Identifying Risk Factors for Metabolic Syndrome in Biomedical Text

Marcelo Fiszman MD PhD, Graciela Rosemblat PhD, Caroline B. Ahlers MD, Thomas C. Rindflesch PhD
National Library of Medicine, Bethesda, MD

Identifying risk factors and biomarkers for diseases is an important aspect of biomedical research. However, much of the underlying information resides in the research literature and is not available in executable form. We propose a methodology based on automatic semantic interpretation (using SemRep) to capture risk factors and biomarkers for diseases asserted in MEDLINE citations. In this initial study, we focus on metabolic syndrome. The performance of SemRep in identifying risk factors and biomarkers for this disorder was 53% recall (CI, 44% to 62%) and 67% precision (CI, 62% to 72%). We discuss how the information captured could assist clinicians in finding current and new risk factors for metabolic syndrome as well as diseases predisposed by this disorder. The availability of this information in executable form can support guideline development and the timely translation of biomedical research into improvements in quality of patient care.

INTRODUCTION

Research on finding risk factors and biomarkers (substances) that predict disease is pervasive in biomedical research. Recently, much attention has been paid to metabolic syndrome [1], a very common, multi-factorial condition that has been implicated in the pathway of several diseases. It has been estimated that over fifty million Americans have metabolic syndrome [2]. As the available literature on this disorder grows, it becomes increasingly useful to develop and maintain knowledge resources that identify risk factors as well as the conditions associated with metabolic syndrome. Most of this information resides in the biomedical research literature and is not immediately available in executable form accessible to advanced information management systems.

There is little research attempting to develop automatic methods for extracting and compiling risk factors for diseases, although Matsunaga uses machine learning techniques to identify genes associated with metabolic syndrome [3]. We propose a method based on semantic interpretation for automatically finding risk factors and biomarkers of disease as they appear in the biomedical literature. Although, the method is applicable to any disease, the focus of this paper is metabolic syndrome. We expanded an existing semantic interpreter called SemRep [4, 5]. In this paper we describe enhancements to SemRep, perform a preliminary evaluation, and discuss clinical implications from the perspective of metabolic syndrome

BACKGROUND

Metabolic syndrome

Metabolic syndrome is a multiplex of clustered risk factors for cardiovascular disease, but evidence is mounting that it is associated with such different conditions as polycystic ovary syndrome [6], Alzheimer's disease [7], and neoplasms [8]. Currently, risk factors and markers for metabolic syndrome are defined by The National Cholesterol Education Program Adult Treatment Panel III (ATP III) [9, 10] as: abdominal obesity, elevated triglycerides, reduced high density lipoprotein (HDL) cholesterol, elevated blood pressure, and elevated glucose. These risk factors are not static and increasingly other factors have been linked to metabolic syndrome.

SemRep

SemRep identifies semantic predications in biomedical text. During processing, a partial syntactic analysis depends on lexical look-up in the SPECIALIST lexicon and the MedPost tagger [11]. MetaMap [12] matches noun phrases to concepts in the UMLS Metathesaurus and determines the semantic type for each concept. Argument identification is based on dependency grammar rules that enforce general syntactic constraints. Indicator rules map syntactic phenomena (verbs, nominalizations, etc.) to predicates in the UMLS Semantic Network, which imposes semantic validation for the relationships constructed.

For example, in (1), an indicator rule links the verb *induce* with the Semantic Network predicate CAUSES. Since the semantic types of the syntactic arguments ('Lipid' and 'Pathologic Function') identified for *induce* in this sentence match the corresponding semantic types in the Semantic

Network, the predication in (2) is constructed.

(1) Chronic elevations of free fatty acids (FFA) induce insulin resistance.

(2) Fatty Acids, Nonesterified CAUSES Insulin Resistance

Crucial to the method employed by SemRep are indicator rules and allowable semantic type arguments for a given relation from the Semantic Network. In a recent paper [13], SemRep was extended to the pharmacogenomics domain. The enhancements required modifications to the Semantic Network. For example, the predicate PREDISPOSES was defined to accommodate genetic etiology of disease, and semantic types allowed as semantic arguments of PREDISPOSES were added. Examples of subject semantic types for this predicate are 'Gene or Genome' and 'Biologically Active Substance'. Object semantic types include those from the semantic group Disorder [14].

In this paper, the relation PREDISPOSES was used to interpret disease risk factor ontological predications of the form "Risk factor PREDISPOSES Disease." This required the addition of indicator rules mapping to PREDISPOSES and semantic types to serve as semantic arguments. We describe these additions in the next section.

METHODS

Adding indicator rules

In order to find indicator rules for disease risk factors, a query was issued to PubMed with the MeSH heading "Risk Factors." Sentences that stated risk factors for diseases were isolated for linguistic analysis. The syntactic patterns listed below (3) were the most productive and form the basis for additional indicator rules in SemRep. The patterns are listed from the most to least frequent. "{*Be*}" means that a form of *be* is optional, and slash indicates disjunction.

