2C4-OS-21a-5

特定患者集団の抽出(Phenotyping)手法確立に向けた技術的課題に関する考察

The Technical Problems for Automated Phenotyping

香川 璃奈[*1] 河添 悦昌[*2] 井田 有亮[*2] 篠原 恵美子[*2] 今井 健[*1] 大江 和彦[*1, 2]
Rina Kagawa Yoshimasa Kawazoe Yusuke Ida Emiko Shinohara Takeshi Imai Kazuhiko Ohe

[*1] 東京大学大学院医学系研究科 Graduate School of Medicine, The University of Tokyo
[*2] 東京大学医学部附属病院 The University of Tokyo Hospital

To improve healthcare quality and advance clinical research, the need for automated phenotyping for patient identification is increasing. However, the phenotyping algorithms have not been matured. We performed the binary annotation for electronic health records of whether a patient is type 2 diabetes mellitus. The diagnoses in electronic health records were limited in terms of accuracy and completeness for automated phenotyping. Physicians checked electronic health records, but some patients were difficult to determine whether they were type 2 diabetes mellitus and therefore made detailed annotation guideline. We assessed the information in electronic health records that was required for determining whether they were type 2 diabetes mellitus and analytic technique in order to get the information.

1. 背景

本邦では電子カルテが普及してきており，蓄積された電子カルテ情報を用いた臨床研究などの需要が高まっている．ここで患者の基本となる情報は患者の病名の情報である．しかし，カルテに記載される病名，構造化データとして登録される病名コードは確実性および網羅性が低いと知られている[McCormick 2014]．そこで，研究対象とする特定の病名や状態を持つ患者集団を自動的に抽出する技術(phenotyping)の需要が高まり，開発が盛んになっている[Jie 2015]．

Phenotyping 手法を開発する際には gold standard となる病名(状態)をカルテに正解付けして決定しておく必要があるが，先述の通り，カルテに書かれている病名，構造化データとして登録されている病名コードは確実性および網羅性が低い．そこで，研究者あるいは専門家である医師が，カルテに正解付けを行う必要性がある．しかし，カルテを読んで患者の疾患を判断することを医療者でない研究者が行うことは難しいと考えられる．カルテは基本的には医療者同士が短時間で読んで理解できるように記録されるため，幅広い検査や処方，病気の知識を補いながら読まないと理解できない記載も少なくない．そこで，可能であれば主治医，そうでなくても専門医やその他の医療者でカルテに対象疾患の有無の正解付けを行う必要がある．しかし，病気の細かいタイプまで厳密に区別する臨床上の必要性がない場合もある，病気の発症や治癒の時期を明確に定義するのは難しい，臨床現場においては診断基準だけではなく病気の定義自体が変更されることが珍しくない中で研究者の意図とすり合わせながら正解付けをしないといけない，などの理由で，医療者であっても厳密な判断に手間と時間がかかる場合がある．

そのため，少しでも自動でカルテへの正解付けを行い，コストを減らす必要がある．しかし現段階で，全症例のうち自動で正解付けできる症例はどれくらいあるのか，そのためにはどのような解析技術を用いれば良いのか，などは明確になっていない．

糖尿病の中の一つのタイプである 2 型糖尿病は，患者数も多く，他の様々な疾患との関連も知られている．また現在も新しい治療薬の開発や診断基準の変更が続いており，phenotyping の需要が高い疾患の一つである．2 型糖尿病の患者を同定する手法は多く開発されているが，いずれも十分な精度に達しているとは言えない [Rachel 2013]．そこで本研究では，2 型糖尿病を例に，電子カルテの構造化データおよび自然文データを用いた phenotyping の技術開発のためのカルテへの正解付け，および phenotyping 技術の応用における問題点を整理し，必要な技術的課題を探る．

連絡先：香川璃奈，東京大学大学院 医学系研究科，東京都文京区本郷 7-3-1 東京大学医学部附属病院 企画情報運営部内，kagawa-r@m.u-tokyo.ac.jp

2. 実験材料

東京大学医学部附属病院(以下，東大病院)を 2012 年に 2 回以上外来受診または入院した患者 104,522 人(平均年齢 52.6(標準偏差 26.5)才，女性 57.2%)を対象とし，2009 年 1 月 1 日～2014 年 12 月 31 日の電子カルテデータを用いた．ランダムに抽出した 611 例について，電子カルテを実際に確認し，医師 3 名(筆者ら)により各患者が 2 型糖尿病か否か判断した．

