ROIS-DS-JOINT
032RP2019

医療アクセス制限研究の属性共起分析による類似オープンデータ順位付けとデータサイエンス応用

神沼英里[1)], 八谷剛史[2)], 田中博[1)], 山本泰智[3)]

1) 東京医科歯科大学 医療データ科学推進室
2) ゲノムアナリティクスジャパン
3) ライフサイエンス統合データベースセンター

提案手法：医療アクセス制限研究の代替オープンデータを類似検索

■提案＝アクセス制限研究の代替オープンデータを、属性類似検索で探す

＜従来のアクセス制限研究利用＞

電子カルテデータセット

医療アクセス制限研究のデータは利用不可

（IRB研究倫理審査の承認）
通常：数カ月

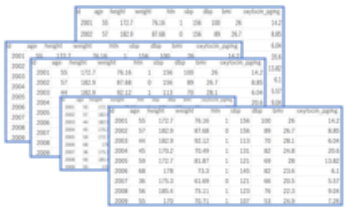
機械学習モデルの構築

アクセス制限研究の機械学習モデルへの応用可能性を検討

＜提案手法＞

①アクセス制限研究(Query)の属性キー取得

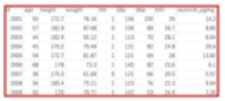
オープンデータ



②複数の公開データセットの属性キー収集

1位
2位
3位

③属性キー類似検索でオープンデータを順位付け



④高順位のオープンデータを用いて属性値予測の機械学習モデル構築

[1] 教材化

[2] 自動化

2019年度の進捗まとめ

■研究発表１件

糖尿病電子カルテを事例とした
MeSH Term注釈に基づくアクセス制限研究のオープンデータ類似検索
神沼英里、山本泰智、田中博
人工知能学会合同研究会2019 SIG-AIMED-008 2019年11月22日

■2019年度の進捗まとめ

＊2019年度：※Google Dataset Search結果からオンラインのオープンデータを
収集するWeb Scrapingプログラムを構築した。
※オープンデータランキング結果から単一データセットを選択して、
機械学習の属性予測モデルを構築、医療・創薬データサイエンス
コンソーシアムの「オープンデータ入門」教材として公開。

＊2018年度：※Google Dataset Search結果から糖尿病電子カルテのデータセット収集。
※データセット属性値を、MeSH Termのマニュアルキュレーション。
※アクセス制限研究をクエリとした代替オープンデータのランキング
プログラム作成、糖尿病電子カルテで試行実験を実施した。

[1] 教材化

アクセス制限研究の
代替オープンデータを用いた
糖尿病電子カルテ属性予測モデルの教材化

2019年度進捗[1] 教材化：
国立国際医療研究センターの糖尿病電子カルテ研修前に学ぶ、オープンデータを用いた糖尿病予測教材を作成する

National Center for Global Health and Medicine

国立国際医療研究センター

J-DREAMS(診療録直結型全国糖尿病データベース事業)を始めとした電気カルテ情報を活用した医療用ビッグデータの構築・管理・運用

Diabetes EMR database

National Center for Neurology and Psychiatry

国立精神・神経医療研究センター

①脊髄小脳変性症、筋ジストロフィー、プリオン病などの疾患データベースを対象とした統計解析。
②疾患データベース、MRI画像データ、脳波データ、髄液データなどを用いた機械学習

