II. 分担研究報告書

厚生科学研究費補助金(第3次対がん総合戦略研究事業)
平成17年度 分担報告書

がん予防対象患者の病院情報システムからの抽出法に関する調査研究

分担研究者　大江和彦　東京大学医学部附属病院・教授

研究要旨【目的】 がん予防対象患者の病院情報システムからの抽出を標準的に実現する手法を開発する。【方法】がん予防対象患者としてハイリスクグループを設定する。具体的には、①家族性因子を有するもの、②ヒトパピローマウイルス (HPV)感染者、C型肝炎ウイルス感染者、を例として病院情報システムからこれらの該当患者を抽出するために何が必要となるかについて検討した。また専用のデータウエアハウスである Clinical Data Repository(CDR)を業務系システムとの間でＨＬ７v2 によりデータ転送する方法により構築した。さらに自然言語処理技術を用いコード化されていないテキストデータの解析手法を検討した。【結果】病院情報システムに入力される病名は標準病名マスタの普及によりある程度、統一的に処理することが可能になった。しかし、入力時に部位の左右情報の有無、疾患ごとの時間的な同時性の有無、など多岐にわたる付加的情報の入力を効果的に誘導するようなシステム機能の開発が求められた。一方、検査結果からの抽出や解析については、ＨＬ７Ver2 に準拠したデータ交換規約を採用し、運用系システムから転送系システムＣＤＲ（データウエアハウス）を構築することによって、効率的なデータ検索が実現可能であることがわかった。また、自然言語処理技術を電子カルテ等の医学文書データに適用して所見や悪性腫瘍の記述を抽出する手法により９０%前後の Recall と Precision が得られることがわかった。

A. 研究目的

オーダリングシステムを含む病院情報システムや電子カルテシステムの普及は2005年度当初で病院10%診療所5%程度である。今後こうしたシステムは次第に普及すると予想される。これらの情報システムに記録・蓄積されたデータ(以下CDR:Clinical Data Repository)からがん予防対策をとるべき対象患者を効率よく抽出する手法を確立することが重要である。本研究では、標準化を踏まえてこうした手法を確立する上での課題等を調査し、解決方法を提示することを目的とする。

B.研究方法

がん予防対象患者としてハイリスクグループを設定する。具体的には、①家族性因子を有するもの、②ヒトパピローマウイルス(HPV)感染者、③C 型肝炎ウイルス感染者、を例として病院情報システムからこれらの該当患者を抽出するために何が必要となるかについて検討し、その抽出システムを試作して課題を考察する。

C. D. 研究結果と考察

1. 家族性因子を有する者の抽出方法

1)病名の表記特性からの抽出方法

明確に判明している家族性腫瘍性疾患を同定するには、CDR中の患者病名または家族歴病名にこれらの疾患を有する患者を抽出することが考えられる。対象となる疾患をICD10対応電子カルテ用標準病名マスターV2. 41(以下標準病名マスター)から「遺伝性」「家族性」の文字列を有する疾患でかつICD10コードが新生物(CまたはD)であるものを抽出し、その後目視で選択した。その結果、家族性を含む新生物疾患は8病名、うち血液系の疾患を除くと家族性大腸ポリポーシスだけであった。また遺伝性を含む新生物疾患は23病名、うち血液系疾患を除くと、大腸多発性ポリープ癌、および遺伝性脳腫瘍、の2病名だけであった。

2)遺伝子診断分類により疾患例示による方法

遺伝子診断の分類 (J Clin Oncol 14:1730-6;1996)によれば、家族性腫瘍疾患は、次のように分類されている。

Group 1: 責任遺伝子が明確に同定されており、検査の結果によって医療方針を決めることができるような疾患。

家族性大腸腺腫症、多発性内分泌腫瘍症MEN2、網膜芽細胞腫、von Hippel-Lindou病など

Group 2: 責任遺伝子と特定の癌への易罹患性との関連がかなりの程度明らかになっているが、研究的側面を残す。

遺伝性非ポリポージス性大腸癌、家族性乳癌、Li-Fraumeni症候群 など

Group 3: 疾患と突然変異との関係が明らかでない場合、あるいは責任遺伝子との関係がごくわずかな家族でしか分かっていない。

末梢血管拡張性運動失調症、家族性黒色腫

これに列挙されている9疾患のうち、家族性大腸腺腫症は前述1)での家族性大腸ポリポーシスと同義である。そのほかでは、多発性内分泌腫瘍2型、網膜芽細胞腫、ヒッペル病(von Hippel-Lindou病)、毛細血管拡張性運動失調症の4疾患は標準病名マスターに収載されていたが、残りの4疾患は収載されていなかった。

3)家族性腫瘍の臨床的定義による推定

日本家族性腫瘍学会のホームページによると、「癌あるいは腫瘍の患者がたくさん発生している家系があります。同じ種類の癌または腫瘍である場合も、ある特定のいくつかの癌(腫瘍)である場合も、いろいろな癌(腫瘍)である場合もあります。このような家系では、大部分は、癌(腫瘍)が遺伝で発生しているのではないかと考えられています。しかし、遺伝ではなく、環境暴露によって家族内に癌(腫瘍)患者が多発している場合もあります。このような場合を、癌(腫瘍)の家族集積、家系内集積、あるいは家族性腫瘍(癌)と呼んでいます」とあり、その臨床的特徴として1)若年発症(一般の癌よりも若くして癌になる)、

2)多重癌・重複癌あるいは両側癌、

の2点が挙げられている。CDRからこの2点の特徴を抽出するには抽出のための計算機処理可能な定義を与える必要がある。本研究では、1)若年発症:腫瘍性疾患の疾患ごとに年齢分布の若年から2.5パーセンタイル点を得てそれよりも若年の患者を若年発症患者とする。2)2つ以上の異なる臓器または両側部位のがんを病名として持つ、と定義して抽出作業ができるかを検討した。

1)疾患別若年発症2.5パーセンタイル点:国立がんセンターのホームページ(http://www.ncc.go.jp/jp/statistics/2005/data07.pdf)には悪性新生物のICD10部位分類別、男女別、年齢階級別の罹患率が公表されているので部位分類別の2.5%タイルを算出することは可能であった。(ただし、このホームページはコピーライト保護のためか、データ表をダウンロードしてテキストデータとして活用することができないようにロックがかけてあるため、別途データの入手申請をするか、手で入力しなおす必要がある)。

2)重複がん・多重がんの抽出

i) 左右情報の問題

CDRには、部位の左右情報が入力されている場合とされていない場合が混在されているため、病名入力時に左右情報が存在すべき部位の疾患においては左右情報の入力を必須とするようなシステム開発が必要であった。

ii) 悪性腫瘍を同時に2箇所以上持つ患者を週出するには、その同時性と転移性の否定の判定がひとつの課題となる。同時性の判定については、理論的には病名開始日と終了日によって既定される罹患期間の重複を判定することが考えられるが、実際には終了日の入力は行われないことが多いことからこの方法では特異度が低いので、さらに重複がん、多重がんは明示的に入力を促すシステムが必要である。

2. ヒトパピローマウイルス(HPV)感染者の抽出

HPV検査は診療でルーチンに実施されているわけではないから、病名に「尖圭コンジローマ」を含む患者を抽出する方法が感度が高いと考えられる。検査が行われた場合のCDRからの抽出は特殊検査であるからロジックとしては容易である。日常のオーダーシステムのデータベースを直接検索することはシステム全体への影響を与える可能性が高いことから専用のCDRへのデータ転送システムの構築が多く