3. 実験手法と結果

3.1 カルテへの正解付けにかかる時間的コスト

実際に医師がカルテを読んで患者が 2 型糖尿病か否かを確認しようとすると，単純には分類できない症例が少なからず存在する．しかしこのような症例をどのように分類したのかについて，明記されている先行研究は調べた限りで見つからなかった．そこで本研究ではカルテの正解付けアルゴリズム(図 1)を作成し，患者を分類した(表 1)．2 型糖尿病と他の型の糖尿病の併発の患者は 2 型糖尿病の患者数にカウントした．これによって，カルテの正解付けを自動で行う際にどのような材料あるいは技術が必要になるかの分析を詳細に行えるようになる．

Phenotyping の開発の際には，専門家の知識や意見を参考にする場面が大きく分けて 2 つあると筆者らは考える．一つは，phenotyping を行うためのルール決定や変数選択を行う際に対象とする病気に関する知識を参考とする場面である．もう一つは，カルテに gold standard となる病名を付与する場面である．前者は，多くの疾患に関しては教科書や診療ガイドラインに基づいた判断が研究者にもある程度可能である．さらに，特殊な疾患でない限りその疾患の専門科の医師であれば十分に知識は持っている．特に 2 型糖尿病ほど一般的な疾患であれば，研修医で

も最低限の答えは可能だろう. しかし後者の場合, どのような疾患を対象とする場合でも, それぞれの疾患に対して必要な症例数のカルテを確認しないといけない. さらに背景に記述したような問題点があり, 時間コストがかかることが予測される. そこで, 東大病院の電子カルテシステムを用いて, カルテの正解付けにかかる時間を評価した. この時間評価実験のみ医師 1 名により, 時間計測もカルテを確認した本人が行った. 表 1 の結果から, 特に糖尿病でないと判断した症例や, 2 型糖尿病か他の型の糖尿病かの判断が難しい症例の正解付けには, 症例数に比例して時間がかかることが分かった.

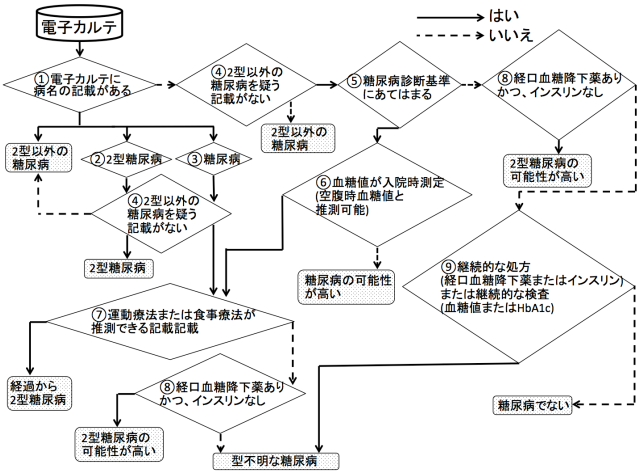

図 1　カルテへの正解付けのアルゴリズム

表 1 カルテの正解付けの結果と用いた時間

正解	2 型糖尿病	経過から 2 型糖尿病	2 型糖尿病の可能性が高い	型不明な糖尿病
患者数	35	26	1	4
時間(秒)	127.1 (105.8)	276.9 (138.0)	170	285.0 (155.7)

正解	糖尿病の可能性が高い	2 型以外の糖尿病	糖尿病でない	合計
患者数	15	9	521	611
時間(秒)	365.5 (122.5)	353.8 (344.8)	252.5 (188.9)	

3.2 カルテの病名情報の正確性と網羅性

カルテに記録されている 2 型糖尿病に関する病名情報で可能となる自動正解付けの範囲を探るため, カルテに記録されている病名情報の正確性と網羅性について調査を行った. 電子カルテには, 自然言語による自由記載部分に病名が自由記載されるが, これとは別に医療費請求のために病名と ICD-10(国際疾病分類)コードを登録するための専用病名登録ページがある. 自由記載部分に 2 型糖尿病の記載がある, 糖尿病の記載がある(2 型糖尿病, 2 型以外の糖尿病の病名記載がある患者は含めない), 後者に登録されている ICD-10(国際疾病分類)コードが E11x(2 型糖尿病), E14x(糖尿病) の 4 種類の病名に関して, 実際に医師がカルテを確認して 2 型糖尿病だと判断した患者とそうでない患者, それぞれについてその病名を持つ患者の数を調査した(表 2 上段). 表 1 における「2 型糖尿病」「経過から 2 型糖尿病」「2 型糖尿病の可能性が高い」に含まれる患者を 2 型糖尿病と見なした.