Analyzing Heterogeneous Medical databases

Japanese Foundation for Cancer Research

がん研究会

病理部におけるデータのデータベース作成と解析

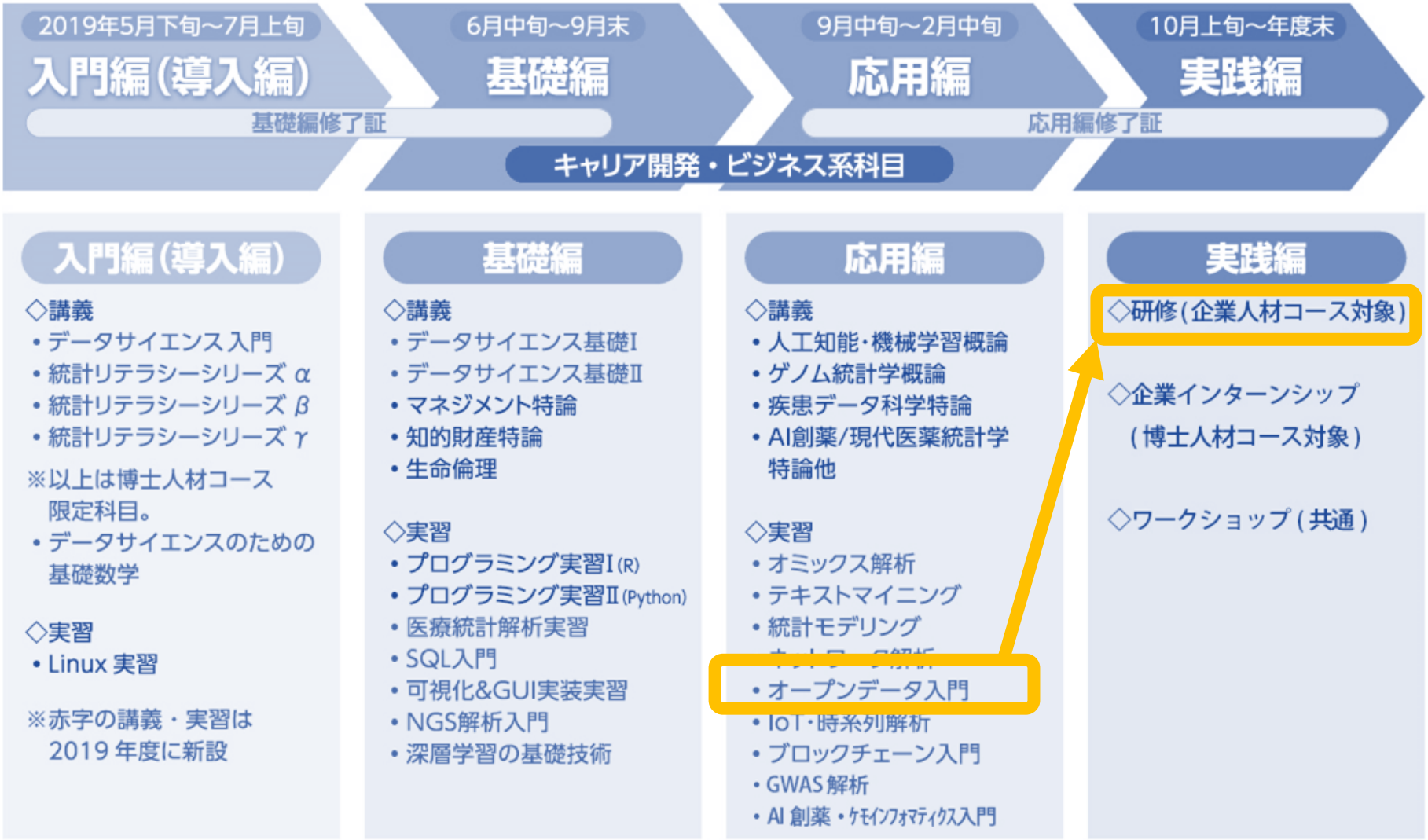
Data Analysis for Diagnostic Pathological Images