の場合に必要となる。

3. C型肝炎ウイルス感染者の抽出
HCV検査は通常の診療で実施されておりCDRに蓄積されているから、検査が行われた場合のCDRからの抽出は特殊検査であるからロジックとしては容易である。専用のCDRの構築の必要性は前述と同様に存在した。

4. 専用CDRシステムの構築
本研究では、病院情報システムのデータベースから1日1回、患者の基本属性、検査結果、病名について専用CDRへ転送するシステムの仕様を検討し、専用CDRの試作を行った。
1)データ転送方法
転送用データ形式はHL7V2. 4をベースとした保健医療情報システム工業会JAHISの臨床検査データ交換規約 <オンライン版> Ver. 2. 0を採用し、検査項目コードには日本臨床検査医学会臨床検査項目分類コードを使用した。
転送は、まず病院情報システム側で検査結果が生成され同システムのデータベースに更新処理が行われるたびに記録されるジャーナルファイルをもとにして、上記HL7形式のデータファイルを1検査報告ごとに1ファイル生成する形式でファイルを生成し、これをftpプロトコルにより専用のCDRに転送する方法をとった。
2)専用CDRの構築
専用CDRにはOS:Windowsserver2003、RDBMS:Oracle V9.x を採用したデータベースシステムを構築し、前記データをCDRに書き込むプログラムにより試作した。同システムではさらに頻繁に実施する抽出作業を簡易化するため mySQL データベースシステムと上記 Oracle データベースシステムを連携させ、mySQL データベースシステムにはビジネスオブジェクト社のOLAPツールを導入して、多角的なリアルタイム分析が可能となるようにした。

5. 自然言語処理による電子化診療情報処理方法
家族歴における家族のがん発生状況や重複癌の検出にかぎらず、高リスク患者の抽出を高精度に実現するには、単にコード化された病名や検査結果情報だけからの条件抽出では限界がある。電子カルテのデータのうち大半を占める自然言語文章データを直接解析できる自然言語処理技術を確立することが必要である。分担研究者は研究協力者の今井らとともに、本研究期間では放射線診断レポートデータを対象に所見の抽出、悪性腫瘍の判定などを実現することを試み、その手法をさらに発展させるため既存の医学教科書データから同様の所見が抽出できるかを検証した。1,155 文から正しく抽出されるべき用語が抽出できた割合 Recall は 87.2%、抽出された所見用語が所見である割合 Precision は 91.4% であった。いずれも 9 割前後の数値を示し、高精度に対象語の抽出と属性付与が行えることが示された[6]。

E. 結論
病院情報システムに入力される病名は標準病名マスターの普及によりある程度、統一的に処理することが可能になった。しかし、入力時に家族性疾患の有無、家族歴の有無、部位の左右情報の有無、疾患ごとの時間的な同時性の有無、悪性腫瘍の場合の転移性情報の有無、など多岐にわたる付加的情報の入力を効果的に誘導するようなシステム機能の開発が求められる。一方、検査結果からの抽出や解析については、HL7Ver2 に準拠した保健医療情報システム工業会JAHISの臨床検査データ交換規約を採用し、運用系システムから転送系システムCDR(データウエアハウス)を構築することによって、コード等の変換作業は必要となるものの、効率的なデータ検索が実現可能であることがわかった。また、自然言語処理技術を電子カルテ等の医学文書データに適用して所見や悪性腫瘍の記述を抽出する手法を試み、90%前後の Recall と Precision が得られることがわかった。

F. 健康危険情報
なし

G. 研究発表
1] N. Shinohara, H. Oyama, S. Matsuya, and K. OHE: A Computational Methd for Identifying Medical Complications based on Hospital Information System Data. Proceedings of CJK-MI Conference 2005(Chinese Hospitals

9(Suppl)), 133-134, 2005.2.
2] 大江和彦:カルテが持つべき機能やデータ規格の標準化が必要. 日本医事新報 No.4213 16, 2005.
3] 大江和彦:医療データの電子化と標準化. 厚生労働統計通信,第 26 号, 2005.
4] 大江和彦: 電子カルテと医療情報. EPS Magazine 遥か 2005, Vol.2, 53-56, 2005.6
5] 大江和彦、山本隆一(対談): 標準化・コスト・セキュリテイから電子カルテを読み解く. 月刊新医療 Vol.32 No.7:40-43,2005.7
6] 波多野賢二、大江和彦:電子カルテと医療情報の標準化. Medical Science Digest, 31(7),9(243)-11(245),2005
7] 大江和彦:我が国の医療情報システムの方向性. 映像情報メディカル, 37(13), 1347-1352,2005.12

学会発表
1] 大江和彦: 医療情報の標準化と普及. 医療情報学,25(Suppl),20, 2005.
2] 大江和彦: 臨床医学オントロジーとターミノロジー. 医療情報学,25(Suppl), 131-132, 2005.
3] 光石豊、遠藤徹、河添悦昌、高田真美、田中勝弥、美代賢吾、大江和彦: 診断報告書における病理医から臨床医へのリクエスト表現の分析. 医療情報学,25(Suppl), 688-691, 2005.
4] 波多野賢二、田代朋子、大江和彦: 合成語病名用語に対する ICD コードマスターの開発. 医療情報学,25(Suppl), 943-944, 2005.
5] 荒牧英治、今井健、柏木聖代、梶野正幸、美代賢吾、大江和彦: 自然言語処理による臨床医学オントロジーの自動構築の試み. 医療情報学,25(Suppl), 966-969, 2005.
6] 今井健、荒牧英治、柏木聖代、梶野正幸、美代賢吾、大江和彦: 自然言語処理を用いた画像診断所見オントロジー構築の試み. 医療情報学,25(Suppl), 972-975, 2005.
7] 篠原信夫、石坂崇、石井義興、小山博史、大江和彦: 時制データベースを用いた検体検査結果データウェアハウスの構築. 医療情報学,25(Suppl), 998-999, 2005.

H. 知的財産権の出願・登録状況(予定を含む)
なし

厚生労働科学研究費補助金(第3次対がん総合戦略研究事業)
平成17年度 分担研究報告書

がん予防に関する知識の体系化に関する研究

分担研究者 小野木雄三
所属:東京大学大学院医学系研究科・特任助教授

研究要旨:がん一次予防を支援するために、がん予防情報に関するオントロジーを作成し、これを利用して日本語の検索語から英語のがん予防情報に関する文献検索を行うことを目的とし、また得られた知識を有効に利用するために、その記述手法を検討した。 UMLSを利用することにより、医学概念に関するオントロジーを構築し、「がんの予防に関連する」などの複雑な関係を使った検索を行うことができるようになった。また、得られた知識を有効に利用するために必要となる知識の蓄積と利用方法に関する一定の基盤を得た。

A. 研究目的

がん一次予防を支援するために、がん予防情報に関するオントロジーを作成し、これを利用して日本語の検索語から英語のがん予防情報に関する文献検索を行うことを目的とする。また得られた知識を有効に利用するために、その記述手法を検討する。

B. 研究方法

昨年度の研究では統制用語集としてSNOMED-CTの概念間関係を利用したが、本年度はこれに加えてMeSHの階層関係自体を解析対象として利用した。一般に概念間関係は階層関係を記述するisa以外の関係が有用であるとされているが、詳細に見るとisaの関係は常に同じ観点から見た包摂関係ではないため、この階層関係の意味を明確に記述することによって、概念間関係を増やすことが可能となると考えられる。このisaで隠されている階層間の概念間関係は、UMLSのSemantic Networkを利用することによって収集することができる。これにより、従来得られていた概念間関係をより拡張した医学概念オントロジーを構築することができる。

このオントロジーを利用して、日本医学会医学用語辞典中に存在するが、既存の統制用語集には含まれていない各医学用語に対し、例えばそれが薬物名あるいはサプリメントなどの食品名に属する可能性が高いUMLS概念であるか否かなど、どの範疇に属する概念であるのかを推定する。

