Automating Identification of Multiple Chronic Conditions in Clinical Practice Guidelines

Tiffany I. Leung, MD, MPH[a,b,c], Hawre Jalal, MD, PhD, MSc[a,b], Donna M. Zulman, MD, MS[a,c],
Michel Dumontier, PhD[d], Douglas K. Owens, MD, MSc[a,b], Mark A. Musen, MD, PhD[d],
Mary K. Goldstein, MD, MS[a,b]

[a]*Department of Veterans Affairs, VA Palo Alto Health Care System, Palo Alto, CA*
[b]*Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA*
[c]*Division of General Medical Disciplines, Stanford University, Stanford, CA*
[d]*Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA*

Abstract

Many clinical practice guidelines (CPGs) are intended to provide evidence-based guidance to clinicians on a single disease, and are frequently considered inadequate when caring for patients with multiple chronic conditions (MCC), or two or more chronic conditions. It is unclear to what degree disease-specific CPGs provide guidance about MCC. In this study, we develop a method for extracting knowledge from single-disease chronic condition CPGs to determine how frequently they mention commonly co-occurring chronic diseases. We focus on 15 highly prevalent chronic conditions. We use publicly available resources, including a repository of guideline summaries from the National Guideline Clearinghouse to build a text corpus, a data dictionary of ICD-9 codes from the Medicare Chronic Conditions Data Warehouse (CCW) to construct an initial list of disease terms, and disease synonyms from the National Center for Biomedical Ontology to enhance the list of disease terms. First, for each disease guideline, we determined the frequency of comorbid condition mentions (a disease-comorbidity pair) by exactly matching disease synonyms in the text corpus. Then, we developed an annotated reference standard using a sample subset of guidelines. We used this reference standard to evaluate our approach. Then, we compared the co-prevalence of common pairs of chronic conditions from Medicare CCW data to the frequency of disease-comorbidity pairs in CPGs.

Our results show that some disease-comorbidity pairs occur more frequently in CPGs than others. Sixty-one (29.0%) of 210 possible disease-comorbidity pairs occurred zero times; for example, no guideline on chronic kidney disease mentioned depression, while heart failure guidelines mentioned ischemic heart disease the most frequently. Our method adequately identifies comorbid chronic conditions in CPG recommendations with precision 0.82, recall 0.75, and F-measure 0.78. Our work identifies knowledge currently embedded in the free text of clinical practice guideline recommendations and provides an initial view of the extent to which CPGs mention common comorbid conditions. Knowledge extracted from CPG text in this way may be useful to inform gaps in guideline recommendations regarding MCC and therefore identify potential opportunities for guideline improvement.

Introduction

Multiple chronic conditions (MCC) commonly refers to the existence of two or more chronic conditions in the same patient. More than two-thirds of Medicare beneficiaries have MCC in 2010,[1] and among Veterans, more than one-third of Veterans 65 years and above have three or more chronic conditions.[2] Providing guideline-concordant, patient-centered care to patients with MCC is challenging because many clinical practice guidelines (CPGs) typically focus on a single disease. Applying single-disease guidelines to patients with MCC increases medication treatment and self-management complexity, risk of interactions between guideline recommendations, and potential adverse events, hospitalization, and poorer outcomes.[3–6]

Suggested CPG improvements towards meaningful guidance in managing patients with MCC include cross-referencing single-disease guidelines and developing a clearer understanding of common chronic condition clusters.[7–12] However, the extent to which CPGs mention comorbid chronic conditions is unclear; and, when this has been done, such investigation has been done manually and for a single condition, such as diabetes.[13] Reviewing all guideline recommendations for multiple chronic conditions is potentially labor-intensive, and we sought to automate the process of identifying chronic conditions in CPG text. In addition to cross-referencing comorbid chronic disease guidelines, it has also been suggested that prevalence of MCC may be able to guide the development of the next generation of CPGs in explicitly addressing common comorbid condition clusters.[14] The frequency with which guidelines explicitly mention comorbid chronic conditions and its relationship to the prevalence of MCC has not been measured.