東京医科歯科大学 Tokyo Medical and Dental University

Deep Learning等によるAI創薬プログラムを使用した計算創薬演習

GOAL: 国立国際医療研究センターの「アクセス制限研究(J-DREAMS)」研修の前に、オープンデータの学習教材を用意する

2019年度進捗[1] 教材化：
「オープンデータ入門」教材として公開する



https://md-dsc.com/curriculum31.php
医療・創薬データサイエンスコンソーシアムより

アクセス制限研究の代替オープンデータ探索①
クエリとなるアクセス制限研究「J-DREAMS」

2018年度成果

Basic information
Year/month of birth
Sex
Hospital code
Laboratory data
Blood samples
Blood cell count
Total protein
Aspartate transaminase
Alanine transaminase
Gamma-glutamyl transpeptidase
Creatine kinase
Total cholesterol
High-density lipoprotein cholesterol
Low-density lipoprotein cholesterol
Triglycerides
Blood urea nitrogen
Creatinine
Potassium
Hemoglobin A1c
Glycoalbumin
1,5-Anhydroglucitol
Blood glucose
Cancer antigen 19-9
Brain natriuretic peptide
Cystatin C
Carcinoembryonic antigen
Thyroid-stimulating hormone
Free triiodothyronine
Free thyroxine
Insulin
C-peptide
Anti-glutamic acid decarboxylase antibodies
Anti-islet antigen 2 antibody
Islet cell cytoplasmic antibody
Zinc transporter 8 antibody
Anti-insulin antibody
Hepatitis B surface antibody
Hepatitis C antibody
Urine samples
Qualitative urinary test
Protein
Albumin
Creatinine
C-peptide
Prescription
All of the patient's prescription information obtained from the participating facility

Diabetol Int (2017) 8:375–382
DOI 10.1007/s13340-017-0326-y

ORIGINAL ARTICLE

Design of and rationale for the Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System (J-DREAMS)

Takehiro Sugiyama[1,2] · Kengo Miyo[3] · Tetsuro Tsujimoto[4] · Ryota Kominami[3,5] · Hiroshi Ohtsu[6] · Mitsuru Ohsugi[1,4] · Kayo Waki[7] · Takashi Noguchi[8,9] · Kazuhiko Ohe[9] · Takashi Kadowaki[10] · Masato Kasuga[11] · Kohjiro Ueki[4,12] · Hiroshi Kajio[4]

Received: 30 March 2017 / Accepted: 12 June 2017 / Published online: 27 June 2017

アクセス制限研究属性情報＝43項目

全国糖尿病患者電子カルテ
「J-DREAMS」プロジェクト

The variables collected through J-DREAMS are listed in Table 2 (the basic information, prescription history, and clinical laboratory data stored in the SS-MIX2 standardized storage) and in Supplementary Fig. 1
(the clinical information collected using the SDMT and stored in the SSMIX2 extended storage).

アクセス制限研究の代替オープンデータ探索②
オープンデータ17件の属性情報にMeSH Term割当

2018年度
成果

■ 専門家によるMeSH Term手作業注釈

A	B	C	D	E
CURATED QUERY	QUERY(ORIGINAL VAR)	VAR Category	MeSH TERM(curated)	MESH UNIQID(curated)
Birth	Year/month of birth	Basic	Term Birth	D047929
Sex	Sex	Basic	Sex	D012723
Hospital code	Hospital code	Basic	Hospitals	D006761
Blood	Blood samples	Laboratory	Blood	D001769
Blood cell count	Blood cell count	Laboratory	Blood Cell Count	D001772

MeSH TermとUnique ID

①QUERY
手作業キュレーション

②MeSH TERMのUnique IDを
手作業キュレーション

■MeSH Term
Medical Subject Headings の略語。
米国国立医学図書館 (NLM)が
提供する生命科学用語集。

オープンデータの属性Keyを、ユニークMeSH Termsに紐づけた

アクセス制限研究の代替オープンデータ探索③
J-DREAMSクエリからオープンデータセットをランキング

2018年度成果

■MeSH Term間の類似度は、Hamming距離で定義

ハミング距離

$$d_{st} = (\#(x_{sj} \neq x_{tj})/n).$$

■JDREQMSクエリに対する、17データセットのTop7ランキング結果

Rank	ID	Hamming 距離	Open Dataset Name	Number of Attributes	Number of Instances
1	6	0.657	Eisenberg et al., PLoS One, 2018. PMID: 29300770	221	92
2	16	0.714	Okamura et al., Int J Obes, 2019. PMID:29717276	31	15,464
3	10	0.800	Heier et al., PLoS One, 2018. PMID: 30359432	39	78
4	7	0.829	McCracken et al., BMJ Open, 2017. PMID:28801438	55	214
5	9	0.829	Andersson et al., PLoS One, 2015. PMID:26186716	33	63
6	2	0.857	Strack et al., Biomed Res Int 2014. PMID:24804245	50	101,766
7	17	0.857	Karakonstantis et al., Mendeley Data, 2018.	12	55