知識表現の記述手法に関して、材料となる知識はオントロジーとは別に既存のガイドラインを利用する。診療ガイドラインに記述されている知識をGLIFに変換し、これを知識データベースに蓄えて利用するための手法を、専門家システム記述言語であるJessを利用して検討する。

(倫理面への配慮)本研究は実際の患者情報を扱うものではなく、公開された医学用語集と文献情報を利用するものであるため、倫理面への配慮を要することはない。

C. 研究結果

UMLSに含まれる既存の統制用語集を利用し、明に記述されている概念間関係だけではなく、階層関係のisaを注意深く分類することによって、より多くの関係を抽出することができた。例えば薬物の構造から見た包摂関係と機能から見た包摂関係は、しばしば近い階層においても混在してisaとして記述されていることがあった。これらのisa関係をSemantic Networkにおける概念間関係で置換することにより、従来のオントロジーよりも概念間関係を拡充することが可能であることが示された。

本手法によって得られたオントロジーを利用し、任意の物質や食物(実際には任意の医学概念、およびそれを表す医学用語)が、がんの予防に関連するか否か、つまりがん発生の原因となるか、あるいはがんの発生を抑制するか、といった関係で関連する概念を探索することが可能となった。さらにその両者(任意の物質と関連する概念と)を用いて文献検索を行うことにより、詳細にその関連性を検討することも可能となった。

知識表現については、診療ガイドラインに記述された知識を例としてGLIFに翻訳し、専門家システムであるJessで知識を利用することの可能なエンジンを試作した。実際に「小児喘息急性発作への対処」、「乳がん術後患者のフォローアップ」、「高血圧の薬物療法」など複数のガイドラインをこのシステムに実装し、GLIFへの翻訳さえできていればエンジン側の変更がほとんど必要なくガイドラインの判断を実行することができることを確認した。これにより、がん予防に関する知識が得られた際に、その知識

を利用するための基盤ができたと考えられる。

D. 考察

作成した医学概念オントロジーでは概念間関係を既存のものよりも拡充し、これにより近接した概念間では関連づけの精度が向上した。しかし2階層以上の概念間関係の評価はかえって複雑になった。これはもとの概念そのものの規定や粒度に曖昧性が残っているためであり、より明確な論理に基づいた概念の構築が必要と考えられた。

また知識表現については、単なる 2 項間関係を超えた複雑なルールを処理するエンジンを構築することはできたが、中間形式として選択した GLIF に翻訳するコストが大きい。そのためオントロジーなどを利用して知識表現に翻訳するためのツールを開発する必要性が示唆された。

E. 結論

UMLS を利用することにより、医学概念に関するオントロジーを構築し、「がんの予防に関連する」などの複雑な関係を使った検索を行うことができるようになった。また、得られた知識を有効に利用するために必要となる知識の蓄積と利用方法に関する一定の基盤を得た。

G. 研究発表

1. 論文発表
なし。

2. 学会発表
張宇, 小野木雄三. 診療ガイドラインに従った患者支援システム開発の試み. 医療情報学, 25(suppl.) (in print) 2005

H. 知的財産権の出願・登録状況

1. 特許取得
なし

2. 実用新案登録
なし

3. その他

平成17年度厚生労働科学研究補助金(第3次対がん総合戦略研究事業(分野3))
(分担)研究報告書

研究課題名:「がん検診画像データからの予防情報抽出に関する研究」

分担研究者 若尾 文彦 国立がんセンター中央病院 放射線診断部医長

研究要旨: 検診の画像診断報告書の効率的作成を支援する所見テンプレートを考案し、所見テンプレーを実装した検診レポートシステムを構築し、PET 検診診断レポート作成システムに組み入れて,その利用状況について、解析を行った。診断-所見サマリー-所見テンプレートによる所見入力-サムネイル画像を用いることで、検診の際に発生する大量の画像情報を効率的にデータベース化する事が可能となり,画像情報の抽出に有用であると考えられた。

A. 研究目的

検診の画像診断報告書の効率的作成を支援する所見テンプレートを考案し、所見テンプレートを実装した検診レポーティングシステムを構築し、実際に、検診の診断レポート作成に用いて,その有用性について評価することを目的とした。

B. 研究方法

がん検診で発生する大量の画像情報から簡便な操作で、診断報告書を作成する「検診レポーティングシステム」および、診断報告書から検診結果報告書を作成する「判定登録システム」を構築し、検診業務の中で利用し、がん検診画像情報の登録状況を解析した。今年度は、PET 検診診断レポートを対象とし、テンプレートシステム導入前後の所見登録状況についても検討をおこなった。

検診レポーティングシステムでは、入力の手間を軽減するために、臓器ー診断ー判定ー判定コメントー所見サマリーで構成されるセットを作成し、診断ー判定のセットを選択することにより、他の項目が選択され、簡単な操作で入力できる設計した。さらに、詳細な所見を入力するために、所見テンプレートとして、病変の詳細情報を簡単に記載できるシステムを構築した。

C. 研究結果

本検診レポーティングシステムを検診業務の中で利用し、その利用状況について解析を行った。対象は、PET 検診診断レポートにテンプレートシステムの使用を開始した 2004 年9月 21 日から 2005 年 11 月 11 日までに作成された PET 検診診断レポート 2029 例とした。

このレポートに登録された総診断数は 4,024 件で、異常無し:528 件、PET による診断:2,575 件、CT (PET-CT として実施)による診断:921 件であった。

PET による診断の内訳を表1に示す。PET による診断では、腸管生理的集積や、胃生理的集積、肩関節炎症性集積が多かった。

表1:PETによる診断

診断名	件数
異常なし	530
腸管生理的集積	854
胃生理的集積	463
肩関節炎症性集積	417
肩甲部集積	104
慢性甲状腺炎	70
甲状腺腫瘍	67
肺門集積	49
肺炎症性集積	44
股関節炎症性集積疑い	41
頚部炎症性集積	40
高血糖による高バックグラウンド	31
甲状腺生理的集積	28
頚部機能的集積	27
肩甲部機能的集積	21
乳腺集積	18
喉頭生理的集積	13
腋窩炎症性集積	10
前立腺集積	9
肛門部集積	8

一方、同時に実施されている CT による診断にお

いては、上顎洞炎、上顎粘膜貯留嚢胞が多かった(表2)。

表2:CT(PET−CT)による診断

診断名	件数
上顎洞炎	185
上顎洞粘液貯留嚢胞	127
脂肪肝	89
肺微小結節	75
肝嚢胞	68
甲状腺腫瘍	67
腎嚢胞	63
子宮筋腫疑い	35
前立腺肥大	28
腎結石	21
卵巣嚢胞性腫瘤	16
胆石	13
副脾	12

所見の登録状況では、異常なし以外の診断がつけられた1,501検査中、612例(40.8%)に所見が登録され、異常なしと診断された 528 例中、51 例(9.7%)に比較し、多く登録がされていた。また、異常なし以外の症例では、所見テンプレートを用いた異常集積の所見であったのに対し、異常なしで登録されていた所見は、「子宮摘出後」、「胆摘出後」等の既往の情報が多かった。異常なし以外で登録された所見を表3、4に示す。

表3:所見の登録状況(集積所見別)

所見	件数
異常集積:あり	480
びまん	89
限局性	425
炎症性	2
強	317
中	151
弱	79
多発	75
単発	244

表4:所見の登録状況(部位別)

部位	件数
頭頸部	149
胸部	78
縦隔	40
腹部	66
後腹膜	3
骨盤	41
骨軟部	89
皮膚	3
全身	2
特定困難	0

集積所見別では、限局性、強い、単発の修正が多かった。また、部位別では、頭頸部についての記載が多かった。

また、レポートに添付された画像は、異常なしと診断された症例で1枚から7枚、平均 1.6 枚の画像が添付されていたのに対し、異常なし以外の診断がされた症例では、1枚から9枚、平均 2.2 枚の画像が添付されていた。