In this paper, we developed an automated method to assess how frequently guidelines mention highly prevalent, co-occurring chronic diseases. We performed textual annotation of CPG recommendations using clinical ontologies. Ontologies, which are machine-understandable descriptions of objects in a domain, have been widely used to codify knowledge, annotate textual documents, and perform statistical analyses.[15] Additionally, annotation of clinical notes and electronic health record documents, published scientific literature, clinical trials announcements, and drug labels using ontologies is well-established.[16–21] We focused on 15 common chronic conditions: hypertension, diabetes mellitus,

hyperlipidemia, stroke, asthma, atrial fibrillation, Alzheimer's dementia and senile dementias, osteoporosis, chronic obstructive pulmonary disease, depression, chronic kidney disease, heart failure, arthritis, ischemic heart disease, and obesity. To our knowledge, there are no studies that automate the identification of comorbid diseases in the text of disease-specific guidelines. This approach may offer a novel opportunity to understand and extract knowledge from the text of guideline recommendations. Making such clinical practice guideline knowledge more readily accessible may support efforts towards retrieving guideline knowledge for use by clinicians in decision-making, developers of clinical decision support systems in personalizing care for patients with MCC, and institutional guideline implementation working groups and committees in managing knowledge. We compared the frequency of disease-comorbidity pairs to the prevalence of these comorbid condition pairs in the Medicare population.

Materials and Methods

To analyze CPG text, we used three publicly available repositories of information: (1) guideline summaries from the National Guideline Clearinghouse, (2) epidemiologic data on the co-prevalence of chronic diseases and a data dictionary of chronic conditions from the Medicare Chronic Conditions Data Warehouse, and (3) biomedical ontologies in the Bioportal repository of the National Center for Biomedical Ontology.

Data Sources

National Guideline Clearinghouse

The Agency for Healthcare Research and Quality maintains the National Guideline Clearinghouse (NGC), first developed in 1997, which identifies published CPGs according to specified inclusion criteria, creates guideline summaries that contain structured, standardized sections, and archives them in a publicly accessible website for retrieval.[22] NGC guideline summaries are structured using 54 guideline attributes,[23] such as Guideline Title, Guideline Objective, and Major Recommendations.[23] As of the end of February 2014, the NGC featured more than 2,500 guidelines.

National Center for Biomedical Ontology

The National Center for Biomedical Ontology (NCBO) is an international consortium based at Stanford University, funded by the National Institutes of Health, which provides online tools for accessing and integrating ontological resources, including Bioportal, a repository of biomedical ontologies. As of February 2014, the NCBO Bioportal contained nearly 400 biomedical ontologies. For this study, we utilized Bioportal ontologies to identify disease synonyms for the 15 chronic conditions of interest.

Medicare Chronic Conditions Data Warehouse

The Chronic Conditions Data Warehouse (CCW), administered by the Centers for Medicare and Medicaid Services, is a research database of care provided to Medicare beneficiaries with chronic conditions. The CCW contains a data dictionary that we used to create a list of ICD-9 (9th version of the International Statistical Classification of Diseases and Related Health Problems) codes for 14 of the most prevalent chronic diseases among Medicare beneficiaries: hypertension, diabetes mellitus, hyperlipidemia, stroke, asthma, atrial fibrillation, Alzheimer's dementia and senile dementias, osteoporosis, chronic obstructive pulmonary disease, depression, chronic

kidney disease, heart failure, arthritis, and ischemic heart disease.[24] We added 3 more ICD-9 codes for obesity, which we added as the 15th disease due to its high prevalence and clinical significance, instead of cancer, which is more heterogeneous. The CCW also provided publicly accessible data on the co-prevalence of chronic conditions in the Medicare population.

Guideline Recommendations Corpus

We compiled a corpus of text from the NGC by downloading the XML version of 2,503 unique guideline summaries in March 2014. We obtained the Major Recommendation sections from 268 XML-based NGC guideline summaries. CPGs were included if (1) at least one of the 15 diseases was mentioned in the title and (2) the target population was the general adult, non-pregnant population. We excluded guideline summaries that were not relevant to the 15 chronic conditions of interest based on the guideline title. For those guidelines that were subdivided into separate summaries, we combined these into a single summary; for example, the American Diabetes Association publishes a full guideline document entitled the Standards of Medical Care in Diabetes, which the NGC parses into nine separate guideline summaries.[22] In this study we refer to NGC guideline summaries as clinical practice guidelines, or CPGs.

