生命科学シソーラスに基づいた医療教育ポータルシステムの開発研究

金　子　周　司　　京都大学大学院薬学研究科教授

1 研究の目的

インターネットを介して得られる医療健康情報は，一般市民だけでなく医療従事者および医療系学生にも影響を及ぼしつつある。特に，日本語で記述された解説記事などの増加は，相対的に英語で記述された原著論文など，科学的に評価の高い一次資料が利用されなくなる状況を招いている。しかし，もし英語のリソースを日本語で検索できるポータルが提供されれば，利便性と有益な波及効果が期待できる。そこで本研究では，インターネットで公表される広範な医学関連研究の成果を，日本人が検索あるいは理解しやすくすることを目的として，過去 10 年以上にわたって構築してきたライフサイエンス辞書（LSD, 金子他 1995, 金子 2006）の資源を最大限に利用するシソーラスの制作と日本語ポータルの研究開発を計画した。

本研究のゴールは，第一にゲノム科学の研究成果論文やデータベースを日本語で検索する場合に，入力したキーワードと密接に関連する別のキーワードを同時に提示することによって情報検索を容易にする検索サーバを開発することとした。また第二に，検索結果として表示される英語ページにおいて，利用者が求める箇所でオンデマンドに専門用語の対訳および解説を表示して，利用者の理解を助ける情報ポータルの試作を行った。英語で書かれたゲノム科学情報のあらゆる Web ページを日本語で検索して内容を理解できるサーバを無料で公開することで，医療や医学研究の成果を広く社会に提供する実用的なインターフェースとして幅広い利用と応用が見込める。

2 研究方法

2-1 シノニム辞書の制作

医療情報の理解に必要と考えられる専門用語の異表記を統一するための統制語としては，後で情報検索に利用することを考えた結果，まずはアメリカ National Library of Medicine が構築している Medical Subject Headings (MeSH)に準拠することにした。そこで MeSH 2008 (2007 年 11 月版）より解剖部位 (Tree A01-A17)，生物名 (Tree B01-B08)，病名および症候名 (Tree C01-C23 および精神疾患 F03)，生体分子および医薬品名 (Tree D01-D27 および Supplemental Concepts)，方法および尺度 (Tree E01-E07)，学問領域や現象 (Tree G01-G14) に帰属する専門用語から，上記のカテゴリに帰属できる Descriptor 21,684 語と LSD に同一の見出し語が収録されていた Supplemental Concepts 2,668 語を合わせた計 24,352 語を統制語として採用した（金子，藤田 2005)。これを元に，LSD と MeSH のすり合わせ作業を行い，シノニム辞書を制作した。

2-2 共起する統制語による関連概念データの制作

PubMed 抄録を収集した文献コーパスに対して,シノニム辞書を適用して統制語によるタグ付けを行う Perl スクリプトを開発した。統制語タグが同一抄録中で共起する頻度を解析し，各用語について出現頻度，共起する他の統制語およびその共起頻度を得た。得られたデータが専門的に見て妥当な関連性を表すかどうかを，複数名の研究者による目視によって検討した。この評価に基づいて，検索キーワードの取捨選択を行い，最適化を試みた。

2-3 関連概念を提示する情報検索エンジンの開発

シソーラスと共起解析データをオンライン版ライフサイエンス辞書 WebLSD に実装することによって,日本語および英語のいずれによっても表記のゆれを吸収して統制語による情報検索を可能にするポータルシステムを Perl cgi にて開発した。

2-4 日本語訳を表示する辞書ツールの開発

ウェッブブラウザで表示されるゲノム情報などの英語ページにおいて，可能な限り簡単な操作で専門用語を辞書引きできるツールを開発するため， Mac OS X 10.5 においてシステム標準で利用できる辞書.app での試作と検証を試みた。この辞書.app ではブラウザである Safari からショートカットで複合語レベルでの辞書検索が実現できる。また，辞書を制作するためのアプリケーションやテキスト使用が Apple において公開されている。

