

# 電子化により外来カルテ文章の情報量は増加する傾向がある

荒牧 英治<sup>1)</sup> 増川 佐知子<sup>1)</sup> 山田 恵美子<sup>2)</sup> 脇 嘉代<sup>2)</sup> 大江 和彦<sup>2)</sup>

東京大学 知の構造化センター<sup>1)</sup> 東京大学医学部附属病院<sup>2)</sup>

## Introduction of electronic health record system increases quantity of chart

Aramaki Eiji<sup>1)</sup> Masukawa Sachiko<sup>1)</sup> Yamada Emiko<sup>2)</sup> Waki Kayo<sup>2)</sup>  
Ohe Kazuhiko<sup>2)</sup>

Center for Knowledge Structuring, University of Tokyo<sup>1)</sup> University of Tokyo Hospital<sup>2)</sup>

[Purpose] While Electronic health record (EHR) systems have spread rapidly year by year, several disadvantages are also pointed out. One of the most claimed disadvantages is that keyboard typing is a tough task, decreasing the amount of record description. Another claim is that an EHR contains a duplication of the same information because an EHR enables a copy-paste writing style. Although such disadvantages are often claimed, they are not experimentally validated. To investigate the characteristic of EHR texts, this paper analyzes the EHR texts using various indicators/metrics. [Methods] 85 patient records, written by 44 medical doctors from April 2003 to 2010 February, are randomly selected. The first part of the records (from April 2003 to June 2009) is paper based records that are written by hand. The others (from June 2009 to February 2010) are written due to EHR systems. [Result] While quality measures indicated that there is not significant difference between the paper-based records and EHRs, the quality measure indicated that EHR contains more information than paper-based record, encouraging broad use of EHR.

Keywords:

### 1. はじめに

平成13年度「保健医療分野の情報化にむけてのグランドデザイン」以降、本邦では徐々に診療録の電子化が進んでいる。電子カルテは、手書きカルテに比べ、診療情報の保存や管理が容易で、臨床研究を行う際にも診療情報を効率よく処理でき、利便性に優れていると考えられている。一方で、患者と対面しつつ入力することは負担であり記述量が減るのではないかという指摘や、電子カルテ内の既存の文章を容易にコピー・ペーストして転用できるため、診療情報が重複して記録され、手書きカルテに比べて文章の量は増えたものの、期待されたほど情報の質は向上しておらず、実際の情報量に変化はない、という指摘もある。これまで電子カルテと手書きカルテを比較してカルテに含まれる医療情報を量的、質的に検討したものは少なく、記述されるテキストに電子カルテ特有の特性があるのかもわかっていない。

以上のような背景のもと、本研究では電子カルテと手書きカルテ間での記述されたテキストに含まれる医療情報の質的、量的な差異を明らかにする。

### 2. 材料

倫理委員会の承認を得て、糖尿病・代謝内科の外来受診通院中の患者85名(44名の医師が記載)の手書きカルテ(2003年4月-2009年6月)及び、電子カルテ(2009年6月-2010年2月)の診療録を対象とし

た。

本研究の目的は、自然文の性質の変化を調べたいので、自然文が多いと考えられる患者主訴部分のみを抽出し分析材料とした。患者主訴の抽出は明示的に「S」と記述されていればその箇所を、明示的な記述がない場合は記述内容を参考に人手で判断した。抽出した量を表1に示す。

表1 材料のサイズ

	手書きカルテ	電子カルテ
受診回数	1841	484
文数	4103	1658
文字数	46756	31899

### 3. 方法

コーパス言語学、計量言語学や自然言語処理の分野で提案されている12の尺度を用いてカルテの文章を質的及び量的に定量化し、それらを指標として電子カルテと紙カルテを比較する。比較に用いた尺度を表2に示す。これらは、大別すると次の2つの観点で分けられる。

- 1) 質を測る指標か(質),または、量を測る指標か(量)
- 2) 医学的な指標か(医),または、一般文章用の指標か(汎)