この結果から, 2 型糖尿病患者のうちカルテの自然文に 2 型糖尿病と明記されているのは約 57%に過ぎないことが分かった. またコードで 2 型糖尿病と登録されている(ICD-10 コード E11x)患者のうち実際に 2 型糖尿病である患者は約半数であると分かった. ICD-10 コード E11x を用いて自動で正解付けを行うと 2 型糖尿病患者のうち約 35%を正解付けできないと分かった. E14x を用いても 2 型糖尿病患者の約 11%を正解づけできないだけでなく, 正解付けされた患者のうち約 74%は実際には 2 型糖尿病ではないと分かった.

3.3 2 型糖尿病を特徴づける説明変数の網羅性と, それを用いた自動正解付けの試み

カルテに記録がある病名情報だけを用いても正解付けが十分に行えないことが分かった. しかし, 2 型糖尿病に関連があると考えられるデータは病名以外にも検査や処方など多種存在する. それらを用いて 2 型糖尿病患者のうちどれくらいの患者数を正解付けできるのか調査した.

臨床上特に重要かつ比較的多くの 2 型糖尿病患者が持つと考えられる変数に関して, 実際に医師がカルテを確認して 2 型糖尿病だと判断した患者とそうでない患者, それぞれについてその病名を持つ患者の数を調査した. 血糖値の異常値(200 mg/dL 以上)の有無, HbA1c(国際標準値)または HbA1c (JDS)の異常値(HbA1c(国際標準値)6.5%以上, HbA1c (JDS)6.1%以上)の有無, 経口血糖降下薬処方の有無, インスリン処方の有無について調査した(表 2 下段). この結果から, 糖尿病を特徴づける変数も単一の変数だけを用いて, 2 型糖尿病患者の正解付けを自動では行えないことが分かった.

また, 2 型糖尿病患者について, それぞれの説明変数がある症例の重複は図 2 の通りである. この結果から, これらの説明変数を単純に組み合わせるルールでも 2 型糖尿病患者を正しく自動で抽出することはできないことが分かった.

これらの項目を適切に組み合わせることで自動での正解付けが可能になるかどうか検討するために, 表 2 に示した項目を初めとする, 糖尿病と関連する 19 変数を用いて 2 型糖尿病か否かの分類器を作成した. 特に PPV を上げることを重視したルールベースでは感度 51.0%, PPV 96.2%であった. また, 複雑な症例を分類できることを期待して SVM でコスト考慮型学習を行い 2 型糖尿病か否かの分類を行ったところ, 感度 88.9%, PPV 61.5%であった. 以上のように, 説明変数からだけでは自動では正しく正解付けができない症例が存在する.

表 2 2 型糖尿病患者とそれ以外の患者それぞれについてのカルテにおける情報の記録の有無について

	2 型糖尿病患者		それ以外の患者	
情報	記録あり	記録なし	記録あり	記録なし
2 型糖尿病の記載	36	26	0	549
糖尿病の記載	24	38	6	543
E11x 病名登録	40	22	46	503
E14x 病名登録	55	7	160	389
血糖値	30	32	13	536
HbA1c	52	10	14	535
経口血糖降下薬	31	31	0	549
インスリン	22	40	11	538

3.4 構造化されたデータだけでなくカルテの文脈を把握する必要がある, 正解付けの複雑さの詳細

3.3 で述べたように, 説明変数を用いても自動では正しく分類できない症例が存在する. カルテの正解付けアルゴリズムを参照すると, 構造化データだけでは 2 型糖尿病か否かを判断でき

