

一般口演 | 医療データ解析

一般口演17 データマイニング

2021年11月20日(土) 16:30 ~ 18:00 F会場 (2号館2階224)

[3-F-3-01] 問診回答を利用した健診受診者に対する糖尿病発症日数予測モデル

*畠山 豊¹、安井 繁宏¹、兵頭 勇己¹、檜山 麻里子¹、平井 学²、奥原 義保¹（1. 高知大学医学部附属医学情報センター, 2. 公益財団法人高知県総合保健協会）

*Yutaka Hatakeyama¹, Shigehiro Yasui¹, Yuki Hyodo¹, Mariko Hiyama¹, Manabu Hirai², Yoshiyasu Okuhara¹
（1. 高知大学医学部附属医学情報センター, 2. 公益財団法人高知県総合保健協会）

キーワード：Prediction model, Health Checkup Data, HbA1c

健診受診者における将来の糖尿病発症リスク評価のため、健診受診時の問診回答及び HbA1c 値に基づく HbA1c $\geq$ 6.5 となる日数予測モデルを提案する。高知県の健診施設における 2003 年から 2019 年までの健診受診者を対象データとしてモデル構築を行う。HbA1c の閾値を 6.5 (%) と定義し、状態空間モデルにより閾値を超えるまでの残り日数推定モデルを構築した。状態モデルで残り日数を定義し、観測モデルで各残り日数時点での HbA1c 値を定義した。状態モデルにおける残り日数の時間変移は前回受診時からの経過日数で定義する。観測モデルは閾値を超えた受診者データに基づきモデル同定を行った。残り日数に加えて、年齢、性別、BMI、及び健診問診回答を入力とし、その時点における HbA1c 値を出力する指数関数モデルとして構築した。モデルパラメータ推定は最小二乗法によって行った。実際の状態推定は粒子数 5000 による粒子型フィルタによって行い、粒子数の比率を残り日数となる確率として評価した。閾値を超えた受診者の内 1755 人のデータに基づき観測モデルの同定を行い、閾値を超えた 438 人のデータと超えていない受診者 6184 人を対象に残り日数の推定を行った。3 年以内となる粒子数比率、つまり 3 年以内に閾値を超える確率を算出した。この確率に対する ROC 解析をした結果 AUC が 0.926 (95% 信頼区間：0.937-0.914) となり、カットオフ値は 0.3679 となった。閾値をこえる健診受診者の各残り日数における HbA1c 値が一定の傾向を示しており、ROC から健診受診データで構築したモデルを適用することで適切に受診者の将来発症リスクを提示可能と考える。観測モデルが想定している最大残り日数 (10 年) による算出する HbA1c 値よりも低い受診者には適用できないことが限界である。

問診回答を利用した健診受診者に対する糖尿病発症日数予測モデル

畠山 豊^{*1}、安井 繁宏^{*1}、兵頭 勇己^{*1}、檜山 麻里子^{*1}、平井 学^{*2}、奥原 義保^{*1}

^{*1} 高知大学医学部附属医学情報センター, ^{*2} 公益財団法人高知県総合保健協会^{*1}

Prediction model of the number of days for diabetes onset using health checkup questionnaire data

Yutaka Hatakeyama^{*1}, Sigehiro Yasui^{*1}, Yuki Hyohdoh^{*1}, Mariko Hiyama^{*1}, Manabu Hirai^{*2}, Yoshiyasu Okuhara^{*1}

^{*1} Center of Medical Information Science, Kochi Medical School, Kochi University,

^{*2} Kochi Health Service Association

Abstract: We constructed a prediction model to estimate the number of days until HbA1c $\geq$ 6.5 for health checkup participants in Kochi. The prediction model is defined by state space model with age, sex, BMI, and answers of medical questionnaire. The target participant data were collected from health screening facilities in Kochi between 2003 and 2019 and classified into DM and non-DM groups. The prediction experiments show that AUC for probability of event (HbA1c $\geq$ 6.5) occurring within 3 years is 0.926(95% confidence interval:0.937-0.914), which the result is calculated by constant trend in HbA1c levels for each remaining day. The constructed model is useful to quantitatively present the risk of developing diabetes in health checkup participants.

Keywords: Prediction model, Health Checkup Data, HbA1c.

