病 50 巻 8 号, pp.631-634, 2007.
- [6] 針田信子, et al. , Lower Serum Creatinine Is a New Risk Factor of Type 2 Diabetes, Diabetes Care 誌, pp.424-426, 2009.
- [7] 大関直也, 水上昌文, 糖尿病神経障害者の歩行速度特性, 理学療法科学 33(1), pp. 89-93, 2018

付録 G

分担表

作業の分担を G.1 に示す.

表 G.1 ロジスティック回帰結果

	清水	石山
データ前処理	ヘルスケアデータの整形	データの集計
時系列情報の付加	アルゴリズム作成, プログラムの実装	-
ロジスティック回帰	-	健康診断と傷病の関係を調べるプログラムの実装
特徴量変化の調査	-	該当個人データの取得, 調査
罹患推定モデル	罹患推定モデルの作成	-