さらに、PET レポーティングシステムにおいて、テンプレートシステムが導入される以前と所見の登録状況について比較を行った。テンプレート導入前の 2004 年 2 月 2 日から 2004 年9月 20 日までの 2,070 例において、異常なし以外の診断がつけられた 1,232 例中 5 例(0.4%)に所見が登録、異常なしと診断された 842 例中 4 例(0.5%)に登録されているのみで、テンプレートシステム導入後の 40.8%、9.7%に比較し、有意に低かった。

D. 考察

テンプレートシステムにより、画像診断情報を効率的に登録することが可能で、検診診断レポート作成システムに導入することで、所見情報の登録率が大幅に向上させたことが確認された。

E. 結論

診断−所見サマリー−所見テンプレートによる所見入力−サムネイル画像を用いることで、検診の際に発生する大量の画像情報を効率的にデータベース化する事が可能となり，画像情報の抽出に有用であると考えられた。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

1) 若尾文彦、他:継続性を持った病院情報システムへの展開。IT vision 8:40-44,2005
2) 飯沼 元、若尾文彦、他:消化管造影検査における FPD-DR。カレントテラピー 23:17-21,2005
3) 飯沼 元、若尾文彦、他:胃癌診断の現況と将来 放射線診断(デジタルX線診断・CT診断)。胃と腸 40(1),37-47,2005
4) 若尾文彦、他:がん診療プロセス解析システムの検討 第 25 回医療情報学連合大会論文集。494-495,2005
5) 富松英人、若尾文彦、他:大腸3D 画像の有用性 3D表示ソフトを用いて。新医療 97-100,2005
6) 石川 ベンジャミン光一、若尾文彦、他:病院情報システムデータを利用した肺悪性腫瘍手術診療プロセスの解析第 25 回医療情報学連合大会論文集。268-269,2005
7) 飯沼 元、若尾文彦他:がん取扱い規約からみた悪性腫瘍の病気診断と画像診断 結腸・直腸・肛門。臨床放射線 50,1371-1386,2005
23:17-21,2005

2. 学会発表

1) 若尾文彦、他:がん診療プロセス解析システムの検討 第 25 回医療情報学連合大会。横浜,2005
2) 石川 ベンジャミン光一、若尾文彦、他:病院情報システムデータを利用した肺悪性腫瘍手術診療プロセスの解析第 25 回医療情報学連合大会。横浜,2005

H. 知的財産権の出願・登録状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし

厚生労働科学研究費補助金(第3次がん総合戦略研究事業(分野3))
(分担)研究報告書

臨床疫学手法を用いたがん予防情報解析アルゴリズム開発に関する研究

分担研究者　小出 大介
所属　東京大学大学院医学系研究科
クリニカルバイオインフォマティクス研究ユニット・臨床疫学部門・特任助教授

研究要旨：臨床疫学的に質の高いがん予防情報の収集・分析に資するアルゴリズムを開発するため、胃がんの予防情報に焦点を絞って検討を行った。EBM の手法を適用して厳選された 11 文献を Gold Standard として、Term を切り出し、出現頻度から gastric、cancer、risk、pylori、CI、supplementation、などが良い Term と判断され、これらの組み合わせにより Sensitivity や Specificity の高いアルゴリズムの候補を 3 種類開発した。一方通常エビデンスが高いといわれる systematic review や randomized controlled trial などはそのような確証の高いスタディが少ないがん予防分野では情報が得られにくく、信頼区間を表す CI などを組み合わせるのが良いことが明らかとなった。本研究結果は、現状の知見をもとにアルゴリズムが開発され、今後の胃がん予防情報の新たな知見を組み入れる余地を残す配慮も行った。本研究の一般化可能性については他のがん種についても検討する必要があるが、胃がんに特異的である gastric や pylori などの Term を除けば応用可能と考えられた。

A. 研究目的

本分担研究は、EBM 等で注目を集めている臨床疫学の手法を用い、がん予防情報を解析するためのアルゴリズム開発をすることである。昨年度までにインフラストラクチャーとなるデータソースの構築を行い、検索の効率化などを考慮し標準的なコード化を行ってきた。最終の 17 年度は、今後臨床疫学的に質の高いがん予防情報の収集・分析に資するアルゴリズムについて実際に検討することとした。なおがん予防薬については 1 次～3 次予防の全てを含む広範な範囲で捉えた。そしてサプリメントなどの情報も取り入れている。

B. 研究方法

臨床疫学的に質の高いがん予防情報の収集・分析に資するアルゴリズムについては、Clinical Query に類する Sensitivity や Specificity を考慮した効率の良いアルゴリズムについて検討した。全てのがんについて検討するのは時間的制約から困難であるため、今回は胃がんについて行った。Gold Standard は今年度までの研究で利用した情報ソースから EBM の手法を適用して厳選された 11 文献とした。Term の切り出しは、perl でプログラムを作成して行った。文献情報のデータベース化には GetARef 6.0J を、データの集計には SAS Ver.9 を用いた。集計方法は全 Term に占める特定の Term の出現頻度、Gold Standard の 11 文献のうちで特定の Term が出現する文献の通し番号と文献数、NLM の医療分野の代表的文献データベースである PubMed における特定の Term の該当文献数を求めることにした。次に Gold Standard 文献中の Publication Type による集計、Gold Standard 文献中にふられたシソーラスである MeSH (Medical Subject Headings) による集計と行い、Sensitivity や Specificity の高い Query を実際に PubMed で試行した。

(倫理面への配慮)
本年度の研究では、特に個人に関わる情報の処理解析対象データは取り扱わなかった。

C. 研究結果

まず胃がんの予防薬情報としてリスクを下げるものとしては、アモキシリンとクラブラン酸カリウムの合剤とオメプラゾールそしてメトロニダゾールによるヘリコバクターピロリの除菌、ビタミン C(アスコルビン酸)の食事からの摂取、そしてベータカロチンやビタミン E およびセレンなどが文献から挙げられた。そしてリスクを増加する因子を避ける意味で、ヘリコバクターピロリ

の感染予防、塩分の過剰摂取の回避、新鮮な野菜や果物摂取不足を補うことなどがやはり文献から挙げられた。それらの文献をGetARefでデータベース化し(図 3)、プレインなテキストファイルとした上で、Term の切り出しを行い、出現頻度順に並べたのが表 1 である。例えば一番上の gastric について 11 論文中で 70 回ヒットし、これは全切り出しTermの2.17%であり、また Article No. で 11 論文全てに含まれることから Total 数が 11 となり、さらに調査時点(2005 年 7 月)における NLM の PubMed で gastric の用語で検索すると、124,888 件ヒットしたということである。Cancer も Gold Standard の 11 論文全てに含まれるが、PubMed でヒットする論文は 50 万件となった。さらに risk も同様で 11 論文全てに含まれるが、PubMed でヒットする論文は52万件であった。一方、pylori は含まれる論文は Gold Standard の 11 論文中 8 論文(論文No.4～11)であるが、PubMed でヒットする件数は 2 万件とそれほど多くはなかった。また信頼区間を示す CI は Gold Standard の 11 論文中 4 論文と少ないが、論文 No.2～5 と上記の pylori とは違う種類の論文を検索できることがわかり、PubMed でヒットする件数も 76,000 件程度である。CI と同様に Supplementation も Gold Standard の 11 論文中 3 論文と少ないが pylori とは別種類の論文が検索され、PubMed でヒットする件数も 36,000 件程度とそれほど多くなかった。一方検索 Term として向かないのは、7位の Treatment で、Gold Standard の 11 論文中ヒットが 4 論文と少ないにもかかわらず、PubMed でのヒットが 166 万件のような Term であった。