Algorithm

We first constructed a list of 448 ICD-9 codes for the 15 diseases using the Medicare CCW.[24] In addition, we identified three corresponding ICD-9 codes for obesity (*Figure 1*). Then, we used the NCBO Bioportal Representational State Transfer (REST) services to obtain 1,829 unique preferred labels and synonyms for each ICD-9 code.[25] Finally, we developed a text-mining algorithm to identify disease mentions by exactly matching the preferred labels and synonyms in the text corpus. All algorithms were developed using *Python 2.7.8*.
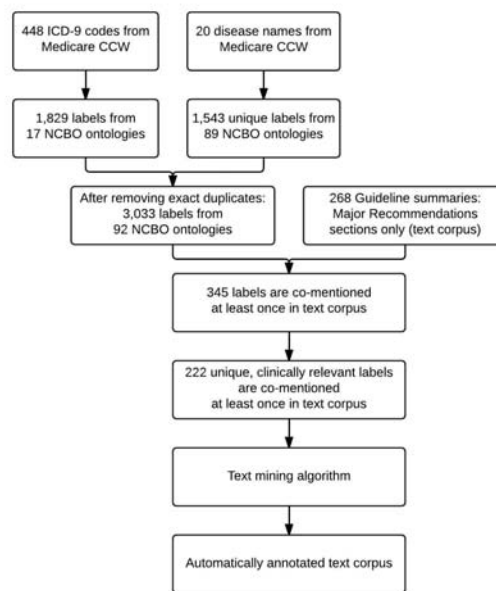

Figure 1. Annotation pipeline for automated identification of disease labels for 15 chronic conditions. Labels include NCBO preferred labels and synonyms.

We counted the frequency of chronic condition mentions in the CPG recommendations (a disease-comorbidity pair). Spe-

cifically, the topic for each disease guideline was the disease in the disease-comorbidity pair, and a mention of a comorbid chronic condition was the comorbidity. Each guideline might have more than one instance of a disease-comorbidity pair if the recommendation text mentioned more than one exact match of the comorbid disease term. For example, in a heart failure guideline, if *coronary artery disease* and *acute myocardial infarction* are each mentioned one time, then there would be two occurrences of the heart failure-ischemic heart disease pair in this guideline.

We evaluated our method by developing a reference standard that summarized the number of disease-comorbidity pairs in the recommendations text of a sample of guidelines. Then, precision, recall, and F-measure were calculated. To compare the prevalence of common chronic condition pairs with the disease-comorbidity frequencies in CPGs, we constructed two network maps. Network maps were used to better illustrate each condition's prevalence and their co-occurrence with comorbid chronic conditions. We obtained epidemiologic data on prevalence and co-occurrence of 14 of the 15 diseases from the Medicare CDW; obesity co-occurrence was estimated from published literature.[26]

Reference Standard

We developed a reference standard to identify comorbid disease concepts in a sample subset of CPG text (*Figure 2*). In the sample, we randomly selected one guideline for each chronic condition for a total of 15 guidelines manually annotated. As an annotation guide for manual annotation, comorbid disease mentions in each guideline, included (1) explicit comorbid disease mentions, including common synonyms (e.g. *left ventricular systolic failure* for congestive heart failure) or guideline-defined acronyms (e.g. *CHF* for chronic heart failure) and (2) statements that referenced or described a mention of one of the 15 chronic conditions, or for which the mention could be inferred from the clinical context of the statement. For example, one coronary artery disease guideline states: "Individuals with established cardiovascular disease, who also have chronic renal disease or diabetes with complications, or target organ damage may be considered for treatment at the lower threshold of *systolic 130 mm Hg and/or diastolic 80 mm Hg*."[27] This statement mentions a description clinically consistent with hypertension, even if *hypertension* is not explicitly stated.
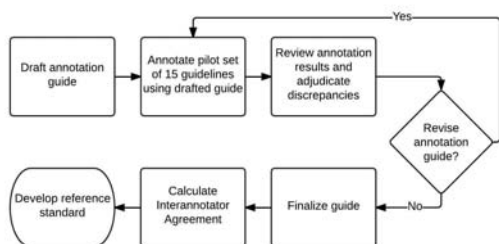

Figure 2. Development of annotation guide and reference standard

Two manual annotators with medical expertise (TL and HJ) performed the annotation using the guide, adjudicated discrepancies, revised the guide and finalized the reference standard. We determined interannotator agreement (IAA) by calculating a ratio of the number of sentence agreements to the total number of sentences, and Cohen's Kappa (κ).