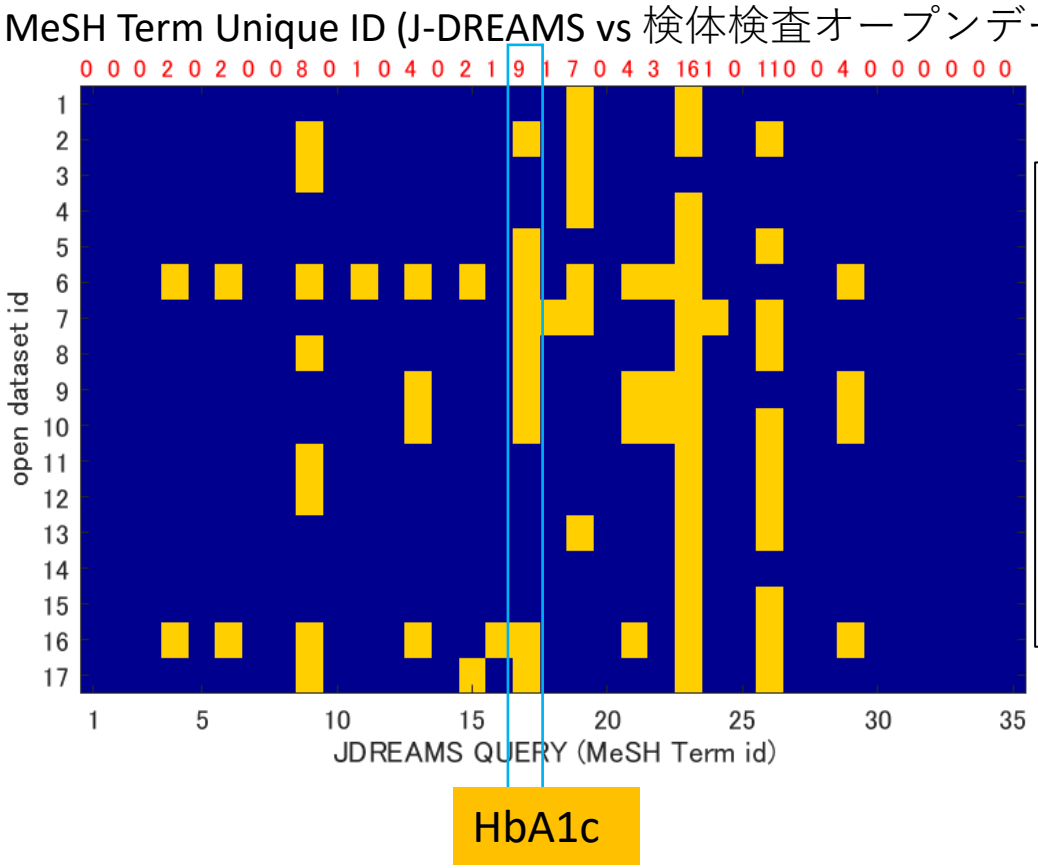
※[Number of Attribute]と [Rank] に有意な相関無し

検査値等項目数

被験者数

教材化への妥協①　複数統合を断念、属性調査へ

■糖尿病オープンデータの場合、どの属性も被覆率が低かった。
複数データセットの統合の前に、属性情報を調査して知見を収集する。



MeSH Terms
（有＝ ，無＝ ）

JDREAMSのMeSH Term 35件中、19件(54%)が17オープンデータセットに存在しない
↓
存在しても被覆率が低い
↓
「HbA1c」属性も、オープンデータ17件の被覆率は50%程度。