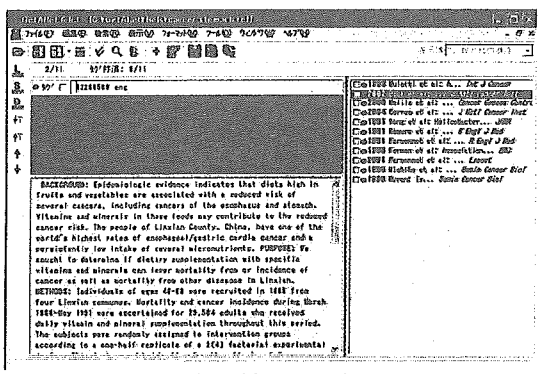

図 3. 胃がん予防薬情報のデータベース化

表 1. 胃がん予防薬情報に含まれるTermの頻度順

Term (Total 3227)	Freq	%	Article No.	Total	PubMed
gastric	70	2.17	1,2,3,4,5,6,7,8,9,10,11	11	124,888
pylori	61	1.89	4,5,6,7,8,9,10,11	8	20,087
cancer	59	1.83	1,2,3,4,5,6,7,8,9,10,11	11	501,938
risk	25	0.77	1,2,3,4,5,6,7,8,9,10,11	11	526,370
infection	24	0.74	4,5,6,7,8,9,10,11	8	429,911
carcinoma	19	0.59	3,4,6,7,9	5	267,264
treatment	17	0.53	4,5,9,10	4	1,655,274
ci	15	0.46	2,3,4,5	4	75,955
helicobacter	15	0.46	4,5,6,7,8,9,10,11	8	19,099
supplementation	15	0.46	2,3,4	3	35,682
beta-carotene	12	0.37	1,3,4	3	5,949
incidence	12	0.37	2,3,5,10	4	292,799
eradication	10	0.31	5,9,10	3	16,327
lesions	10	0.31	4,5,10	3	303,038
odds	10	0.31	6,7,8,11	4	50,248
alpha-tocopherol	9	0.28	1,3	2	8,146
associated	9	0.28	2,5,6,7,8,9	6	1,099,004
precancerous	9	0.28	4,5,10	3	4,811
prevention	9	0.28	3,4,5,9,10	5	186,574
ratio	9	0.28	5,6,7,8,11	5	299,388
risks	9	0.28	1,4,9	3	63,991

Publication Type による集計(表 2)では、Journal Article が Gold Standard の 11 論文全てに含まれたが、それ以外に際立った Type は得られなかった。MeSH による集計では(表3)、Humans という Term が Gold Standard の 11 論文全てに含まれ、次いでMaleが10論文となった。以上の結果を基に、臨床疫学的に質の高いがん予防情報の収集・分析に資するアルゴリズムとしては図4のようになり、1つめの Query で PubMed を検索した結果では 2,954 件、2つめの Query では 2,710 件で、Gold Standard の論文は全て含み、3つめの Query では Gold Standard の論文中で文献 1 が検索漏れとなりやや Sensitivity を犠牲としてしまうが、1,093 件とかなり絞り込んだより Specificity の高い検索となった。

表 2. Publication Type による集計

Publication Type	Count
Journal Article	11
Clinical Trial	5
Randomized Controlled Trial	5
Multicenter Study	1
Review	1

表 3. MeSH による集計

MeSH	Freq
Humans	11
Male	10
Middle Aged	9
Female	6
Aged	5
*Helicobacter pylori	4
Adult	4
Case-Control Studies	4
*Helicobacter pylori/immunology	3
Drug Therapy, Combination	3
Prospective Studies	3
Risk	3
Risk Factors	3
Stomach Neoplasms/epidemiology/*etiology	3

1) gastric[Title/Abstract] AND cancer[Title/Abstract] AND risk[Title/Abstract]
◆2,954 件
2) gastric[Title/Abstract] AND cancer[Title/Abstract] AND risk[Title/Abstract] AND humans[MeSH Terms] AND (Journal Article[Publication Type])
◆2,710 件
3) gastric[Title/Abstract] AND cancer[Title/Abstract] AND risk[Title/Abstract] AND (CI[Title/Abstract] OR pylori[Title/Abstract])
◆1,093 件

図 4.質の高い胃がん予防情報のためのアルゴリズム

D. 考察

まず数あるがんの中で胃がんに焦点を当てたのは、国立がんセンターのがん統計'05 によると、かつて胃がんががん死亡率のトップでよくみられるがんであったが、現在ではその死亡率も減少しつつあり、また性別と年齢を調整した罹患率でみても胃がんの罹患率が減っていることから、予防効果が奏していると考えられたからである。また食塩過多やヘリコバクターピロリが胃がんの原因の一つとの認識も強まり、その予防策も確からしさが高く、取り組みやすいということからである。これは逆にまだ原因や予防が確立していない領域のがんでは難しいということも意味する。しかしそのような別のがんの領域でもがん予防情報が確立すれば本手法は適用可能であろう。

本研究結果でみるように、Gold Standard とする論文中に含まれる率が高ければ Sensitivity も高くなり、PubMed でのヒット数が少ない Term であれば Specificity が高くなる。両者は拮抗することが多いが、可能な限り Gold Standard とする論文中に含まれる率が高く、PubMed でのヒット数が少ない Term を選べば Sensitivity も Specificity 高くすることができる。このような Term としては今回、gastric や cancer や risk などが良く、また publication type としては journal article というのが良いとの結果になった。この中で cancer や risk は他のがん種でも有効ではないかと考えられる。逆に gastric や pylori などは胃がん特異的で他のがんでは利用できないであろう。またエビデンスレベルの視点から必ずしも systematic review や randomized controlled trial などは通常は質が高いと思われる情報が得られる Term ではあるが、まだ確証情報の少ないがん予防分野では好ましくなく、せいぜい CI などの信頼区間が適当であるとの結果になった点は興味深く、この CI であれば他のがん領域でも有効な Term ではないかと思われる。

今回3つのアルゴリズムの候補を示したが、Specificity の高いアルゴリズムであれば図 4 中の 1)、Sensitivity の高いアルゴリズムであれば 3)となる。ただこれらは現状における知見をもとに作られたアルゴリズムであり、新しい知見を積極的に取り入れる仕組みにはなっておらず、そのような新たな知見も拾い上げることができるようにある程度 broad なアルゴリズムとする必要がある。また一般化可能性については他のがん種でも確認し、有効な Term を取り入れる必要がある。また医学辞書を検索にかませるなど、工夫も考えられる。PubMed は英文なので、今回は英語の Term のみでの検討となったが、PubMed 自体が主要な日本の論文も含むため、日本における情報収集にもある程度役立つものと考えられる。

E. 結論

臨床疫学的に質の高いがん予防情報の収集・分析に資するアルゴリズムを開発するため、胃がんの予防情報に焦点を絞って検討を行った。その結果 gastric、cancer、risk、pylori、CI、supplementation、などが良い Term と判断され、これらの組み合わせにより Sensitivity や Specificity の高いアルゴリズムの候補を 3 種類開発した。一方通常エビデンスが高いといわれる systematic review や randomized controlled trial などでは情報が得られにくく、信頼区間を表す CI などを組み合わせるのが良いことが明らかとなった。胃がんに特異的で

ある gastric や pylori などの Term を除けば他のがん種にも応用可能と示唆された。

G. 研究発表

1. 論文発表

1) 小出大介, 陳俊成, 小山博史: がん予防薬情報データベースの開発. 臨床薬理. 36(Suppl):s280. 2005.