Results

Selection of guideline summaries meeting criteria for inclusion in the text corpus are illustrated in *Figure 3*. To ensure we included all relevant titles, we performed both a text search and a manual review of all 2,503 unique guideline summary titles. After inclusion based on text search of the titles, manual review of the titles identified 26 additional guideline summaries where it was unclear if it met inclusion criteria; we manually reviewed the Major Recommendations to identify the disease, and included 23 guideline summaries.
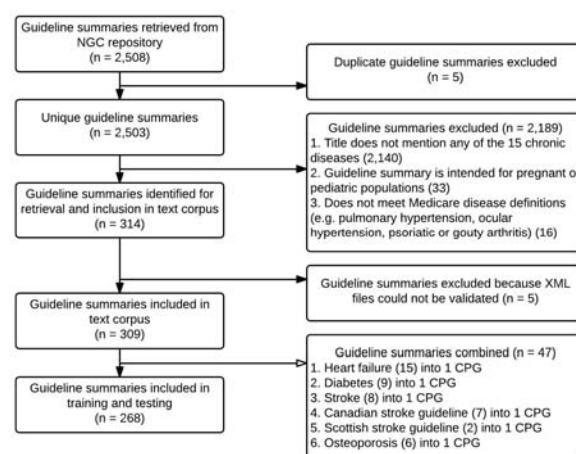

Figure 3. Guideline summaries meeting inclusion criteria for text corpus. CPG = clinical practice guideline

In the reference standard, manual annotators agreed on annotations for 2,890 (97.0%) of 2,981 sentences. In this preliminary evaluation, compared to the reference standard, precision of our approach was 0.82, recall was 0.75, and F-measure was 0.78.

Figure 4 summarizes our preliminary findings on the number of CPGs for each condition (represented by each node) and the number of disease-comorbidity pairs (represented by a directed edge). For example, the *diabetes mellitus-ischemic heart disease* pair occurs 153 times in 43 diabetes CPGs; conversely, the *ischemic-heart disease-diabetes mellitus* pair occurs 323 times in 38 ischemic heart disease CPGs. CPGs for concordant diseases (diseases that are part of the same pathophysiologic risk profile), such as hypertension, diabetes, and ischemic heart disease, mentioned one another most frequently. Hypertension was the only disease mentioned across all CPGs, while Alzheimer's disease and osteoporosis were mentioned the least. *Figure 5* illustrates the prevalence of each chronic disease among Medicare beneficiaries in 2012 (nodes), and the co-occurrence of each condition with another common chronic condition (edges). For example, the prevalence of diabetes mellitus and Alzheimer's disease and dementia are 18% and 11%, respectively, and the prevalence of their co-occurrence is 5.4%.

The network diagrams of disease-comorbidity pairs and the prevalence of co-occurring chronic condition pairs highlights a mismatch between explicit mentions of chronic conditions in CPGs and the prevalence of MCC in Medicare beneficiaries. The mismatch is less pronounced for concordant chronic conditions, such as hypertension, diabetes, and hyperlipidemia. For discordant conditions, the mismatch is more pronounced, such as in hypertension and arthritis, which co-occur in the

Medicare population with a prevalence of 32.2%, but have minimal explicit mentions of one another: the *arthritis-hypertension* pair occurs 5 times in 19 arthritis CPGs and *hypertension-arthritis* pair occurs zero times in 13 hypertension CPGs.

In general, some disease-comorbidity pairs occur more frequently in CPGs than others. Sixty-one (29.0%) of 210 possible disease-comorbidity pairs occurred zero times; for example, no guideline on *chronic kidney disease* mentioned *depression*, while *heart failure* guidelines mentioned *ischemic heart disease* 209 times, the most frequently occurring disease-comorbidity pair among included CPGs.
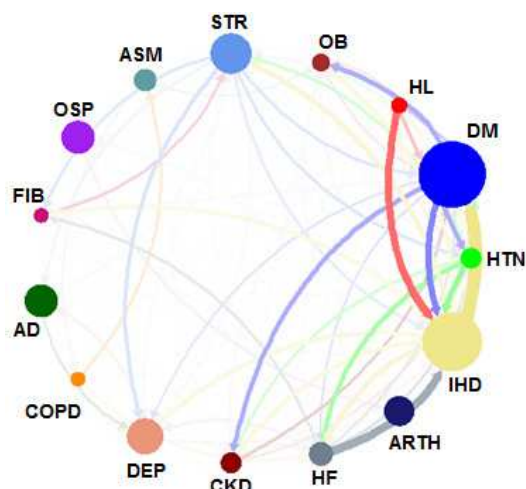