3 結果および考察

3-1 シノニム辞書の制作

2008 年 6 月時点で，表 1 に示すカテゴリの MeSH Descriptor および Supplemental Concepts (SC) に帰属する統制語 2.5 万語の 96%を日本語化し，英語表記と日本語表記を併記できるようにした。その上で，延べ約 16 万語の英語および日本語で記述される LSD 収録語および MeSH 用語を統制語に集約することで，対訳シソーラスとシノニム辞書を制作した。このデータから，16 万語の同義語のうち，LSD 収録の英語と日本語，および新たに加えた MeSH 英語が，それぞれほぼ 3 分の 1 ずつの割合を占めることがわかる。生体分子などの物質名，特に海外での医薬品商品名や化学一般名などの異表記を非常に数多く含む物質カテゴリにおいては，MeSH に由来する名称が半数に及び，これら新しく加えた用語によって欧米の文書に対する網羅性が高まったことが期待できる。

表 1　ライフサイエンス辞書のシソーラス化（サマリー）

Tree	カテゴリ	統制語数(a)	シノニム数(b)	平均異表記数 (b)/(a)	LSD 英語		LSD 日本語		MeSH 独自	
A	解剖部位	1,522	7,022	4.6	3,309	47%	3,060	44%	653	9%
B	生物名	3,478	16,499	4.7	5,157	31%	6,870	42%	4,472	27%
C+F03	病名・症候名	4,339	26,821	6.2	9,319	35%	11,695	44%	5,807	22%
D+SC	物質名	11,250	91,090	8.1	20,571	23%	25,678	28%	44,841	49%
	(医薬品)	3,569	42,387	11.9	7,749	18%	11,072	26%	23,566	56%
E	方法, 尺度	2,185	11,071	5.1	3,663	33%	4,676	42%	2,732	25%
G	知識, 現象	1,578	7,332	4.6	2,734	37%	3,024	41%	1,574	21%
計		24,352	159,835	6.6	44,753	28%	55,003	34%	60,079	38%

百分率はシノニム数(b)に対する割合を表す。

しかしながら一方で，LSD に収録されながら MeSH と照合できないため統制語に帰属されない用語が英語で 1 万語以上も存在することが明らかになった（表 2)。特に，病名・症候名や解剖部位名においては，国内で用いられている標準病名マスターや国際的な有害事象報告のための統制語である MedDRA にも収録されながら MeSH に帰属できない用語が数多く残された。また，医薬品としては国内医薬品において収録されていない用語が多数存在した。今後は，これらの語句を帰属させるためにツリーを拡張していく必要が示された。

表 2　統制語に帰属されなかった LSD 収録語

カテゴリー	統制語に帰属されない語句		例
	LSD 英語	(日本語対訳)	
解剖部位	2,242	2,376	sacral cord, natural killer T-cell, iPS cell
生物名	757	807	Periplaneta japonica, avian influenza virus
病名・症候名	4,871	5,757	varicella zoster, ketoacidosis
物質名	2,719	3,163	hemoglobin A1c, mozavaptan
方法，尺度	1,634	2,140	molecular imaging, chemoradiotherapy
計	12,223	14,243	

MeSH との照合は基本的に英語ベースで行ったため，日本語対訳で帰属されなかった語数は参考値である。

3-2 共起する統制語による関連概念データの制作

PubMed より代表的な学術誌に掲載された 10 年分の論文抄録（600 M バイト）をコーパスとして収集し，シノニム辞書によってテキスト中に最長一致で統制語の XML タグを施した。このタグ付けテキストの内容をブラウザで確認しながら，曖昧性の排除と統制語の最適化を行った（図 1）。この過程において，テキストでの一致のみによって統制語への変換を行う場合，曖昧性を排除するために多義性のある略語や商品名等，一部のシノニムをタグ付け辞書から除外する必要が生じた（約 200 語）。また，「ヒト human」，「病気 disease」，「酸 acid」等のように，非常に大きな概念は関連するキーワードとして不必要あるいは不適切と考えられたため，それら（約 360 語）もタグ付けから除外した。

図 1　英語論文抄録にタグ付けを施した XML データをブラウザで表示した例
テキスト中に出現する専門用語をすべて統制語に自動変換する Perl スクリプトを用いて XML データを作成した。日本語の統制語見出しを，物質や医薬品名は青色で，病名は赤色で，方法や尺度は緑色で表すことによって抄録で述べている内容に関連するキーワードの関係や，統制語の妥当性を判定できるようにした。この抄録の場合，「ベンゾジアゼピン」と「股関節骨折」の関係を述べている論文であることが一見してわかる。しかし「recipient」を「移植」という統制語に翻訳した箇所は誤りであるため，このような対応関係は解析辞書から除外する措置をとることによって最適化を行った。