2) Daisuke Koide, Edward Peskin: New uses for computer in medical education, clinical practice, and patient safety in the US and Japan. Progress in Informatics. 1(2). P.3-15. 2005.

3) 小出大介: 医療安全の実践である薬剤疫学-その発展に必要なファクター-. 薬剤疫学. 10(Suppl): s32-3. 2005.

4) 小出大介: 日本発のエビデンス産生と教育プログラム. Japanese Pharmacology & Therapeutics. 33(5): 413-415.2005.

2. 学会発表

1) 小出大介, 陳俊成, 小山博史:がん予防薬情報データベースの開発. 第26回 臨床薬理学会 年会, 2005 年 12 月 3 日, 別府.

2) 小出大介: 医療安全の実践である薬剤疫学-その発展に必要なファクター-. 第 11 回 日本薬剤疫学会 学術総会. 2005 年 11 月 13 日, 福井.

(発表誌名巻号・頁・発行年等も記入)

H. 知的財産権の出願・登録状況(予定を含む。)

1. 特許取得

なし

2. 実用新案登録

なし

3.その他

なし

厚生労働省科学研究費補助金(第3次対がん総合戦略研究事業)
平成17年度 (分担)研究報告書

研究課題名:「がん予防薬の臨床試験支援用情報システム構築に必要となる機能仕様に関する調査研究」

分担研究者氏名: 山本精一郎
所属:国立がんセンターがん予防・検診研究センター

研究要旨:がん予防薬の開発に向けた臨床試験のシステム構築について必要となる機能仕様を調べるために、がん治療の臨床試験分野で行われているシステムを調査した。がん予防法開発においても、治療法開発と同様な臨床試験組織を作ることによって、臨床試験を行うことができる。しかし、サンプルサイズの増大、エンドポイントの把握、対象者のコンプライアンスの低下など治療の研究よりも困難な点も多く存在するため、治療開発にもまして堅固な組織化が必要である。

代替療法使用のがん予防へ影響や、生活習慣や代替療法のがん予防・再発への効果を調べるための大規模コホートを設立する第一歩として、代替療法使用を把握する質問票を開発した。さらに、この質問票ならびにコホート設定の実現可能性を調べるパイロット研究を国立がんセンターにおいて実施中である。このパイロット研究において、乳がん患者の代替療法の使用実態も明らかになると考えられる。

A.研究目的

がん予防は、がんの治療や早期発見に比べて、がん克服のためのもっとも効率のよい方法である。がんを予防するための方法を開発するためには、がん予防薬や生活習慣改善などががん予防を減らすことを立証するために臨床試験を行うことが必要である。しかしながら、本邦においてはがん予防薬開発の分野ではがん治療の分野のように系統的に臨床試験は行われていない。そこで、がん治療分野の臨床試験システムを検討することにより、がん予防薬開発の臨床試験システムに必要となる機能仕様の検討を行うことを目的とする。このシステムを実装することにより、がん予防を効率的に行うための方法を系統的に開発することができる。

また、がん患者や健康に関心のある人々は、がんやその再発を防ぐために、代替療法や生活習慣の改善をしようとするが、これらに対するエビデンスは非常に少ない。なぜなら、代替療法使用のがん予防へ影響や、生活習慣や代替療法のがん予防・再発への効果を調べることを目的としてわが国で行われた臨床試験やコホート研究はほとんどないからである。来年度以降、乳がん患者に対し、生活習慣や代替療法使用の有無が乳がん予防・予後に与える影響を調べる大規模コホート研究を行うことを計画しているため、そこで本年度は、この研究で曝露評価として用いることのできる質問票の開発およびパイロット研究を行う。このパイロット研究によって、日本における代替療法使用の実態が明らかになるとともに、将来のコホート研究によってがん患者がほしい代替療法についての情報をシステマティックに提供することができるようになる。

B.研究方法

がん予防薬の開発に向けた臨床試験のシステム構築について必要となる機能仕様を調べるために、がん治療の臨床試験分野で行われているシステムを調査する。

生活習慣や代替療法使用のがん予防や再発を調べる研究を行うために、過去に行われた生活習慣の疫学研究や代替療法の調査研究を下に、このコホート研究で用いる生活習慣に関する質問票と代替療法に関する質問票を開発する。また、乳がん患者を対象としてこれらの生活習慣質問票と代替療法質問票の実施可能性を調べるパイロット研究を実施する。

(倫理面への配慮)

パイロット研究は国立がんセンターの施設倫理委員会の承認を受け、ヘルシンキ宣言ならびに疫学研究の倫理指針に沿って行う。

C.研究結果

臨床試験には、それが評価するmodalityが治療であっても予防であっても、研究者グループ、データセンター、独立して研究をモニターする委員会機能が必要であることがわかった。研究者グループと独立委員会は研究者で構成することができるため、データセンターを自前で持つか委託する必要があり、いずれにしても試験あたりデータマネジメント費用として数百万円(同時に行う試験の数によって変動する)が必要となることが判明した。研究者グループ、データセンター、独立委員会のどの部分が担ってもよいが、研究を円滑に進めるには3つの部門間を調整する事務局機能が重要であることもわかった。事務局は臨床試験について通暁している必要があり、ここの部分がしっかりしていないと業務委託も円滑に行かないことが判明した。

生活習慣や代替療法使用のがん予防や再発を調べる研究を行うために、過去に行われた生活習慣の疫学研究や代替療法の調査研究を下に、このコホート研究で用いる生活習慣に関する質問票と代替療法に関する質問票を開発した。乳がん患者を対象としてこれらの生活習慣質問票と代替療法質問票の実施可能性を調べる研究が、国立がんセンター施設倫理委員会に承認され、現在調査実施中である。

D.考察

がん予防の臨床試験は治療の臨床試験を行うよりもより大きなサンプルサイズを必要とするため、しっかりした組織が必要となるが、組織の形態についてはがん治療の臨床試験組織と同様な形態で運営できるといえる。予防研究のほうが治療研究よりもエンドポイント把握が困難になる、対象者のコンプライアンスを維持するのが難しい、サンプルサイズが大きくなるなどの問題があり、世界的に見ても成功例はなかなかないことがわかった。

代替療法の質問票の開発は、これらについてのがん予防情報の正確な解釈を支援するための科学的評価に必要な調査項目の設定につながるといえる。また、この研究結果から乳がん患者における生活習慣と代替療法使用の有無の現状、対象者の得たい情報についても知ることができる。

E.結論

がん予防法開発においても、治療法開発と同様な臨床試験組織を作ることによって、臨床試験を行うことができる。しかし、サンプルサイズの増大、エンドポイントの把握、対象者のコンプライアンスの低下など治療の研究よりも困難な点も多く存在するため、治療開発にもまして堅固な組織化が必要である。

代替療法使用のがん予防へ影響や、生活習慣や代替療法のがん予防・再発への効果を調べるための大規模コホートを設立する第一歩として、代替療法使用を把握する質問票を開発した。さらに、この質問票ならびにコホート設定の実現可能性を調べるパイロット研究を国立がんセンターにおいて実施中である。このパイロット研究において、乳がん患者の代替療法の使用実態も明らかになるであろう。