ず，カルテのコンテキストの理解が必要となる症例が存在することが分かる．正解付けアルゴリズム(図 1)の④，⑤，⑥，⑦，⑨の過程が該当する．

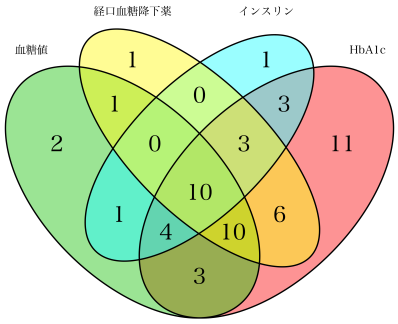

図 2　2 型糖尿病患者の説明変数の重複

④二次性糖尿病など 2 型以外の糖尿病の除外

正解付けのためにカルテを読む中で，二次性糖尿病など 2 型以外の糖尿病の可能性を除外するために必要な情報を得る必要がある．二次性糖尿病とは，血糖上昇作用を持つ薬の内服，膵癌により膵臓を摘出したこと，遺伝性疾患や内分泌疾患など，他の病気やそれに対する治療によって血糖値が上昇し糖尿病になった状態のことである．1 型糖尿病という生来インスリン分泌が不足しているため糖尿病である可能性も除外する必要がある．

1 型糖尿病であると判断できた 3 例(うち 1 例は 2 型糖尿病と併発)は，すべてカルテの本文中に病名が明記されており，さらに ICD-10 コードで 1 型糖尿病(E10x)が登録されていた．二次性糖尿病は確認した限りで，実際には二次性である，あるいは二次性の可能性があると判断した患者は 9 例(うち 2 例は 2 型糖尿病と併発)であった．その中でカルテの自然文に「ステロイド糖尿病」「糖尿病(LC による)」など病名が明記されているのは 4 例，「糖尿病」と書かれているのが 3 例，糖尿病に関する病名が明確には書かれていないのが 2 例であった．また，ICD-10 コードで二次性糖尿病(E12x または E13x)が登録されているのは 1 例であった．すなわち，二次性糖尿病という比較的稀で特殊な病態も，カルテに病名が正しく記載されているわけでも病名コードが正しく登録されているわけでもないことが分かった．

病名が明確に記録されていない中でカルテを読んで二次性糖尿病と推測できた記載は，下記のようなものであった．

◆「PSL10mg を内服していた際に血糖高値」(血糖値が上昇する二次的な要因があることが明記されている):4 例

◆二次性糖尿病をもたらす可能性のある治療を受けているため二次性糖尿病であると推測可能である．ただし「二次性糖尿病」というような明確な記述はない．:1 例

特に他院での治療内容は，構造化されてカルテ内に記録されていることが少なく，さらに東大病院では他院から紙で提供される診療情報はスキャンして画像として保管されていることが多い．そのため，電子化されていないデータにも全てに目を通す必要がある．また，全て東大病院で治療されている患者であっても，高血糖の原因となりうる出来事のあとに始めて高血糖を示しただけでは二次性糖尿病とは言い切れない．以前に高血糖を示したことは一度も無かったのか，高血糖を示した際に一過性高血糖を示しうる状況ではなかったのか(後述)，をカルテから把握する必要がある．

⑥血糖値が入院で測定されている，あるいは空腹時血糖値であるとことを推測可能

血糖値は空腹時には低く食後には高くなるため，糖尿病の診断を行うためには，空腹時血糖値か食後血糖値かを把握した上で血糖値が異常かだけで否かを判断する必要がある．しかし東大病院の血糖値のデータはその区別ができない．そこで便宜的に，入院時に測定されているデータか，カルテに空腹時血糖値だと解釈できる記載があるかどうかを確認する．入院時に測定されたかどうかは構造化データである入院記録を参照できる．外来患者であってもカルテの記載からその血糖値が空腹時か食後か推測できた記載として，下記の例があった．

◆「朝食前空腹時も高い」(自己血糖測定している場合など，血糖値が空腹時か食後か明記されている場合):2 例

⑦食事療法，運動療法が行われていることが分かる記載があるか否かの判断

糖尿病という病名がカルテ本文中に明記されている患者について，食事療法，運動療法が行われていることが分かる記載を電子カルテの自然文から拾うことができると，糖尿病の中でも 2 型糖尿病の可能性が高いと考えることができる．この情報は東大病院のカルテでは構造化されて格納されていない．おそらくどこの病院でも構造化されていないだろう．さらに，食事療法や運動療法は脂質異常症など他の疾患に対しての指導として記録されていることもあるため，明らかに他の疾患に対してのコメントではないこと，糖尿病に関連する記載であることを確認しなければいけない．よってこれらの情報は構造化データからだけでは判断できず，カルテの自然文の記載を用いる必要がある．