1. はじめに

生活習慣病の予防や早期発見などを目的として健診受診が行われている。早期発見の対象疾患の一つとして糖尿病があり、一般に HbA1c 値によって評価され糖尿病の診断基準としては HbA1c が 6.5 以上と定義されている。HbA1c 値は診断基準値に近くづくにつれて急に値が上昇することが知られているため、早期発見が必要となる。また、そのような境界域の受診者には生活習慣の改善指導が望ましく、発症リスクを提示することが期待される。

本論文では健診受診者の受診時の HbA1c 値と回答した問診項目から HbA1c $\geq$ 6.5 を超えるまでの日数を推定するモデルの構築を行う。

健診受診は年 1 回の間隔で行われるため時系列変化を扱うことは困難であるため、受診日の HbA1c 値及び問診情報から日数を推定するモデルを構築する。実際に基準値を超えた受診者の日数と HbA1c 値の分布の特徴に基づきモデルパラメータを同定し、実際の受診者に適用し推定した日数の妥当性を評価することで構築した予測モデルを評価する。

2. 方法

2.1 対象データ

高知県の健診施設における 2003 年から 2019 年までの健診受診者データを対象として予測モデルの構築を行う。HbA1c の検査結果が 1 回以上存在する受診者を初期コホートとして抽出を行う。解析対象データの抽出条件として、HbA1c 値を測定した受診日に年齢、性別、BMI の記録が存在し、さらに問診項目に全て回答していることを定義する。問診項目は、(1)脂質異常の有無、(2)脳卒中の有無、(3)心臓病の有無、(4)腎疾患の有無、(5)貧血の有無、(6)たばこの有無、(7)10キロ増加の有無、(8)運動1年の有無、(9)身体活動の有無、(10)歩行速度、(11)体重増減の有無、(12)食べる速度(3 段階)、(13)就寝前夜食の有無、(14)朝食抜きの有無、(15)飲酒の頻度(3 段階)、(16)酒量(4 段階)、(17)睡眠休養の有無、(18)生活改善の意識(5 段階)、(19)保健指導の意識、(20)糖尿病の有無の 20 項目である。糖尿病の有無については無しと回答している受診者に限定する。

HbA1c の閾値を 6.5(%)と定義し、HbA1c $\geq$ 6.5 を超えた受

診者を糖尿病群とし、一度も閾値を超えていない受診者を非糖尿病群とする。糖尿病群における閾値を超えた初回日の HbA1c が 8 以上の受診者は除外した。対象データとして、糖尿病群は閾値を超えた初回日、非糖尿病群は最終受診日から 10 年間で 3 回以上受診を行っていることとし、予測対象期間は 10 年以内とした。さらに、糖尿病群の受診者データを予測モデルの学習に使用する学習データ群と予測モデルの検証に使用するテストデータ群に分割を行う。

2.1 状態空間モデルに基づく予測モデル構築

受診日から HbA1c が閾値を超える日までの残り日数を予測対象とするモデルを構築する。本論文では、状態空間モデルに基づく予測モデルとして構築を行う。予測対象受診日における健診受診回数を $t(\geq 1)$ と定義した際の状態空間モデルを以下に定義する。

状態モデル: $Days(t) = Days(t-1) + interval(t) + \omega$

観測モデル: $HbA1c(t) =$

$$f(Days(t), Questionnaire(t)) + \omega$$

閾値を超えるまでの残り日数を $Days(t)$ 、前回受診からの経過日数を $interval(t)$ とする。対象受診日の年齢、性別、BMI、及びダミー変数化した各問診回答を $Questionnaire(t)$ と記述し、残り日数 $Days(t)$ 及び受診日の状況 $Questionnaire(t)$ から受診日の $HbA1c(t)$ を推定する関数を f と記述する。正規ノイズを ω とする。

観測モデルにおける関数 f を先行研究^{1),2)}の結果に従い指数関数として定義する。

$$HbA1c(t) = \exp(a_0 * Days(t) +$$

$$\sum_{i=1}^{29} a_{i*} Questionnaire_i(t)) +$$

$$b_0 * Days(t) + \sum_{i=1}^{29} b_{i*} Questionnaire_i(t) + c$$

各モデルパラメータを糖尿病群の学習データを用いて最小二乗法によって推定を行う。さらに、AIC による変数選択を行い最適な問診回答項目を決定する。

構築した状態空間モデルの状態推定は粒子数 5000 による粒子型フィルタ³⁾によって行い、粒子数の比率を残り日数となる

確率として評価を行う。

糖尿病群及び非糖尿病群の比較はマンホイットニーの U 検定で行い、有意水準を 5%とした。実際の計算にはR(4.0.3)で行い、モデルパラメータの推定は nls 関数で行った。