F.健康危険情報

G.研究発表

1.論文発表

1) Abe M, Ohira M, Kaneda A, Yagi Y, Yamamoto S, Kitano Y, Takato T, Nakagawara A, and Ushijima T. CpG island methylator phenotype is a strong determinant of poor prognosis with neuroblastomas. Cancer Research 2005;65(3):828-34.
2) Hanaoka T, Yamamoto S, Sobue T, Sasaki S, Tsugane S, for the Japan Public Health Center-based prospective study on cancer and cardiovascular diseases Group. Active and passive smoking and breast cancer risk: observational cohort study. Int J Cancer 2005;114(2):317-22.
3) Horstmann E, McCabe M S, Grochow L, Yamamoto S, Rubinstein L, Budd T, Shoemaker D, Emanuel E J, Grady C. Risks and Benefits of Phase 1 Oncology Trials: 1991-2002 New Engl J Med 2005;352:895-904.
4) Kodera Y, Sasako M, Yamamoto S, Sano T, Nashimoto A, Kurita A on behalf of the Gastric Cancer Surgery Study Group of Japan Clinical Oncology Group. Identification of risk factors for the development of complications following extended and super-extended lymphadenectomies for gastric cancer. Br J Surg 2005;92:1103-9.
5) Tsubono Y, Otani T, Kobayashi M, Yamamoto S, Sobue T, and Tsugane S for

the JPHC Study Group. No Association between Fruit or Vegetable Consumption and the Risk of Colorectal Cancer in Japan: JPHC Study. Br J Cancer. 2005;92(9):1782-4.
6) Ishikura S, Tobinai K, Ohtsu A, Nakamura S, Yoshino T, Oda I, Takagi T, Mera K, Kagami Y, Itoh K, Tamaki Y, Suzumiya J, Taniwaki M and Yamamoto S. Japanese Multicenter Phase II Study of CHOP Followed by Radiotherapy in Stage I-II1, Diffuse Large B-cell Lymphoma of the Stomach. Cancer Science 2005;96, 6.
7) Kabuto M, Yamamoto S., et al. A Case-Control Study of Childhood Leukemia and Residential Power-Frequency Magnetic Fields in Japan. Int J Cancer (in press)
8) Takano T, Ohe Y, Sakamoto H, Tsuta K, Matsuno Y, Tateishi U, Yamamoto S, Nokihara H, Yamamoto N, Sekine I, Kunitoh H, Shibata T, Sakiyama T, Yoshida T, Tamura T. Epidermal growth factor receptor gene mutations and increased copy numbers predict gefitinib sensitivity in patients with recurrent non–small-cell lung cancer. J Clin Oncol 23(28):6829-37, 2005.
9) Tateishi U, Hasegawa T, Yamamoto S, Yamaguchi U, Yokoyama R, Kawamoto H, Satake M, Arai Y. Incidence of multiple primary malignancies in a cohort of adult patients with soft tissue sarcoma. Jpn J Clin Oncol. 2005;35(8):444-52.
10) Ishihara J, Yamamoto S, Iso H, Inoue M, Tsugane S. Validity of a self-administered food frequency questionnaire (FFQ) and its generalizability to the estimation of dietary folate intake in Japan. Nutrition Journal 2005;4:26.
11) Yamamoto S, Tsugane S. Soy and breast cancer prevention : SOY in Health and Disease Prevention, Sugano M (Ed.), CRC Press, Boca Raton, 2005.
12) Hashimoto K, Yamamoto S. Learning from a visit to the JNCI editorial office. Jpn J Clin Oncol 2005;35:162-164.

2.学会発表

1) Matsumura Y, Hayashi K, Liang CY, Yamaji Y, Marui E, Yamamoto S, Sugishita C, Sugai Y. Relationship between alcohol consumption and cognitive function in the community living elderly people in Japan. Health and Nutrition, Japan. IEA, August, 2005
2) Marui Eiji, Liang Chun Yu, Yamaji Yoshio, Matsumura Yasuhiro, Hayashi Kunihiko, Yamaji Yoshio, Yamamoto Seiichiro, Sugai Yuichi, Sugisita Chieko. Daily Life Styles and Intellectual Functions in Community-living Elderly People. IEA, Augsut, 2005.

H.知的財産権の出願・登録状況

1.特許取得 なし
2.実用新案登録 なし
3.その他 なし

厚生労働科学研究費補助金(第3次対がん総合戦略事業)
(分担)研究報告書

研究課題名:「がん予防薬の薬物動態関連酵素とSNPとの関連データベースの開発」
分担研究者氏名: 日紫喜 光良
所属:産業技術総合研究所 生物情報解析研究センター

研究要旨
がん予防薬と代謝酵素等の遺伝子・タンパク質との既知の相互作用(物理的な結合だけでなく、他のメカニズムを介して結果的に遺伝子の機能を調節するものを含む)をテキストから抽出してデータベース化するための情報収集支援システムを構築した。薬剤の分子レベルでの作用についての情報とその表現が多様であるため、当データベースの作成に必要な情報の収集には自動的な情報抽出の手法よりはむしろ人手によるアノテーション(テキストから、必要な情報を有する部分を読み取り、含まれる情報を解読して、テキスト中の根拠とともに記録すること)を採用した。そのために、遺伝子・遺伝子産物辞書を作成し、薬剤と遺伝子・タンパク質との間の分子レベルでの作用を表す用語を定義するとともに、それらの用語を自動的にテキスト中に同定して強調表示する機能をもったアノテーションシステム(アノテータ(アノテーションをおこなう人員)の作業を容易にするシステム)を開発した。収集した情報を「GenoCache」(Genomics and Cancer Chemoprevention)データベースに収納し、遺伝子を多型情報や既知のパスウェイに関連づけた。

A. 研究目的

ほぼ完全に解読されたヒトゲノム配列に加えて、その他のゲノム情報ー例えば遺伝子配列の多様性ーについてのデータが充実しつつあり、それらを導入することでがんの化学予防についてのさまざまな知識が統合されると期待される。現時点では、そのような視点から構成されたがんの化学予防についてのWebサイトはまだないと考えている。

本研究ではがん予防薬と代謝酵素等の遺伝子・タンパク質との間の相互作用(物理的な結合だけでなく、他のメカニズムを介して結果的に遺伝子の機能を調節するものを含む)についての情報の収集をおこなうことによって、がん予防薬の分子メカニズムの研究に寄与するリソースを構築する。そのために、既知のがん予防薬を含む薬剤のリスト、および、遺伝子・タンパク質のリストを作成するとともに、それらを利用して研究者ががん予防薬と遺伝子・タンパク質との間の相互作用についての情報を収集することを支援するシステムを開発する。

B:研究方法

遺伝子辞書の作成:パブリックドメインのヒト遺伝子データベース(HUGO、LocusLink ならびに UniProt)を用いて、遺伝子の ID とその代表的な名称、ならびに代替の名称を収集した。それらのうち、NCBI の dbSNP に対応のあるものを用い、およそ 23,000 の遺伝子座に対する遺伝子名を収集した。

物質辞書の作成:がん予防物質についての総説を収集し、そこの表から物質名を収集した。次に、その名称を MedIDPlus データベースで検索して、対応する物質の CAS registry number を取得した。なお、UMLS Metathesaurus には、用語の意味カテゴリー(Semantic Type)が付与されているが、がん予防薬をあらわすものはない。もっとも近いものは、たとえば次の Semantic Type をもつものと考えられる:T121 (Pharmacologic Substance)ならびに T200 (Clinical Drugs)。

分子作用についての語彙の作成:遺伝子転写物やタンパク質の量の増減をあらわすために使うことが可能な語(increase, decrease, up-regulate, down-regulate 等)、タンパク質の作用の増強・減弱をあらわすために使うことが可能な語(inhibit, enhance 等)、タンパク質の作用の種類をあらわす語(phosphorylate 等)についての語彙を作成した。

がん予防物質について記述している可能性のある文書の取得:PubMed データベースをがん予防物質の CAS registry number で検索することによって、がん予防物質について記述している可能性のある文書を取得した。