今年度は、複数オープンデータの統合を断念して、１オープンデータのみの教材で、糖尿病属性予測モデルに対する「属性情報を調査」する。

教材化への妥協②　被験者数が多く、属性数が少ない
単一データセットで属性調査＋教材化

■複数データセットでの共通属性の選択が容易ではないので、被験者数が多く属性数が少ない単一のデータセットを選択した。

	距離	属性数	
rank=1	0.65714	221	
rank=2	0.71429	31	
rank=3	0.80000	39	
rank=4	0.82857	55	
rank=5	0.82857	33	
rank=6	0.85714	50	被験者数
rank=7	0.85714	12	55名
rank=8	0.88571	15	1415名
rank=9	0.91429	16	
rank=10	0.91429	19	
rank=11	0.91429	17	
rank=12	0.91429	52	
rank=13	0.94286	9	
rank=14	0.94286	30	
rank=15	0.94286	9	
rank=16	0.94286	48	
rank=17	0.97143	101	

被験者1,415名の内訳
糖尿病患者＝95名、健常者＝1,320名

※被験者数多でも、糖尿病患者数は少ない

PLoS One. 2017 Sep 14;12(9):e0184840. doi: 10.1371/journal.pone.0184840. eCollection 2017. PMID:28910380

Validation of the diabetes screening tools proposed by the American Diabetes Association in an aging Chinese population.

Woo YC[1], Lee CH[1,2], Fong CHY[1], Tso AWK[1], Cheung BMY[1,2], Lam KSL[1,2].

Author information

1 Department of Medicine, The University of Hong Kong, Hong Kong, Hong Kong SAR.
2 Research Centre of Heart, Brain, Hormone and Healthy Aging, The University of Hong Kong, Hong Kong, Hong

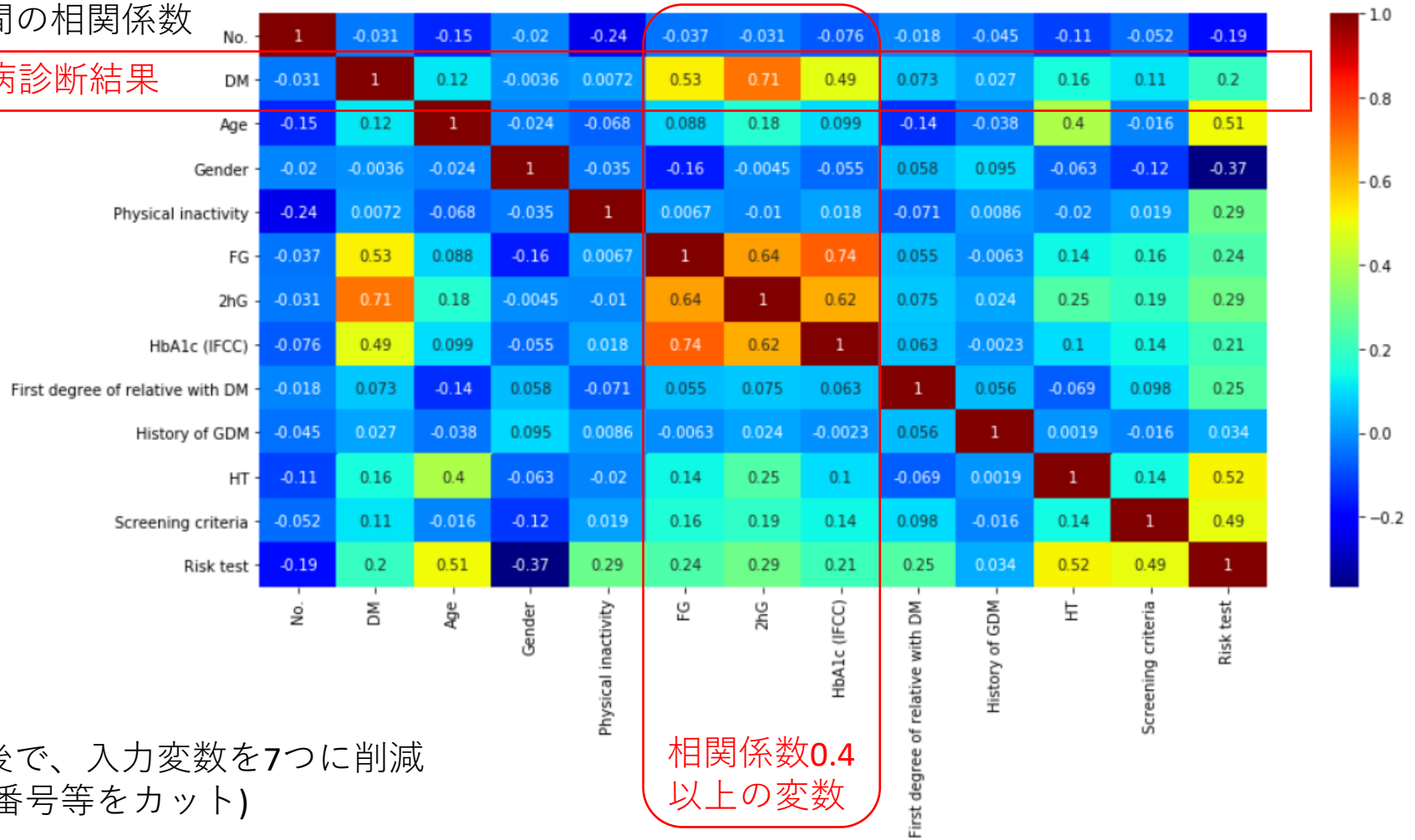
※2010～2012年に実施された高齢化集団対象の糖尿病有病率調査のデータ(CRISPS4)