3. 結果

初期コホート 238104 人に対して、抽出条件を満たす糖尿病群 2193 人(学習データ 1755 人、テストデータ 438 人)、非糖尿病群 6184 人を解析対象とした。各条件による対象データの推移を図 1 に示す。各群における受診初回時の年齢、HbA1c、BMI及び問診回答状況を表 1 に示す。睡眠及び保健指導以外の生活習慣の問診項目に有意差が示された。

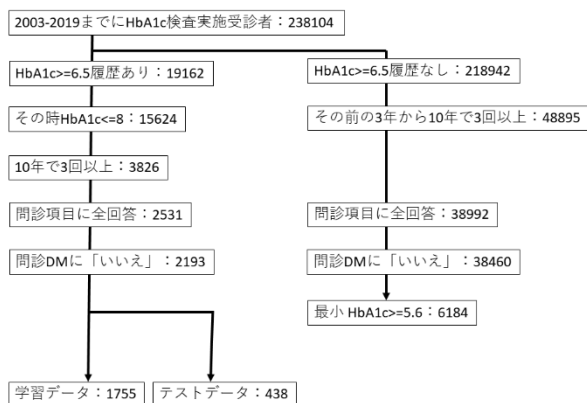


図 1 対象データの抽出条件と人数

糖尿病群(HbA1c≥6.5 履歴あり群)の学習データをモデルパラメータ同定にテストデータと非糖尿病群(HbA1c≥6.5 履歴なし群)をモデル評価に利用する

表 1 糖尿病群と非糖尿病群における初回受診時特徴量

	糖尿病群	非糖尿病群	P
n	2193	6184	
sex:female	846 (38.6%)	3766 (60.9%)	<0.001
age	59 (65-49)	62 (67-55)	<0.001
HbA1c	6.0 (6.2-5.8)	5.9 (6.0-5.8)	<0.001
BMI	25.1 (27.6-22.7)	23.1 (25.3-21.1)	<0.001
脂質異常	251 (11.4%)	718 (11.6%)	0.835
脳卒中	56 (2.6%)	124 (2%)	0.128
心臓病	81 (3.7%)	205 (3.3%)	0.402
腎疾患	2 (0.1%)	10 (0.2%)	0.453
貧血	43 (2%)	144 (2.3%)	0.316
たばこ	604 (27.5%)	1045 (16.9%)	<0.001
10キロ増加	1230 (56.1%)	2274 (36.8%)	<0.001
運動1年	661 (30.1%)	2278 (36.8%)	<0.001
身体活動	1082 (49.3%)	3463 (56%)	<0.001
歩行速度	968 (44.1%)	3044 (49.2%)	<0.001
体重増減2	552 (25.2%)	1015 (16.4%)	<0.001

食べる速度：遅い	212 (9.7%)	655 (10.6%)	0.222
食べる速度：速い	775 (35.3%)	1600 (25.9%)	<0.001
就寝前夜食	557 (25.4%)	1222 (19.8%)	<0.001
朝食抜き	334 (15.2%)	554 (9%)	<0.001
お酒：時々	621 (28.3%)	1532 (24.8%)	0.001
お酒：毎日	741 (33.8%)	1599 (25.9%)	<0.001
お酒量：1～2合未満	511 (23.3%)	1088 (17.6%)	<0.001
お酒量：2～3合未満	276 (12.6%)	436 (7.1%)	<0.001
お酒量：3合以上	180 (8.2%)	244 (3.9%)	<0.001
睡眠休養	1574 (71.8%)	4484 (72.5%)	0.508
生活改善：改善するつもりである（概ね6ヶ月以内）	959 (43.7%)	2657 (43%)	0.535
生活改善：近いうちに（概ね1ヶ月以内）改善するつもりである。	238 (10.9%)	493 (8%)	<0.001
生活改善：既に改善に取り組んでいる（6ヶ月未満）	163 (7.4%)	319 (5.2%)	<0.001
生活改善：既に改善に取り組んでいる（6ヶ月以上）	235 (10.7%)	723 (11.7%)	0.217
保健指導	1289 (58.8%)	3678 (59.5%)	0.567