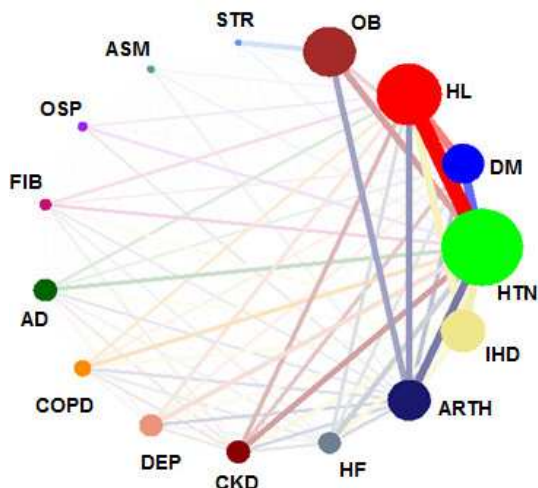*Figure 4. Frequency of disease-comorbidity pairs*

Figure 5. Co-occurring chronic diseases among Medicare beneficiaries, 2012

HTN Hypertension	**OSP** Osteoporosis	**HF** Heart Failure
DM Diabetes Mellitus	**FIB** Atrial Fibrillation	**ARTH** Arthritis
HL Hyperlipidemia	**COPD** Chronic Obstructive Pulmonary Disease	**IHD** Ischemic Heart Disease
OB Obesity	**AD** Alzheimer's and Dementia	
STR Stroke	**DEP** Depression	
ASM Asthma	**CKD** Chronic Kidney Disease	

Discussion

We developed an automated method that identifies comorbid conditions in CPG recommendations. Our results suggest that highly prevalent co-occurring pairs of conditions are variably addressed explicitly in corresponding disease guidelines. For concordant conditions, which also co-occur commonly in the Medicare population, the high frequency of disease-comorbidity pairs suggest that more guidance exists in CPG recommendations for these combinations of MCC. However, for discordant conditions, less guidance may exist, although such conditions commonly co-occur among patients with MCC. These represent potentially high-impact areas of opportunity for guideline improvement in order to provide clinicians with explicit guidance on providing care for these patients with common MCC combinations.

Limitations of this study include the use of NGC guideline summaries, rather than original, full-text CPGs. Additionally, the text corpus is small, totaling only 268 guideline summaries. In a larger text corpus inclusive of full-text guidelines, disease-comorbidity pairs may occur more frequently and potentially provide a more accurate representation of MCC in CPG recommendations. A more extensive evaluation of our method using a reference standard including more CPGs also could be informative, allowing for a more detailed understanding of how the method performs for the identification of each chronic condition in CPG text. Future research could iteratively improve the text-mining algorithm developed, explore alternative analytic methods, evaluate alternative annotation tools for biomedical concept annotation, and perform natural language processing on the text corpus.

Further investigation is needed to understand the observed variation in the frequency of disease-comorbidity pairs chronic disease CPGs. Our methods can be used to extract knowledge from CPGs about chronic conditions towards facilitating future research and identify guideline improvement opportunities. Making such clinical practice guideline knowledge more readily accessible may support efforts towards retrieving guideline knowledge for use by clinicians in managing care for patients with MCC, developing clinical decision support systems personalizing care for patients with MCC, and assisting institutional guideline working groups in managing knowledge.

Conclusion

Our approach adequately identifies explicit knowledge about comorbid chronic conditions currently embedded in the free text of clinical practice guideline recommendations. Our findings may help prioritize opportunities for single-disease chronic condition guidelines to be improved towards providing specific guidance for patients with MCC, especially for discordant comorbid conditions.

Address for correspondence

Correspondence to Tiffany I. Leung, MD, MPH, Division of General Medical Disciplines, Stanford University School of Medicine, 211 Quarry Road, Suite #102, Palo Alto, CA 94305. E-mail: tileung@stanford.edu

References

1. Centers for Medicare and Medicaid Services. Chronic Conditions Among Medicare Beneficiaries, Chartbook. 2012th ed. Baltimore, MD: 2012.

2. Yoon J, Zulman D, Scott JY, Maciejewski ML. Costs Associated With Multimorbidity Among VA Patients. *Med Care.* 2014 Mar;52 Suppl 3(3):S31–6.

3. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical Practice Guidelines and Quality of Care for Older Patients. *JAMA.* 2005;294(6):716–724.