がん予防薬情報アノテーションシステムの作成:予備的な調査により、がん予防物質と遺伝子との関係は種類が多様であり自動的な収集

が困難であることが示唆されたので、文章中の情報の収集を人手でおこなう方針を採用した。その作業を能率化するために、テキスト中にあり情報のてがかりとなる語句を強調して認識しやすくするとともに、登録作業を簡略化することが可能な「がん予防薬情報アノテーションシステム」を作成した。このシステムは、WWW インターフェースを介して、作業者（アノテータ）が、がん予防薬情報を参照・登録する Web サーバアプリケーションとして構成された。

B. 研究結果

がん予防薬情報アノテーションシステム：がん予防薬情報アノテーションシステムは、以下のサブシステムから成る。アノテータは、「アノテータ管理サブシステム」によって、がん予防薬情報アノテーションシステムにアクセスする。アノテータからの要求を受けたがん予防薬情報アノテーションシステムは、「予防薬情報管理サブシステム」に登録されたがん予防薬を対象として、「予防薬関連論文収集サブシステム」、「予防薬・遺伝子・効果辞書マッピングサブシステム」によって、論文の収集および論文に記載された予防薬に関する情報を収集し、「予防薬情報アノテーションサブシステム」によって、収集された情報を参照しつつ、論文からがん予防薬ー遺伝子関係情報の抽出・編集・蓄積を実施する。また、「SNP情報収集サブシステム」によって、収集された遺伝子上のSNPを探索し、関連情報と共にがん予防薬と関連付けて蓄積する。

予防薬情報アノテーションサブシステムのクライアント側のユーザインターフェースは図1に示すとおりである。ユーザは、次のような段階を経てアノテーションをおこなう。

1. がん予防薬を選択する
2. 文献をリストから選択する
3. 文をスクロールしながら検討する
4. 文が薬剤の分子レベルの作用について言及しているか判断する

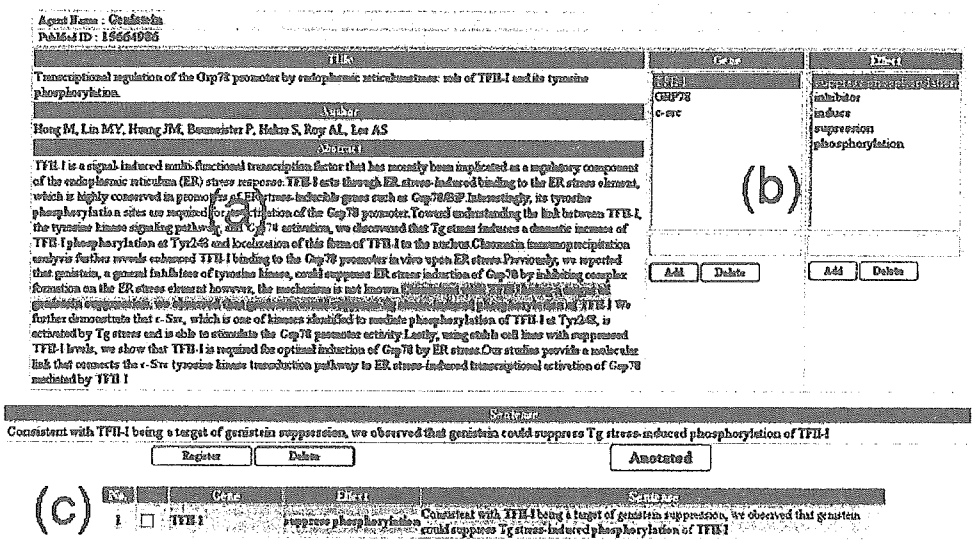

図1. 予防薬情報アノテーションサブシステム画面。(a) 個々の文をスクロールすると、注目する文がハイライトされる。遺伝子/タンパク質名ならびに 効果を表現する用語がマークアップしてある。(b) このテキスト中に同定された遺伝子名称ならびに分子レベルでの作用を表現する単語またはフレーズのリスト。注目する文をスクロールするたびに、もしもその中に含まれる用語があれば、選択状態としてハイライトされる。画面左下方の「Register」ボタンをクリックすると、選択された用語のペアが記録される。(c) 薬剤と遺伝子・遺伝子産物間に作用があると判断された文が表に取り込まれる。

GenoCache データベース：このように蓄積されたデータを検索可能な Web サイト「 GenoCache 」 (Genomics and Cancer Chemoprevention) を構築中である。

GenoCache データベースの構成要素は大きく分けて物質 DB 検索結果、文献 DB 検索結果、物質の効果の3種類の表から成る。物質 DB 検索結果と個々の文献 DB 検索結果とは CAS Registry Number でリンクされ、文献 DB 検索結果と物質の効果とは PubMed ID でリンクされる。標的遺伝子あるいはタンパク質、ならびにその標的への効果は、物質の効果の表に記述される。標的への効果としては、タンパク質あるいは mRNA 量の増減、リン酸化/脱リン酸化の促進/阻害、酵素反応の基質としての作用、構想反応の阻害薬としての作用、転写因子の転写活性化促進/阻害などがある。また、標的遺伝子/タンパク質は、Entrez Gene ID や Genbank の Accession number を介して dbSNP 等の公共データベースにリンクされる。このサイトは、物質ならびに遺伝子から検索可能である。

C. 考察

発がんあるいはがんの阻害については、毒性学的観点から個体あるいは組織レベルの多くの知見が蓄積されている。一方で、ほぼ完全に解読されたヒトゲノム配列に加えて、その他のゲノム情報－例えば遺伝子配列の多様性－についてのデータが充実しつつあり、それらを導入することでがんの化学予防についてのさまざまな知識が統合されると期待される。現時点では、そのような視点から構成されたがんの化学予防についてのWebサイトはまだないと考えている。

薬剤が作用する、あるいは薬剤を代謝するタンパク質の同定のためには、文章から同定するよりは、もし可能ならば KEGG あるいは BIND といった、分子間の作用についてのデータベースから取得することができれば、ある程度正確な情報がより簡単に入手できるであろう。しかし、どちらのデータベースでも、検討の対象としたがん予防物質の多くについて、同定が困難であった。その他 PubMed, PubChem, OMIM 等のデータベースの検索をおこない比較検討した結果、ChemIDplus が、薬剤の一意な ID(CAS registry number)を得るためにもっとも便利であると判断し、それを手がかりとして PubMed を検索するという方法を選択した。

GenoCache データベースにおいては、アノテーションの効率を検討しながら、個々の物質の効果に関連した情報（例えば実験に使用した細胞やがん予防物質の使用方法など）を拡張することも考えられる。

D. 結論

がん予防薬の遺伝子・タンパク質への作用についての情報を PubMed テキストからデータベース化するためのアノテーションシステムを開発し、情報を収集している。それらの情報は GenoCache (Genomics and Cancer Chemoprevention) データベースに収納された。このデータベースは、がん予防薬の分子メカニズムの研究に寄与するリソースとなることが期待される。また、本研究で収集されるような多様な情報の収集にあたっては、情報へのてがかりとなる用語を自動的に同定したあと、それを情報の抽出を自動的に行う情報抽出に用いるよりはむしろ、人手のアノテーションのサポートに用いるアプローチのほうが適していると考えられた。

E. 研究発表:
なし

F. 論文発表:
なし

G. 学会発表:

1. 日紫喜光良、小山博史. がん予防薬ターゲット探索のためのサイト「GenoCache」の開発. 第12回日本がん予防研究会. 岐阜. 2005/7/14
2. Hishiki, T., Oyama, H.: GenoCache: a Genomics and Cancer Chemoprevention Portal. American Medical Informatics Association (AMIA) 2005 Annual Symposium. Washington D.C., 2005/10/25
3. 日紫喜光良、小山博史:がん予防薬ターゲット探索のためのサイト「GenoCache」の開発. 第 25 回医療情報学連合大会. 横浜. 2005/11/25

H. 知的財産権の出願・登録状況:
なし