観測モデルにおける関数 f においてAICによる変数選択を行った結果、脂質異常の有無、脳卒中の有無、貧血の有無、10キロ増加の有無、運動1年の有無、身体活動の有無、体重増減の有無、食べる速度、就寝前夜食の有無、朝食抜きの有無、酒量(4段階)の項目が選択された。モデルパラメータを表 2 に示す。学習データに対する関数 f によって算出された推定 HbA1c 値と実際の HbA1c 値の分布を図 2 に示す。推定値が図 2(a) となる。横軸は実際に閾値を超えた日の日何日前かの日数であり、縦軸が HbA1c 値である。年数ごとの頻度分布を図 2(b)に示す。実際の測定値と推定値の相関係数は 0.63 となった。

糖尿病群のテストデータ及び非糖尿病群に適用した予測結果の例を図 3、4に示す。図 3 は糖尿病群、図 4 は非糖尿病群の例を示している。それぞれの図は各時刻での粒子数のヒストグラムであり、その比率が残り日数の確率として評価する。つまり、推定残り日数が 3 年以内の粒子数の比率が 3 年以内に閾値を超える確率として評価する。3 年以内での閾値を超えるかどうかについてのROCを図 5 に示す。AUC は

0.926 であり 95%信頼区間は(0.937-0.914)となった。この時のカットオフ値は 0.368 であり、カットオフ値を変動させたときの感度、特異度、陽性的中率、陰性的中率を表 3 に示す。この結果から 3 年以内の確率が 0.8 の時、陽性的中率が 83.3% 陰性的中率 98.1% であることが得られた。

	a	b	c	
Days(t)	0.003983	0.00007848		5.814
BMI	0.008847	-0.001314		
脂質異常	-0.1012	0.04061		
脳卒中	0.08803	-0.04524		
貧血	0.03258	-0.0514		
10キロ増加	0.04293	0.00862		
運動1年	-0.007182	-0.01327		
身体活動	-0.006258	0.01669		
体重増減	0.1068	-0.05361		
食べる速度:遅い	0.129	-0.004917		
食べる速度:速い	-0.04565	0.02307		
就寝前夜食	0.011	0.01195		
朝食抜き	0.0942	-0.05117		
お酒量:1〜2合未満	-0.0138	-0.01526		
お酒量:2〜3合未満	-0.005637	-0.01945		
お酒量:3合以上	0.08772	-0.0445		
性別	-0.1668	0.03133		
age	-0.01393	0.005785		

$$HbA1c(t) = \exp \left(a_0 + Days(t) + \sum_{i=1}^{17} a_{1i} Questionnaire_i(t) \right) + b_0 + Days(t) + \sum_{i=1}^{17} b_{1i} Questionnaire_i(t) + c$$

表 2 推定された観測モデルパラメータ

4. 考察

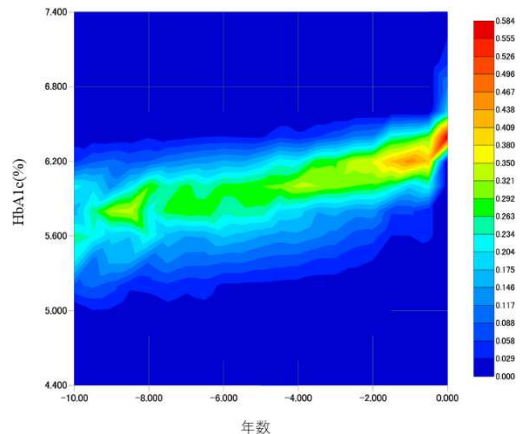
HbA1c 値が閾値をこえる健診受診者の各残り日数における HbA1c 値は閾値直前で急激な上昇を示すなどの一定の傾向を示しており、その傾向を考慮した予測モデルを適用することで適切に受診者の将来発症リスクを提示可能な結果が示されたと考える。

初回受診時の段階で糖尿病群は生活習慣が悪化していることが問診項目の回答から示されており、そのため HbA1c 値の推定関数において AIC による変数選択で生活習慣の項目が選択されたと考える。推定 HbA1c 値は相関係数で評価すると必ずしも精度が高いはない。これは閾値を超えない高 HbA1c 値で維持する受診者も存在するため、ある程度誤差が発生すると考える。しかしながら、図 2 で示すように推定 HbA1c 値を生活習慣の違いによって個人差を反映することが可能であるため健診データに対しては有効に適用されていると考える。