(3) Patterns for indicator rules

a. RiskF {be}*risk factor for/of* Disorder
b. RiskF *risk for/of* Disorder
c. RiskF *predict* Disorder
d. RiskF *marker of* Disorder
e. RiskF *determinant of/for* Disorder
f. RiskF *contribute* Disorder
g. RiskF *promote* Disorder

"RiskF" and "Disorder" represent risk factors and disease terms, respectively. Verbs (c, f, and g) are listed in infinitival form; however, when constructing indicator rules all possible inflections are included. In addition, nominalization forms (such as *prediction* and *contribution*) were added, along with their prepositional argument cues. Thirty two indicator rules were added. For example the sentence in (4) is interpreted as the predication in (5), because of the indicator rule *prediction* and the prepositional argument cue *of.*

(4) Some experts propose C-reactive protein as a screening tool for prediction of cardiovascular disease.

(5) C-reactive protein PREDISPOSES Cardiovascular Disease

Adding types as semantic arguments

After further analysis of the risk factor sentences, the semantic types in (7) (categorized in groups) were allowed to serve as subject semantic arguments of the predication in (6). Permissible object arguments remain the semantic types of the UMLS semantic group Disorder.

(6) {Risk Factor} PREDISPOSES {Disorder}

(7) Risk Factor:
('Gene or Genome')
('Amino Acid, Peptide, or Protein', 'Antibiotic', 'Biologically Active Substance', 'Carbohydrate', 'Eicosanoid', 'Element, Ion, or Isotope', 'Enzyme', 'Hazardous or Poisonous Substance', 'Hormone', 'Immunologic Factor', 'Inorganic Chemical', 'Lipid', 'Neuroreactive Substance or Biogenic Amine', 'Nucleic Acid, Nucleoside, or Nucleotide', 'Organic Chemical', 'Organophosphorus Compound', 'Pharmacologic Substance', 'Steroid', 'Vitamin')
('Disease or Syndrome', 'Finding', 'Injury or Poisoning', 'Mental or Behavioral Dysfunction', 'Neoplastic Process', 'Pathologic Function', 'Sign or Symptom')
('Clinical Attribute', 'Organism Attribute')
(Food')
('Daily Recreational Activity', 'Individual Behavior')
('Laboratory or Test Result')

Accommodating polarity

The syntactic patterns in (3) are frequently expressed in text modified by polarity terms (8 and 9) and negation (10). SemRep is able to interpret negation and generates (11) for (10).

(8) Intake of acetylsalicylic acid is associated with a **decreased** risk of colorectal cancer

(9) Moderate alcohol consumption was independently associated with **lower** risk for cardiac mortality

(10) Former pregnancies are **not** a risk factor for giant cell arteritis.

(11) Pregnancy NEG_PREDISPOSES Giant Cell Arteritis

In order to interpret sentences such as (8) and (9), we added polarity words (e.g. *lower*, *decreased*, and *reduced*) to SemRep's negation machinery. These sentences are thus interpreted as though they were negated.

Evaluation

We used two test sets to evaluate the performance of SemRep in identifying risk factors for metabolic syndrome. The first was used to calculate recall and the second to calculate precision. For both sets, a query was issued to PubMed using the MeSH headings "Metabolic Syndrome X" and "Risk Factors," limited to citations with abstracts in English.

The first one hundred sentences from the first set that expressed risks factors for metabolic syndrome were scrutinized by one of the authors (GR) and annotated with the correct PREDISPOSES predication. SemRep output was compared against this set and recall was measured.

To calculate precision, SemRep output for the first three hundred and fifty sentences from the second set that produced PREDISPOSES predications were assessed for correctness by two of the authors (GR and CBA). 95% confidence intervals (CI) were calculated for both recall and precision.

RESULTS

Out of the 100 sentences in the pre-tagged (recall) sample, the total number of PREDISPOSES predications marked was 118. SemRep missed 56 of these, resulting in recall of 53%, with a 95% confidence interval ranging from 44% to 62%.

SemRep produced 396 PREDISPOSES predications from the 350 sentences in the second (precision) set. Of these, 122 were marked as incorrect by the evaluators. Therefore, precision was 67%, with 95% confidence interval ranging from 62% to 72%.

DISCUSSION

The enhancement of SemRep to identify risk factors for diseases produced encouraging results for the MEDLINE citations on metabolic syndrome. In the sentences used to calculate recall, most mistakes were due to missing indicator rules. Two sentences were responsible for 13 false negatives. For example, SemRep missed all nine risk factors in (12) because it does not have *increased odds* as an indicator rule.