4. Zulman DM, Asch SM, Martins SB, Kerr E a, Hoffman BB, Goldstein MK. Quality of Care for Patients with Multiple Chronic Conditions: The Role of Comorbidity Interrelatedness. *J Gen Intern Med.* 2013 Oct 1

5. Tinetti ME, Bogardus ST, Agostini J V. sounding board Potential Pitfalls of Disease-Specific Guidelines for Patients with Multiple Conditions. 2004;2870–2874.

6. Hughes LD, McMurdo MET, Guthrie B. Guidelines for people not for diseases: the challenges of applying UK clinical guidelines to people with multimorbidity. *Age Ageing.* 2013 Jan;42(1):62–9.

7. Uhlig K, Leff B, Kent D, Dy S, Brunnhuber K, Burgers JS, et al. A Framework for Crafting Clinical Practice Guidelines that are Relevant to the Care and Management of People with Multimorbidity. *J Gen Intern Med.* 2014 Jan 18;7–9.

8. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical Practice Guidelines and Quality of Care for Older Patients. *JAMA.* 2005;294(6):716–724.

9. Ickowicz E. Guiding Principles for the Care of Older Adults with Multimorbidity: An Approach for Clinicians. *J Am ... [Internet].* 2012 Oct [cited 2013 Dec 6];60(10):E1–E25.

10. Eckstrom E, Feeny DH, Walter LC, Perdue L a, Whitlock EP. Individualizing cancer screening in older adults: a narrative review and framework for future research. *J Gen Intern Med.* 2013 Feb;28(2):292–8.

11. Steinman M a, Lee SJ, John Boscardin W, Miao Y, Fung KZ, Moore KL, et al. Patterns of multimorbidity in elderly veterans. *J Am Geriatr Soc.* 2012 Oct;60(10):1872–80.

12. Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity. *BMJ.* 2012 Jan;345(October):e6341.

13. Wyatt KD, Stuart LM, Brito JP, Carranza Leon B, Domecq JP, Prutsky GJ, et al. Out of context: clinical practice guidelines and patients with multiple chronic conditions: a systematic review. *Med Care.* 2014 Mar;52 Suppl 3(3):S92–S100.

14. Steinman M a, Lee SJ, John Boscardin W, Miao Y, Fung KZ, Moore KL, et al. Patterns of multimorbidity in elderly veterans. *J Am Geriatr Soc.* 2012 Oct;60(10):1872–80.

15. Musen M a, Noy NF, Shah NH, Whetzel PL, Chute CG, Story M-A, et al. The National Center for Biomedical Ontology. *J Am Med Inform Assoc.* 2012;19(2):190–5.

16. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012 Jun;13(6):395–405.

17. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *J Am Med Inform Assoc.* 2014 Oct 21;1–8.

18. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen M a. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics.* 2009 Jan;10 Suppl 9:S14.

19. Guo Y, Korhonen A, Liakata M, Silins I, Hogberg J, Stenius U. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics.* 2011 Jan;12(1):69.

20. Boyce RD, Horn JR, Hassanzadeh O, Waard A De, Schneider J, Luciano JS, et al. Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. *J Biomed Semantics.* 2013 Jan;4(1):5.

21. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform.* 2007 Sep;8(5):358–75.

22. Agency for Healthcare Research and Quality. National Guideline Clearinghouse. Last accessed 2014 Feb 14; Available from: http://www.guideline.gov/

23. Agency for Healthcare Research and Quality. National Guideline Clearinghouse: Template of Guideline Attributes. Last accessed 2014 Feb 14; Available from: http://www.guideline.gov/about/template-of-attributes.aspx

24. Centers for Medicare and Medicaid Services. Chronic Condition Warehouse Categories. Last accessed 2014 May 26. Available from: https://www.ccwdata.org/cs/groups/public/documents/document/ccw_conditioncategories2011.pdf

25. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 2009 Jul;37(Web Server issue):W170–3.

26. Doshi J, Polsky D, Chang V. Prevalence and trends in obesity among aged and disabled US Medicare beneficiaries, 1997–2002. *Health Aff.* 2007;26(4):1111–1117.

27. Scottish Intercollegiate Guidelines Network (SIGN). Risk estimation and the prevention of cardiovascular disease. A national clinical guideline. *Natl. Guidel. Clear.* 2007. Last accessed 2014 Sep 24;Available from: http://www.guideline.gov/content.aspx?id=10588