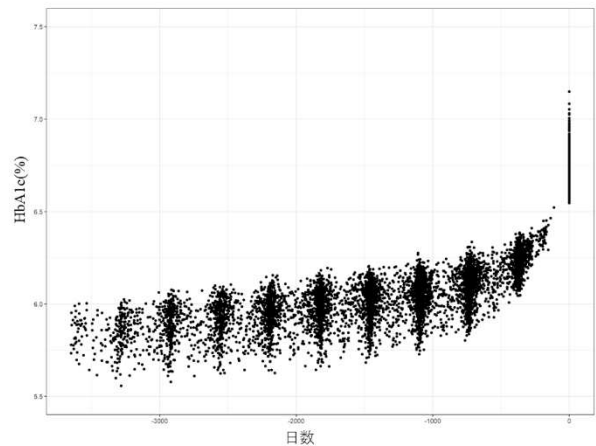
図 3 及び図 4 で示すように粒子数の比率が残り日数に対する確率として表現されている。初回受診日にはある程度 HbA1c 値が高くても数年以内に閾値を超える確率は高く推定されていない。これは学習データにおいて HbA1c 値が高いまま維持するデータが存在するため、このような結果になっている。HbA1c 値が高い値を維持しつづけると短い残り日数の確率が高くなり、値が低下すると確率が低くなる傾向となっている。これは図 2 で示された残り日数と測定された HbA1c 値の関係が反映されている。図 4 データの受診者は最終受診日まで閾値を超えていないが、最終測定日における推定残り日数は比較的短い結果となっている。これは過剰な推定結果というよりも、HbA1c 値が 6.0 あたりで長期間推移しているため、急激に HbA1c 値が上昇する可能性が常にある程度存在していることを示していると考ええる。

粒子型フィルタによって算出した 3 年以内で閾値を超えるかの確率では高い精度が示されたと考える。これは実際の HbA1c 値分布が残り日数ごとに異なっていることが分布から示され、それを利用した状態空間モデルによって予測したためと考える。陰性的中立に比べて陽性的中率が低い結果となっているが、これは予測対象となるテストデータ群が非糖尿病群に比べて受診者が少ないためと考える。しかしながら、

ROC の各カットオフ値の結果から、受診者の糖尿病発症のリスク提示は行えると考ええる。



(a)



(b)

図 2 観測モデルによる学習データ適用結果
(a)観測モデルパラメータ同定に用いた学習データの適用した HbA1c 推定結果(b)適用結果に対する 1 年ごとの推定 HbA1c 分布

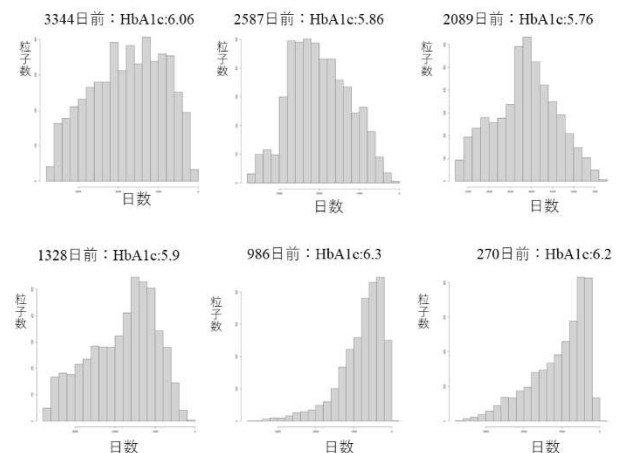


図 3 糖尿病群の 1 受診者に適用した際の粒子分布推移
HbA1c≥6.5 を初めて満たした受診日を 0 日とし、それ以前の受診データに対して構築予測モデルを適用した際の粒子分布