次に，同一抄録中で共起する統制語のペアを収集することによって計 100 万対以上の統制語の共起頻度を求め，出現した 2 万語の統制語ごとに上位 30 対までの共起概念データを得た。ここで解析に用いるコーパスによって得られる共起概念のリストは大きく異なった。例えば，1.3 G バイトの臨床研究抄録を用いた解析では，ある薬物と共起する概念はほとんど医薬品名で占められ，標的となる生体分子や作用メカニズムを示唆するキーワードが得られない等，必ずしもコーパスが大きいからと言ってデータが適切にならない場合があることがわかった（図 2）。

本研究は医療系学部あるいは大学院に所属する学生による教育現場での利用を想定していたため，上述した代表的学術誌にこの 10 年間に発表された先端的研究成果を記述する広範な学術分野の論文抄録に限ることとした。試行錯誤の末，主観的に見てほぼ全ての用語カテゴリーにわたってバランス良い共起結果を得ることができたと考えている。しかし今後，コーパス母集団を変えることでさらに専門家の知識を反映するような最適化を試みる必要があると思われる。

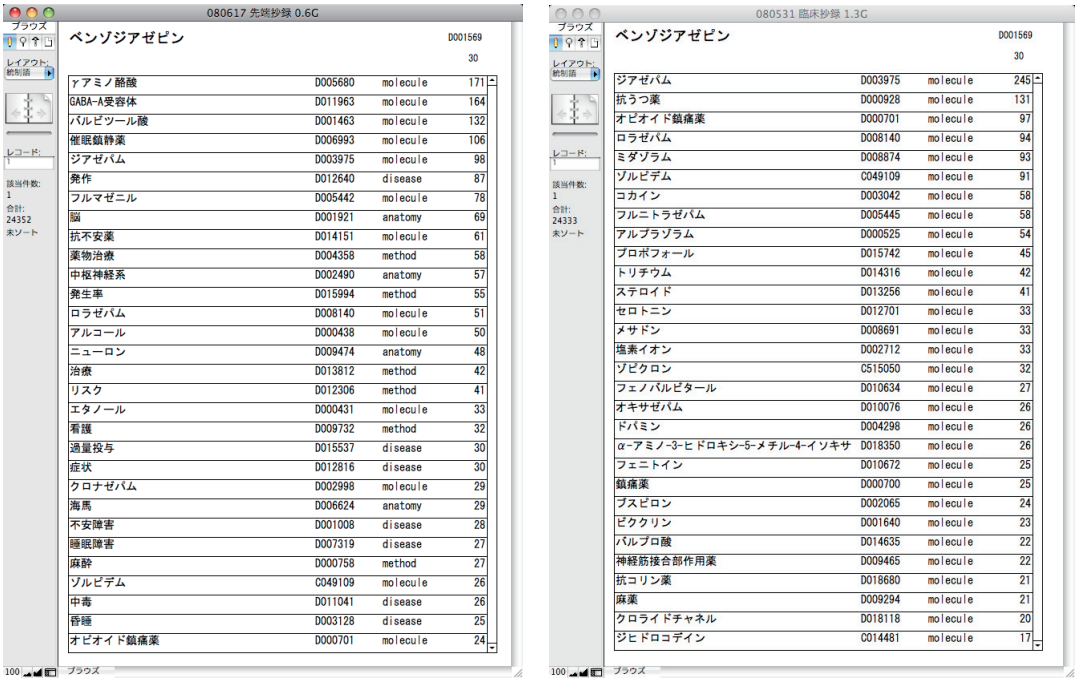

図 2 解析するコーパスによる共起概念の差異の例
（左，先端研究の学術誌抄録 0.6 G バイト，右，臨床研究の学術誌抄録 1.3 G バイト）

左の例では，解剖部位や疾患など多岐にわたる概念と共起しているが，右例では併用薬がほとんどを占める結果となり，検索キーワードの組み合わせとしては適切ではない。

3-3 関連概念を提示する情報検索エンジンの開発

このようにして得た共起概念データをシソーラスのツリー表やシノニム表示と組み合わせることによって，検索語として入力した日本語あるいは英語を自動的に統制語に直して表示するだけでなく，ツリーによって上位や下位の概念を探索できるようにしたり，関連性の高い共起概念を表示することで既存ポータルに適切なキーワード対を検索語として渡したりするためデータを XML 形式で制作した。