これらの区別もあわせて表2に示した。

表2 比較に用いたテキスト評価尺度

評価尺度		概要	(1)	(2)	評価尺度		概要	(1)	(2)
STR	Script type Ratio	文字種の割合	質	汎	SynComp	Syntactic Complexity	統語的複雑さ	量	汎
POSR	Part Of Speech Ratio	品詞の割合	質	汎	DisComp	DisComp	談話的複雑さ	量	汎
CaseR	Case Ratio	格助詞の割合	質	汎	#Chars	Discourse Complexity	受診毎の文字数	質	汎
FLUr	Fluency (R Value)	流暢さ [Guiraud1954]	質	汎	#MWs	Number of Medical Words	受診毎の医学用語数	質	汎
FLUc	Fluency (C Value)	流暢さ [Herdan1960]	質	汎	SimSR	Similar Sentence Ratio	前回受信時と類似した文の割合	質	医
FLUα	Fluency (α2Value)	流暢さ[Maas1972]	質	汎	SameSR	Same Sentence Ratio	前回受信時と同じ文の割合	質	医
MR	Machine Readability	機械処理可能な医療表現数	量	医					

### 3.1 基本統計量

文字種の割合 (STR), 品詞割合 (POSR), 格助詞の割合 (CaseR) は, 文体を計量化する最も基本的な尺度であり, 古典文学の著者推定などで頻繁に使われる質的指標である。

### 3.2 語彙

FLUは第2言語教育/翻訳において用いられる指標である。流暢でない話者は少ない語彙しか持たないという仮定に基づいた, 語彙の豊富さを測る指標であり, 複数の定義が提案されているが, 本研究ではFLUr<sup>1)</sup>, FLUc<sup>2)</sup>, FLUα<sup>3)</sup>の3つを用いる。式中のNは延べ語数であり, どれだけ多くの語彙が標本に含まれているかである。V(N) は異なり語数, すなわち, 何種類の語彙が標本に含まれているかである。

$$R = \frac{V(N)}{\sqrt{N}} \quad [\text{Guiraud1954}]$$

$$\alpha^2 = \frac{\log N - \log(V(N))}{\log N} \quad [\text{Maas1972}]$$

$$C = \frac{\log(V(N))}{\log N} \quad [\text{Herdan1960}]$$

### 3.3 機械可読性

MRは機械処理可能な医療表現数である。本研究では機械学習により構築した用語抽出器<sup>4)</sup>を用いて算出した。対象とする用語は病名, 検査名, 検査値や施設名など表3に示した用語を対象とした。

表3 指標MR計測の対象用語

D	疾患/症状表現, 疾患を示す表現.
T-NAME	検査名.
T-NUM	検査値.
M-NAME	医薬品名, 一般名も含む.
CHANGE	患者状態の変化を示す表現.
R	治療/処置を表す表現.
ACTION	入院イベントに関する表現.
M-NUM	医薬品の投与量及びその単位.
A-LOC	病院・診療所などの施設名.
B	人体部位表現.

### 3.4 複雑さ

SynCompは, 統語構造 (文内の関係) の複雑さを計測する。これには文に含まれる平均用言数を用いた。

文の構造が複雑であれば1文 (1主訴) 当たりの命題数が多いという仮定に基づいている。

DisCompは, 談話構造 (文間の関係) の複雑さを計測する。これには接続助詞の出現する割合を用いた。文間の構造が複雑であれば接続助詞 (しかし, だから,

～が etc) が多く出現するという仮定に基づいている。

### 3.5 情報量

受診毎の文字数 (#Chars) と受診毎の医療表現数 (#MWords) によって情報量を測る。医療表現数の計測にあたっては, 医学書院医学大辞典の見出し語, 南山堂医学大辞典の見出し語, MedDRA/J, 標準病名にマッチした語 (列) をカウントした。

SimSR, とSameSRはコピー・ペーストの割合を測る量で, 前回受診時の主訴と今回受診時の主訴を比較して計測する。

SameSRは両主訴間で同一であった文の割合であり, SimSRは両主訴間で類似していた文の割合である。ここで, 類似している文とは編集距離による類似度が0.8以上のものとした。

例えば図1のような主訴が与えられたとする。4つの主訴文 (b1~b4) のうち1文が前回と同じ (a1=b1) であるためSameSRは0.25 (1/4) となる。

同様に, 2文が類似度が高い (a1=b1, a2≐b2) ためSimSRは0.50 (2/4) となる。

2007/2/2 (a1) かわりなし HbA1c 5% 台  
(a2) 血糖コントロール良好

2007/3/4 (b1) かわりなし HbA1c 5% 台  
(b2) 血糖コントロールは良好  
(b3) インフルエンザ予防注射希望  
(b4) 本日入荷していないとご説明