属性予測モデル①属性変数間の相関係数

■オープンデータ（論文の表S1）を利用

Woo YC et el., "Validation of the diabetes screening tools proposed by the American Diabetes Association in an aging Chinese population", PLoS One. 2017 Sep 14;12(9):e0184840

■変数間の相関係数



※この後で、入力変数を7つに削減
(被検者番号等をカット)

属性予測モデル②糖尿病判定モデルの構築

■勾配ブースティング決定木法で、糖尿病の有無を判定 する

＊糖尿病(DM)か否かを判別する2値クラス分類モデル
＊７属性を入力変数に設定（元論文の主題：リスクテスト属性は削除）
＊XGBoost のPythonライブラリを使用
＊GridSearchで木構造Depth探索＝最適値： 2
＊データ分割(訓練：テスト)＝７：３
＊テスト評価結果＝Accuracy 1.00

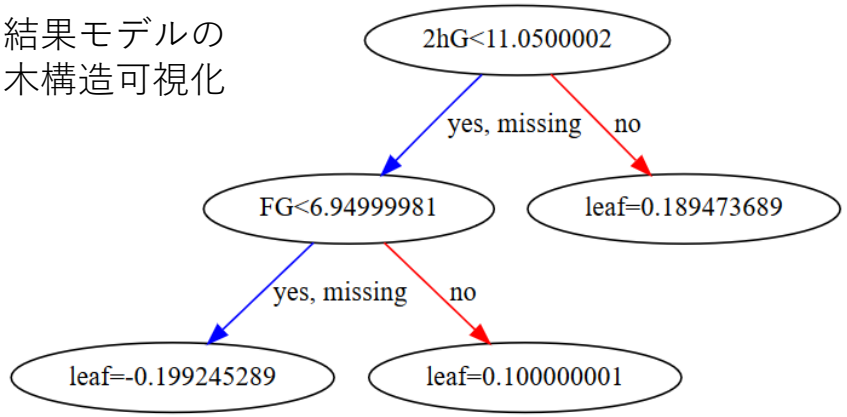
＊属性変数の重要度計算

入力属性変数	変数重要度
2hG [ブドウ糖負荷試験の2時間後血糖値]	0.9056
FG [空腹時血糖値]	0.0801
Age	0.0117
HbA1c [ヘモグロビンA1c]	0.0026
Gender	0.0000
smoking	0.0000
Physical inactivity	0.0000