(12) Older age, postmenopausal status, Mexican American ethnicity, higher body mass index, current smoking, low household income, high carbohydrate intake, no alcohol consumption, and physical inactivity were associated with **increased odds** of the metabolic syndrome.

In analyzing the results of processing the precision set, as in previous evaluations of SemRep, inadequate handling of word sense ambiguity accounted for the majority of the false positives. For example, in (13), SemRep wrongly interpreted *gap* as an acronym for a protein, rather than a statistical method of biomedical analysis, and thus produced the false positive predication (14).

(13) Gap analysis of pediatric reference intervals for risk biomarkers of cardiovascular disease and the metabolic syndrome.

(14) GTPase-Activating Proteins PREDISPOSES Metabolic syndrome

In this paper, we focused on metabolic syndrome to illustrate the methods used. However, analysis of text involving other diseases indicates that the language expressing risk factors is not peculiar to a particular disease. It appears that the method applies equally effectively to disorders other than metabolic syndrome, although we have not yet demonstrated this.

Our results show that we need additional indicator rules in order to increase recall for finding PREDISPOSES predications. We are currently working on this issue and also on other problems revealed by the error analysis.

Implications for metabolic syndrome

We conducted a PubMed search with MeSH heading "Metabolic Syndrome X" limited to citations with abstracts in English. This search yielded 2745 citations, which we processed with SemRep followed by an automatic summarization system [15, 16]. Our intention was to investigate and isolate information in current research about metabolic syndrome.

SemRep retrieved 143 unique risk factors for metabolic syndrome and 53 diseases predisposed by this disorder. The fifteen most frequent are presented in Table 1, risk factors in column 1 and diseases in column 2. The current risk factors that constitute diagnostic criteria for metabolic syndrome are marked with an asterisk. In addition to finding these, SemRep was able to identify risk factors often discussed but not currently considered as diagnostic of the syndrome. Two of these, C-reactive protein and stress, are noteworthy.

Table 1 – Most frequent risk factors for and diseases predisposed by metabolic syndrome in MEDLINE.

Risk Factors	Diseases
Obesity*	Cardiovascular Diseases
Waist circumference*	Diabetes Mellitus
Uric Acid	Coronary heart disease
Sedentary	Atherosclerosis
Hypertriglyceridemia*	Kidney Failure
Body mass index*	Myocardial Infarction
Increase in blood pressure*	Hypertensive disease
Hyperglycemia*	Cerebrovascular accident
C-reactive protein	Ischemic stroke
High Density Lipoproteins*	Heart failure
Low Birth Weights	Vascular Diseases
Stress	Myocardial Ischemia
Cigarette Smoking	Hypogonadism
Adiponectin	Erectile dysfunction

There is considerable controversy in the biomedical literature whether C-reactive protein is or is not a risk factor for metabolic syndrome [17]. The International Diabetes Federation has proposed a new definition for this disorder that includes C-reactive protein, but this is not widely accepted [18].

Stress has not traditionally been considered a risk factor for metabolic syndrome. However, recent studies [19, 20], the latter in the British Medical Journal provide evidence that chronic stress is an important risk factor for metabolic syndrome. In addition, the authors propose a biological link between the psychological stress of everyday life and cardiovascular disease.

In considering the diseases predisposed by metabolic syndrome (Table 1, second column), cardiovascular disease is, not surprisingly, the most common. However, diseases such as erectile dysfunction [21] and hypogonadism [22] have also been discussed. Less frequent diseases that are not listed in Table 1, such as rheumatoid arthritis [23] and sensorineural hearing loss [24], were also found.

Summarizing the SemRep output for metabolic syndrome highlighted some interesting current research, for example, the association between hepatitis C virus infection and metabolic disease. One form of liver disease known as non-alcoholic steatohepatitis is intractable and progressive and is considered to be one of the phenotypic features of metabolic syndrome. The fact that hepatitis C virus infection causes steatohepatitis and induces metabolic syndrome is surprising. When looking at the summary we found direct links between hepatitis C virus infection, insulin resistance and metabolic syndrome [25, 26]. The authors state that hepatitis C virus infection may directly disturb insulin signaling pathways independent of hepatic steatosis. They conclude that the infection must be viewed not only as a liver disease but also as a metabolic disorder.

CONCLUSION

We investigate the use of semantic interpretation for identifying disease risk factors and biomarkers in MEDLINE citations. We expanded an existing semantic processor and automatically identified risk factors for metabolic syndrome and diseases predisposed by this disorder in the biomedical research literature. Our results suggest that the information extracted might be useful for biomedical researchers in understanding the connections among the variables associated with this complex syndrome. It could also be potentially useful for developing clinical guidelines to improve patient care. Although the paper focuses on metabolic syndrome, the methodology proposed is applicable to any disease.