図1 主訴記述の例

## 4. 結果

文字種の割合 (STR), 品詞割合 (POSR), 格助詞の割合 (CaseR) を図2に示す。また, 流暢さの指標 (FLUr, FLUc, FLUα) を図3に示す。これらはテキストの質を捉える尺度であるが, これらに有意な差は見られない。

### 3-D-1-6 一般口演/3-D-1:一般口演42

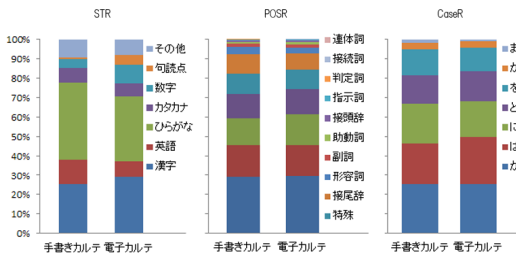


図2 文字種の割合(STR), 品詞割合(POSR), 格助詞の割合(CaseR)

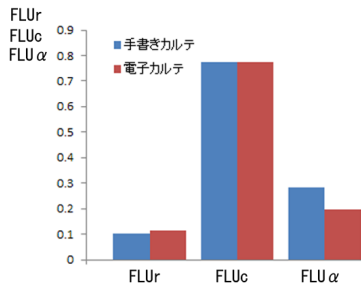


図3 流暢さの指標 (FLUr, FLUc, FLUα)

自動抽出できる医療表現の量(MR)を図4に示す。図に示されるようにすべての医療表現カテゴリにおいて電子カルテにおいて医療表現が多く抽出されている。よって、現状の言語処理技術では、電子カルテテキストの方から多くの医療情報を抽出できると考えられる。ただし、自動抽出された医療表現が多いことがすなわち実際の医療表現の量に差が結びつく訳ではないことに注意されたい(手書きカルテ文章には略語/略記など自動抽出が困難な表現が多く、これらが考慮されないからである)。

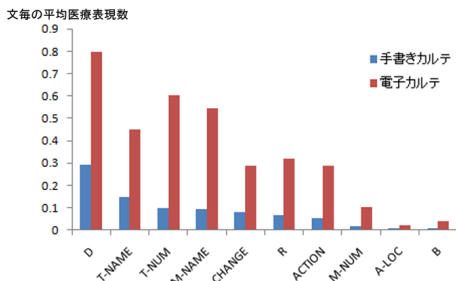


図4 自動抽出できる医療表現の量(MR)

統語構造の複雑さ(SynComp)と談話構造の複雑さ(DisComp)をそれぞれ図5左と図5右に示す。これらのいずれも電子カルテが有意に高く、より複雑な構造を持っていることが分かる。これは、手書きカルテが名詞句の羅列となる傾向が高く複雑さを持たないのに対し、電子カルテ上では文として記述されることが多いためだと考えられる。

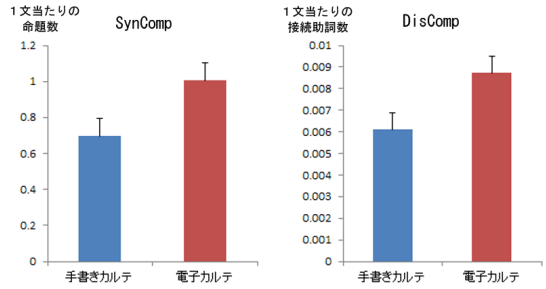


図5 統語構造の複雑さ(SynComp)と談話構造の複雑さ(DisComp)

受診毎の文字数(#Chars)と医療表現数(#MWords)を図6上に示す。

#Charsは62.0文字から164.4文字へ有意に増加し,#MWordsにおいても1.38から4.05に有意に増加している。

また、医師毎の手書きから電子カルテ移行後の#Charsと#Wordsの推移を図6下に示す。#Charsにおいては手書きの電子カルテの相関係数は0.254(p=0.05で有意,p=0.01で有意でない)であり,#MWordsにおいても手書きの文字数と電子カルテ文字数の相関係数は0.322(p=0.01で有意)と概ねの相関が得られた。このことから、手書きにおいて多く記載する医師は、電子カルテにおいても多く記載していることが分かる。