これらデータをまずウェッブブラウザで検索可能にするため，公開している WebLSD のサブセットとして，英和・和英対訳辞書と一体で使うことができるような cgi を制作し，2008 年 6 月より公開した。

この WebLSD に実装したシソーラスを用いることによって，任意に入力する日本語の検索語が英語に訳されるだけでなく，MeSH に準拠した統制語について，シノニム，ツリー，共起概念が表示される（図 3）。ツリーでは表示している統制語が赤字で表示され，その上位と下位に位置する概念へクリックで自由に移動することができる。共起概念は日本語と英語で最大 30 種類がそれぞれの統制語ごとに表示され，日本語をクリックした際には選んだ用語と統制語との組み合わせで Google へ検索キーワードが渡される。また，英語のリンクからは PubMed にキーワード対が渡されるようにしてある。基本的には URL を明示できる検索エンジンやデータベースに対して，このインターフェースを介してデータを渡すようにカスタマイズすることは容易にできるため，汎用性や応用性にも優れている。

3-4 日本語訳を表示する辞書ツールの開発

Mac OS X 10.5 には標準で辞書ツールである辞書.app が付属している。この辞書.app は日本語にも対応しており，キーワード入力に応じて結果を表示する incremental search を可能にした特徴を持っている。また，辞書.app は Mac OS X 標準ウェッブブラウザである Safari からショートカットキー（Command + Control + D）によって呼び起こすことができる。さらにこの時，カーソルが置かれている単語の前後を最長一致で判定し，最もその場所にふさわしい複合語を選び出して表示する他の辞書には見られない機能を有している。Apple 社は辞書.app に対応する辞書を制作するための技術資料を公開しているため，今回，この辞書.app を用いる辞書を制作した。

その結果，WebLSD で実装したシソーラスとほぼ同様の機能を有するスタンドアロン辞書を制作することができた（図 4）。辞書.app は検索語の途中でも先読みでキーワードを表示するため，前方一致するキーワード

リストを見ながら，適切な用語のシソーラスを見ることができる。シソーラス内での操作はほぼ WebLSD と同様であり，ツリーの上下移動や共起概念からの外部リンクを装備することができる。また，このようにして制作した辞書は Safari に表示された html ページのカーソル位置からショートカットキーで呼び出すことができるため，英和の用語検索が容易に行える。今後，さらに辞書.app および Safari の連携が簡単かつ高機能になることを期待したい。