Acknowledgments

This study was supported in part by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine.

References

1. Metabolic syndrome. Nat Med. 2006 Jan;12(1):26-80.

2. Galassi A, Reynolds K, He J. Metabolic syndrome and risk of cardiovascular disease: a meta-analysis. Am J Med. 2006 Oct;119(10):812-9.

3. Matsunaga T, Muramatsu MA. Knowledge-based computational search for genes associated with the metabolic syndrome. Bioinformatics. 2005 Jul 15;21(14):3146-54.

4. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J of Biomed Inf. 2003 Dec;36(6):462-77.

5. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In Chen, Fuller, Hersh, and Friedman, eds. Medical informatics: Knowledge management and data mining in biomedicine. Springer, 2005, pp. 399-422

6. Cussons AJ, Stuckey BG, Watts GF. Cardiovascular disease in the polycystic ovary syndrome: new insights and perspectives. Atherosclerosis. 2006 Apr;185(2):227-39.

7. Razay G, Vreugdenhil A, Wilcock G. The metabolic syndrome and Alzheimer disease. Arch Neurol. 2007 Jan;64(1):93-6.

8. Chiu HM, Lin JT, Shun CT, et al. Association of metabolic syndrome with proximal and synchronous colorectal neoplasm. Clin Gastroenterol Hepatol. 2007 Feb;5(2):221-229.

9. Grundy SM. Metabolic syndrome: a multiplex cardiovascular risk factor. J Clin Endocrinol Metab. 2007 Feb;92(2):399-404.

10. McNeill AM, Rosamond WD, Girman CJ, et al. The metabolic syndrome and 11-year risk of incident cardiovascular disease in the atherosclerosis risk in communities study. Diabetes Care. 2005 Feb;28(2):385-90.

11. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics. 2004;20(14):2320-1.

12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001;17-21.

13. Ahlers CB, Fiszman M, Demner-fushman D, Lang F, Rindflesch TC. Extracting semantic predications from MEDLINE citations on pharmacogenomics. Pac Symp Biocomput. Jan, 2007;(12): 209-220.

14. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo. 2001;10(Pt1):216-20.Medinfo. 2001;10(Pt1):216-20.

15. Fiszman M, Rindflesch TC, Kilicoglu H. Summarization of an online medical encyclopedia. Medinfo. 2004;11(Pt 1):506-10.

16. Fiszman M, Rindflesch TC, Kilicoglu, H. Abstraction summarization for managing the biomedical research literature. Proc of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004;76-83.

17. Ridker PM, Wilson PW, Grundy SM. Should C-reactive protein be added to metabolic syndrome and to assessment of global cardiovascular risk? Circulation. 2004 Jun 15;109(23):2818-25.

18. Lin MS, Shih SR, Li HY, et al. Serum C-reactive protein levels correlates better to metabolic syndrome defined by International Diabetes Federation than by NCEP ATP III in men. Diabetes Res Clin Pract. 2007 Jan 15.

19. Brunner EJ, Hemingway H, Walker BR. Adrenocortical, autonomic, and inflammatory causes of the metabolic syndrome: nested case-control study. Circulation. 2002 Nov 19;106(21):2634-6.

20. Chandola T, Brunner E, Marmot M. Chronic stress at work and the metabolic syndrome: prospective study. BMJ. 2006 Mar 4;332(7540):521-5.

21. Muller A, Mulhall JP. Cardiovascular disease, metabolic syndrome and erectile dysfunction. Curr Opin Urol. 2006 Nov;16(6):435-43.

22. Laaksonen DE, Niskanen L, Punnonen K, et al. The metabolic syndrome and smoking in relation to hypogonadism in middle-aged men: a prospective cohort study. J Clin Endocrinol Metab. 2005 Feb;90(2):712-9.

23. Dessein PH, Stanwix AE, Joffe BI. Cardiovascular risk in rheumatoid arthritis versus osteoarthritis: acute phase response related decreased insulin sensitivity and high-density lipoprotein cholesterol as well as clustering of metabolic syndrome features in rheumatoid arthritis. Arthritis Res. 2002;4(5):R5.

24. Barrenas ML, Jonsson B, Tuvemo T, et al. High risk of sensorineural hearing loss in men born small for gestational age with and without obesity or height catch-up growth: a prospective longitudinal register study on birth size in 245,000 Swedish conscripts. J Clin Endocrinol Metab. 2005 Aug;90(8):4452-6.

25. Sanyal AJ, Contos MJ, Sterling RK, et al. Nonalcoholic fatty liver disease in patients with hepatitis C is associated with features of the metabolic syndrome. Am J Gastroenterol. 2003 Sep;98(9):2064-71.

26. Koike K. Hepatitis C virus infection can present with metabolic disease by inducing insulin resistance. Intervirology.2006;49(1-2):51-7.