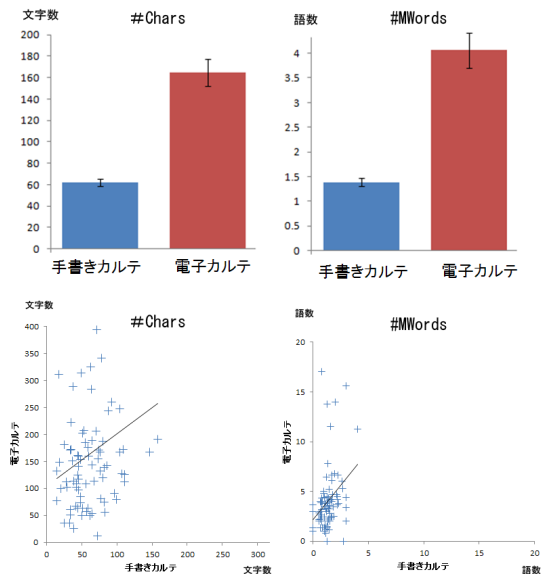


図6 受診毎の文字数(#Chars)と医療表現数(#MWords)

コピー・ペーストの割合を測る尺度(SimSRとSameSR)を図7に示す。SameSRにて12ポイント、SimSRにて10ポイントほど電子カルテでは高くなっ

ており、類似・同一文が増えていることが分かる。ただし、この増加分が実際にコピーペーストによるものかどうかは分からない。また、コピーペーストされたとしても、必要に応じてなされた可能性もある。したがって、本指標はコピーペーストが行われたとしても、ただだか12ポイントであるということを示している。

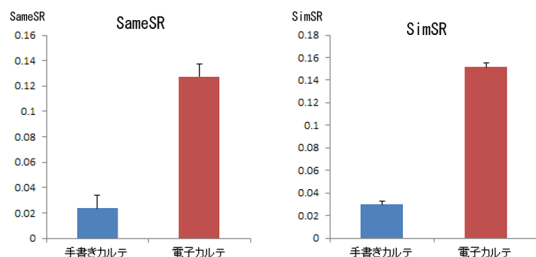


図7 コピー・ペーストの割合を測る尺度 (SameSR とSimSR)

## 5. 考察

文字種割合、品詞割合をはじめとする多くの質的尺度において手書きカルテと電子カルテとの差は見られなかった。

例外は文の複雑さ(SynCompとDisComp)だけであり、両尺度は電子カルテの方が複雑な構造をもつことを示している。これは紙カルテが名詞句の羅列なのに対し、電子カルテの方が自然文らしい記述が多いことに起因すると思われる。

また、量的尺度においては、あらゆる尺度で電子カルテの記述量が多いという結果が得られた。

この増加量はコピー・ペースト可能な割合の増量(SimSRの増加量12ポイント)をディスカウントしても高く、記述量は増えていることを示している。

また、この手書きカルテ時代と電子カルテ以降の記述量に相関があることから、医師の記述を増幅させる装置として機能しているとも考えられる。

最後に、増加した記述が臨床的に意味があるかどうか重要であるが、これについては多くの観点があるため、本研究では触れない。

## 6. おわりに

本研究では電子カルテと紙カルテを比較してカルテに含まれる医療情報を量的、質的に検討した。

質的尺度を用いた検討では紙カルテと電子カルテに複雑さ以外の有意な差は認められなかった。量的尺度を用いた検討では、紙カルテに比較し、電子カルテの記述量が高い傾向にあった。

## 7. 謝辞

本研究の一部は、日本学術振興会科学研究費補助金 若手研究(A)「非文法的かつ断片化したテキストからの情報抽出に関する研究」及び挑戦的萌芽「ダミー診療録の構築および自動構造化に関する研究」の援助を得てなされた。

## 参考文献

- [1] Guiraud P. Les Caracteres statistiques du vocabulaire. Essai de methodologie. Paris, 1954.
- [2] Herden G. Small particle statistics. 2nd edition: London, Butterworths Scientific Publications, 1960.
- [3] Maas H.D. Zusammenhang zwischen Wortschatzumfang und Lange Eines Textes. Literaturwissenschaft und Linguistik 1972; 8: 73-79.
- [4] Aramaki E, Miura Y, Tonoike M, Ohkuma M, Mashuichi H, Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification. Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2009) Workshop on BioNLP, pp.185-192, 2009.
- [5] Jurafsky D, Martin JH. Speech and Language Processing. pp.74, Prentice Hall, 2009.
- [6] similarity.pm. <http://search.cpan.org/~mlehmann/String-Similarity-1.04/Similarity.pm>.