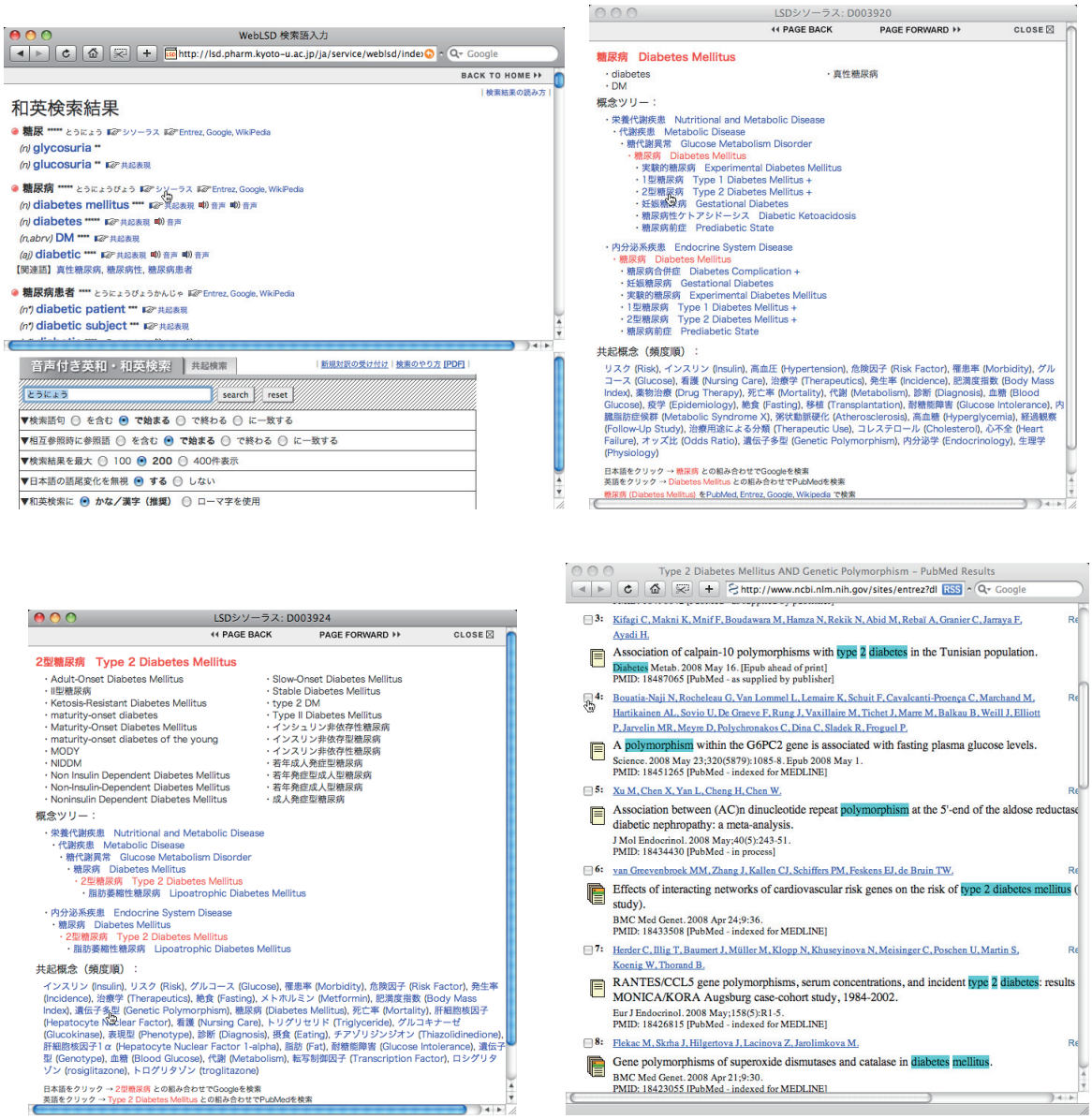

図 3　WebLSD に実装したシソーラスと共起概念による医療情報ポータルの使用例
（左上）「とうにょう」と入力したときに表示される和英辞書で「糖尿病」シソーラスをクリック
（右上）用語ツリーから下位概念である「2 型糖尿病をクリック」
（左下）2 型糖尿病の共起概念リストから「遺伝子多型 Genetic polymorphism」をクリック
（右下）PubMed に「type 2 diabetes mellitus」と「genetic polymorphism」の 2 つの MeSH が渡されて，ヒットする文献リストが表示される