この情報に該当する実際の記載は下記の通り(重複は除く)．

◆「食事療法」「運動療法」:5 例

◆「エネルギー制限」「栄養指導」(別の言い回し):4 例

◆「血糖値は内服なし」「糖尿病食事のみ」(解釈が必要):1 例

◆「糖質を減らしたいが，ご飯が大好きで難しい」(患者の発言内容をカルテに記載):1 例

⑤検査値が糖尿病診断基準に当てはまる

⑨継続的な治療薬の処方あるいは検査

糖尿病は本質的に血糖値が高いか否かで診断する病気である．しかし血糖値は糖尿病以外にも，様々な要因で一時的に高くなりうるため，血糖値が高いからといって必ずしも糖尿病とも限らない．したがって，継続的な治療薬の処方あるいは検査が行われている場合，これが糖尿病に対して行われているものなのか，あるいは一過性の高血糖に対して行われているものなのかを判断しなければならない．一過性高血糖に関しては，手術の後，ショック，グルコースを含む点滴によるものなど，原因として考えられる状況がとても多く，かつ臨床的にも複雑である．さらにカルテの自然文にも「一過性」と明記されている例は調べた限りでは見つからなかった．手術前後やショック状態という情報も，構造化されていることは少なく，カルテの自然文に直接明記されていることも少ない．さらに一過性高血糖であることを確認する場合には，その原因となる状態が解除されたときに血糖値が正常に戻っていることも確認する必要があるため，その状態が終了した時期の確認までしなければならない．これらの情報は構造化されて記録されていることはなく，カルテの自然文から判断しなければいけない．

血糖値が高くてもそれが一過性のものであれば，2~3 ヶ月の血糖値の平均値を反映する HbA1c は低値を示すことが多い．しかし，実際には高血糖が長い期間続いていても，HbA1c が見かけ上低値に見える場合が存在する．溶血性貧血や輸血などが該当し，血糖値が高値かつ HbA1c が低値だから一過性高血糖とも断言することはできず，HbA1c がどのような状況下で測定さ

れているのかの情報も必要である. この情報も, 病名の記録の確実性と網羅性が低いというカルテの特性上, カルテの自然文から読み取る必要がある.

これらの情報を推測できた情報源には, 下記の例があった.

◆「腹腔鏡下右腎尿管全摘後 POD1」(一過性高血糖を来しうる状態の明確な記載がある. POD1 は術後 1 日を意味):6 例

◆手術後, ショック状態など(カルテ本文に明確な記載はないが, 救急車で運ばれてきて意識清明でない状態での血糖値であることから推測できる, など):13 例

◆「HbA1c は貧血だろう」(HbA1c が見かけ上低値であると判断できる記載):2 例

4. 考察

4.1 正解付けに時間がかかる理由

3.1 の結果のとおり, 特に糖尿病でない症例や,「経過から 2 型糖尿病」などの 2 型糖尿病か他の型の糖尿病かの判断が難しい症例をカルテから正解付けを行うことには時間がかかる.

「2 型糖尿病」など病名が明記されているカルテが存在する場合でも, そこに十分な治療の経過が書かれていないときには, 2 型糖尿病という記載に誤りがないかを確認するために他のカルテを参照する必要がある. これは図 1 の④の過程に相当し, このために病名が明確に書かれているカルテであっても正解付けに時間を要する場合がある. 2 型糖尿病か他の型の糖尿病かの判断が難しい症例の正解付けに時間がかかるのも, 同じ原因が考えられる. さらに「糖尿病でない」症例を正解付けする際にも, 過去のカルテ全てに目を通して該当する記載がないことを確認するため, 時間がかかると考えられる. また, 検査値が正常値であっても, 糖尿病であるが治療によるコントロールがとても良い患者である可能性がある. そのため, 検査値だけでは糖尿病でないと判断できないのも, 正解付けに時間がかかる一因である.

4.2 病名や説明変数で正解付けができない医療的背景

医療現場では保険請求上の病名の記録が必要であるため, 糖尿病でない患者にも糖尿病の病名コードがついている(表 2)例が少なくない. それは, 血糖値を検査して実際には糖尿病ではなかった人に対しても, 検査に保険請求するために糖尿病に関連する病名コードをつける必要があるからである. また, 日本では患者がいつどこの病院を受診するかは患者の自由である. そのため患者の情報が様々な病院に分散することがあり, 罹患期間が長い患者の発症経緯や病気の経過を細かく把握できずそもそも厳密な診断が不可能である, という問題が存在する.