属性予測モデル③糖尿病判定結果とWHO診断基準の比較

■勾配ブースティング決定木モデルによる糖尿病/健常の判定結果を考察。

結果モデルの
木構造可視化



今回の木構造分析結果は、世界保健機関（WHO）1998の診断基準**（空腹時血糖値（FG）≥7 mmol / LまたはOGTT 2時間後血糖値（2hG）≥11.1 mmol / L）**と値が近かった。

空腹時血糖値，75g糖負荷試験（OGTT）2時間値の組み合わせにより，**表1**のごとく糖尿病型，正常型，境界型に分ける．随時血糖値≧200mg/dLも糖尿病型とする[4]．

表1 空腹時血糖値および75g糖負荷試験（OGTT）2時間値の判定基準
（静脈血漿値，mg/dL，括弧内はmmol/L）

	正常域	糖尿病域
空腹時血糖値	<110（6.1）	≧126（7.0）
75g OGTT 2時間値	<140（7.8）	≧200（11.1）
75g OGTTの判定	両者を満たすものを正常型とする	いずれかを満たすものを糖尿病型とする
	正常型にも糖尿病型にも属さないものを境界型とする	

随時血糖値≧200mg/dL（≧11.1mmol/L）の場合も糖尿病型とみなす．
正常型であっても，1時間値が180mg/dL（10.0mmol/L）以上の場合は，180mg/dL未満のものに比べて糖尿病に悪化する危険が高いので，境界型に準じた取り扱い（経過観察など）が必要である．
（文献4から引用）

参考資料：糖尿病域の基準
https://minds.jcqhc.or.jp/n/med/4/med0004/G0000107/0010

※世界では、血糖値の単位はmmol/Lが採用されている。日本で単位(mg/dL)は18倍の数値のため、データ統合には注意が必要になる。

属性予測モデルは、医療・創薬データサイエンスコンソーシアムより
e-Learning教材「オープンデータ入門」として公開した。

[2] 自動化

Google Dataset Searchの結果から、
代替オープンデータを
自動取得する。

DatasetAutoScan：データセットの自動ダウンロードプログラムを開発

■Google Dataset Searchの検索結果から、figshareの表データセットを自動ダウンロードするGoogle Colabプログラム「DatasetAutoScan」を構築した。

■DatasetAutoScanの流れ

1. Google Dataset Searchのキーワード検索結果を取得（キーワード例：Diabetes）
2. 検索結果からURLのみ抽出
    - ＊GDS検索結果から、URLのみ抽出
    - ＊URLリストから「figshare.com」を含むURLのみ残す
    - ＊クレンジング処理
    - ＊figshare.comのURL数をカウント
3. URLのうち、https://figshare.com/のみ抽出
4. figshare.comのサイトからダウンロードURLを取得
    - ＊表形式データを、data_XXディレクトリでColabディスクに取得
    - ＊表形式と共に.bibデータもdata_XXディレクトリに取得
    - ＊tar+gzipでまとめて、ローカルディスクに落とす
5. Colabディスクのデータファイル(csv,xlsx等)をLocal PCにダウンロード

謝辞・参考文献など

■成果プログラムの公開

https://github.com/ekaminuma/ROIS-DS-JOINT/

■Acknowledgements

This work was supported by ROIS-DS-JOINT (032RP2019, 029RP2018).

■References

- Sugiyama T, et al., Diabetol Int, 8:375, 2017.
- J-DREAMS (http://jdreams.jp/)
- Google Dataset Search (https://toolbox.google.com/datasetsearch)
- Woo YC et el., PLoS One, 12:9, e0184840, 2017.
- Jimeno-Yepes AJ, et al., BMC Bioinformatics, 14:1471, 2013.
- MeSH Browser (https://meshb.nlm.nih.gov/search)
- Eisenberg et al., PLoS ONE, 13: e0190301, 2018.
- Okamura et al., Int J Obes (Lond), 43:139, 2019.