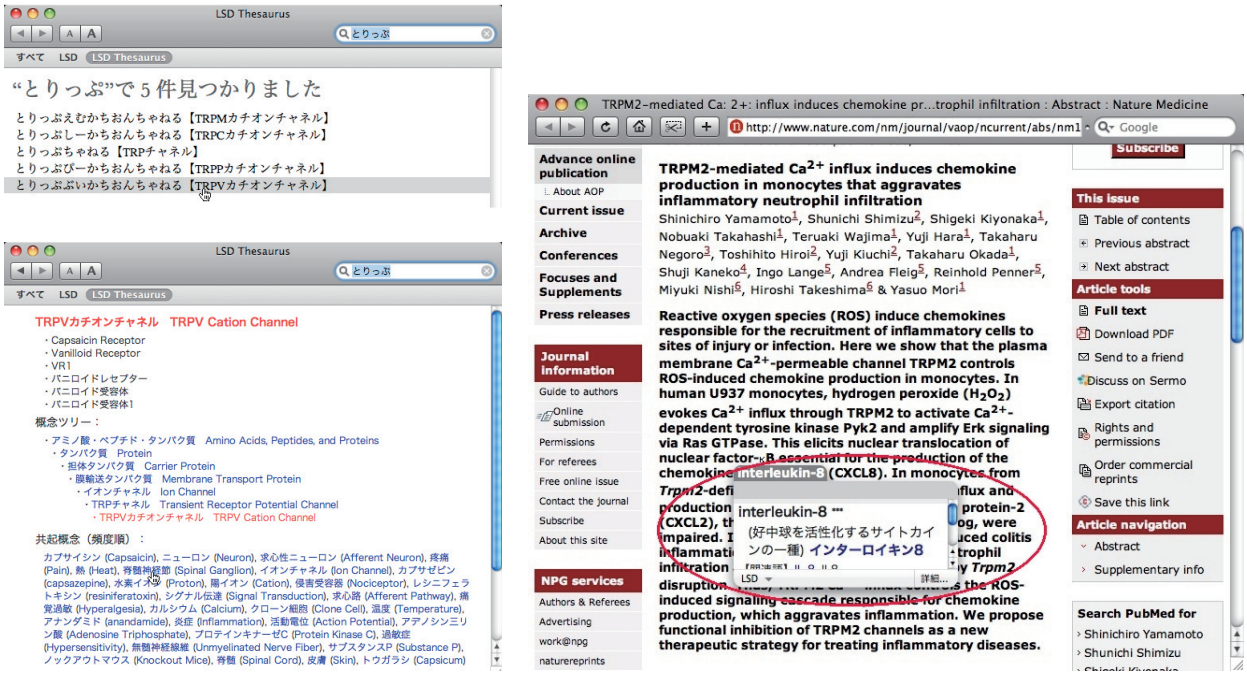

図 4　Mac OS X での検索ポータルの実装
（左上）辞書. app で「とりっぷ」と入力した段階で表示される候補語
（左下）TRPV カチオンチャネルのシノニム，ツリーと共起概念
（右）Safari で電子ジャーナルを検索した際にショートカットキーで辞書. app を起動した画面

4 結語

以上のように，本研究では当初の計画で予想した以上に有用なポータルを開発することができた。今後さらにデータの最適化を計り，公開ポータルとしての利便性を向上させる予定である。また，提示する共起概念の視覚的な表示技術についても検討していきたい。しかし，教育における本シソーラスの利用経験はまだ浅いため，将来的にこれらを医療情報教育に活用し，客観的な評価も進めたいと考えている。

謝辞　本研究は（財）電気通信普及財団研究調査助成（平成 18 年度）および（独）日本学術振興会科学研究費研究成果公開促進費（平成 17-19 年度，177002）の研究助成を受けて行われた。また，本研究にあたって，Perl スクリプトの制作とサーバおよび辞書の制作および公開には，藤田信之氏に大変，ご尽力いただいた。さらに，辞書. app については中村浩之氏から制作するきっかけとなる示唆をいただいた。以上，ここに記して深く感謝の意を表したい。

【参考文献】

金子周司, 鵜川義弘, 大武博, 河本健, 竹内浩昭, 竹腰正隆, 藤田信之
　ライフサイエンス辞書 2 の制作と公開, コンピュータサイエンス, Vol. 2, No. 2, 135-142, 1995.
金子周司
　文献情報の解析に基づく対訳シソーラスの評価, 医療情報学, Vol. 25, No. 6, 475-483, 2005.
金子周司
　ライフサイエンス辞書とは, 情報管理, Vol. 49, No. 1, 24-35, 2006.
ライフサイエンス辞書ホームページ
　http://lsd-project.jp/

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
医薬品名の同義語辞書と自動分類タガーの開発	日本薬学会第 128 年会	2008 年 3 月
国内外の医療用医薬品名を網羅する同義語辞書の制作	第 11 回日本医薬品情報学会学術大会	2008 年 7 月
医薬品名称の同義語解決による有害事象データベース AERS の情報活用	第 18 回日本医療薬学会年会	2008 年 9 月
医学用語シソーラスに基づく効率的医療情報検索システムの開発	第 28 回医療情報学会連合大会	2008 年 11 月