糖尿病に特徴的な問題として, 患者数が多い慢性疾患であるため, 必ずしも専門家が診断しているわけではなく「糖尿病」「2 型糖尿病」「耐糖能異常」などの類似する診断名(状態名)が厳密な診断基準に基づいていない場合がある. また, 糖尿病はタイプが異なっていても, 血糖値を下げる, という治療方針が大きくは変わらないためわざわざ詳細に検査を行い細かいタイプまで診断する必要性が臨床上低い場合も少なくない. さらに, 糖尿病を他院で治療中であれば自院で細かい検査や処方を行う必要がない. このような理由で, 診断名やカルテの記載が曖昧になり, 検査値や処方が記録に残らない場合がある. このことも, 正解付けのためのカルテの自然文理解の必要性と正解付けの時間コストの一因となる.

4.3 正解付けに必要な技術

3.4 の結果より, 症例が少ない複雑な条件も適切に把握することが, カルテへの正解付けには必要であることが分かった. このような症例は数が十分に集まれば自動で分類することも可能であろう. しかし現状では, 3.4 の結果のようにそれぞれの症例の数は少ない. そのため, カルテの自然文に書かれている詳細なコンテキストを抽出することで, 情報量を増やして, 自動での正解付けが少しでも可能になるようにする必要がある. 特に患者の発言そのままの記録には含意関係認識などが有用な可能性がある[金子 2013]. また, 関連する他の病名(状態)の phenotyping 技術の開発も重要であろう.

4.4 今後の課題

病名情報だけで自動で正解付けが可能な患者と, そうでない患者とのあいだに, データの偏りがあるのかどうかについては, 現時点で解析できていない. 併存疾患の割合, 受診科の偏り, 検査値, 治療方針の偏りなどがこの 2 者の間に仮に無いとするならば, 利用目的に応じて, 病名情報だけで自動で正解付けが可能な患者のみを研究に用いる選択肢もありうるだろう.

本研究では 2 型糖尿病を例に解析を行ったが, 他の病態や状態を推測する上では, 今回の解析とは全く異なる問題点があがる可能性がある. 例えば不妊治療の胚移植の正解付けを考える. これは医療行為の実施の有無に関する正解付けなので, 実際の答えは明確なタスクである. この場合, そもそも 2 型糖尿病における「厳密に診断されていないので答えが曖昧にならざるを得ない」という問題点は存在しない. しかし, 不妊治療は保険適用外の治療であるため, 保険請求上の病名登録が無い問題がある. 胚移植に特異的な検体検査などは無く, 正解付けが電子カルテの自然言語による自由記載に大きく依存することになる. 新たに医事会計システムのデータの利用を考える必要もあるだろう. さらに, 1 患者につき短期間しか起こりえない状態名であるため,「この日に胚移植を行った」「この日が胚移植の最終日である」という日にちの推測まで求められるだろう. 2 型糖尿病の推測のような長い病歴は必要ない代わりに, 様々な略語や類似する表現を全て抽出し, さらにカルテには明記されない開始と終了の日にちを判断するために胚移植以外の不妊治療に関連する単語を抽出する必要があり, 特に辞書が充実した NLP 技術の開発が必要となると考えられる.

本研究は東京大学医学系研究科・医学部倫理委員会により承認を得ています(承認番号 10791).

参考文献

[Jie 2015] Jie X, Luke VR, Pamela LS, Guoqian J, Richard CK, Huan Mo, *et al.* Review and evaluation of electronic health records-driven phenotype algorithm authoring tool for clinical and translational research. J. Am. Med. Inform. Assoc. 2015.

[McCormick 2014] McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. PLOS One. 2014.

[Rachel 2013] Rachel LR, Shelley AR, Douglas W, Bryan CB, Mark NF, Marie LM, *et al.* A comparison of phenotype definitions for diabetes mellitus. J. Am. Med. Inform. Assoc. 2013

[金子 2013] 金子貴美, 宮尾祐介, 戸次大介. 基本文関係に分解した含意関係認識日本語評価データの構築. 自然言語処理学会第 19 回年次大会発表論文集, 2013.