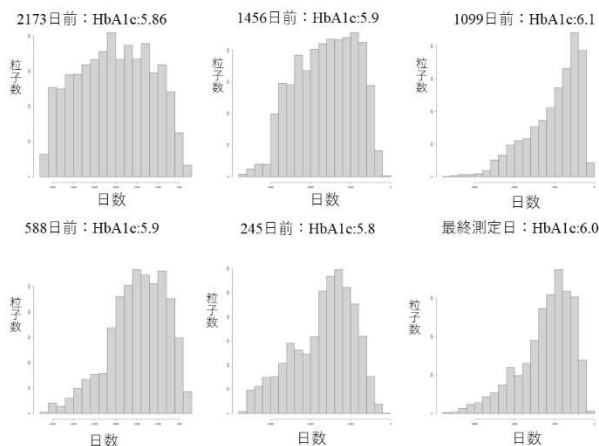


図 4 非糖尿病群の 1 受診者に適用した際の

粒子分布推移

最終受診日を 0 日とし、それ以前の受診データに対して構築
予測モデルを適用した際の粒子分布

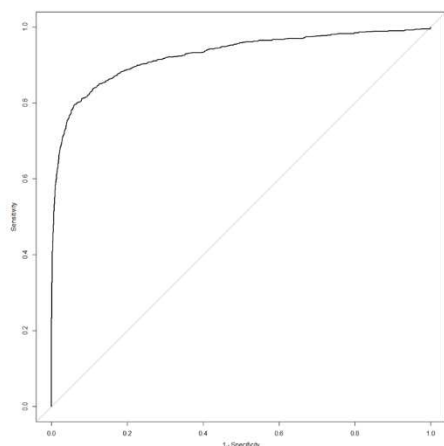


図 5 対象受診者に適用し 3 年以内での閾値を超えるか
についての ROC

推定残り日数が 3 年以内の粒子数比に対する ROC

HbA1c 値の変化自体を対象とした予測モデルについて、様々なアプローチ^{4),5)}が提案されている。これらの先行研究は HbA1c 値の測定が数カ月単位で実施されているデータを対象としている。しかしながら、本論文で対象とする健診受診データは 1 年単位でのデータであり、HbA1c 値の変化量を扱うには受診間隔が大きすぎると考える。そのため、受診時のデータ及び過去データを反映した状態モデル値による予測モデルは、健診データに対して有効だと考える。また、他の検査値を組み合わせた手法も存在しているが、必ずしも受診時に測定しているとは限らないため、欠損が多くなり予測精度の低下が起きる可能性があるため、構築モデルは HbA1c 値と問診項目のみで構築を行った。

本モデルの限界は、HbA1c 値が高くない(HbA1c<5.6)受診者には適用できないことがある。これは観測モデルの構築に利用した糖尿病群の最低値に依存している。つまり、健診受診者の学習データで 5.6 未満の受診者は 10 年以内で閾値を超えた受診者は存在しなかったことを意味する。そのため、これらの受診者が閾値を超えることは稀であるため実運用上問題ないと考える。

表 3 ROC のカットオフ値に対する感度、特異度、陽性的
中率、陰性的中率

カットオフ値	感度	特異度	陽性的中率	陰性的中率
0.1	0.994	0.031	0.028	0.995
0.2	0.967	0.389	0.043	0.998
0.3	0.907	0.742	0.09	0.997
0.368	0.85	0.874	0.159	0.995
0.4	0.815	0.909	0.202	0.994
0.5	0.715	0.968	0.385	0.992
0.6	0.582	0.989	0.597	0.988
0.7	0.449	0.996	0.74	0.985
0.8	0.308	0.998	0.833	0.981
0.9	0.182	1	0.941	0.978
1	0.016	1	1	0.973

5. おわりに

高知県の健診受診者のデータを対象として、糖尿病発症のリスク評価を行うため HbA1c 値が 6.5 を超えるまでの残り日数を推定するモデルを、実際の受診者データに基づき状態空間モデルに基づき構築した。実際の受診者データに対して適用した結果、一定期間内に閾値を超える確率を適切に推定することが可能な結果が得られた。この結果から、構築した予測モデルによって受診者の糖尿病発症のリスクを定量的に提示することが可能になった。

参考文献

- 1) Tabak AG, Jokela M, Akbaraly TN, Brunner EJ, Kivimaki M, Witte DR. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet* 373: 2215-2221. 2009.
- 2) 畠山 豊、兵頭 勇己、奥原 義保. 経過記録情報を用いた深層学習による欠損値補間した HbA1c 予測モデルの構築. *医療情報学* 40(5):231-238.2020.
- 3) Kitagawa G. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J Comp Graph Stat* 5: 1-25. 1996.
- 4) Chen-Ling Huang, Usman Iqbal, Phung-Anh Nguyen, et al.: Using hemoglobin A1C as a predicting model for time interval from pre-diabetes progressing to diabetes. *PLoS One*. 5;9(8), 2014.
- 5) Sunil B Nagaraj , Grigory Sidorenkov , Job F M van Boven , Petra Denig. Predicting short- and long-term glycated haemoglobin response after insulin initiation in patients with type 2 diabetes mellitus using machine-learning algorithms. *Diabetes Obes Metab*. 21(12):2704-2711. 2019.