Computing Conceptual Distances between Breast Cancer Screening Guidelines: An Implementation of a Near-Peer Epistemic Model of Medical Disagreement.

Hossein Hematialam∗c, Luciana Garbayo∗f, Seethalakshmi Gopalakrishnan∗c, Wlodek Zadrozny∗cd

College of Computing, UNC Charlotte, Charlotte NC
School of Data Science, UNC Charlotte, Charlotte NC
Departments of Philosophy & Medical Education, U. of Central Florida, Orlando, FL

Abstract

Using natural language processing tools, we investigate the differences of recommendations in medical guidelines for the same decision problem – breast cancer screening. We show that these differences arise from knowledge brought to the problem by different medical societies, as reflected in the conceptual vocabularies used by the different groups of authors. The computational models we build and analyze agree with the near-peer epistemic model of expert disagreement proposed by Garbayo. Even though the article is a case study focused on one set of guidelines, the proposed methodology is broadly applicable.

In addition to proposing a novel graph-based similarity model for comparing collections of documents, we perform an extensive analysis of the model performance. In a series of a few dozen experiments, in three broad categories, we show, at a very high statistical significance level of 3-4 standard deviations for our best models, that the high similarity between expert annotated model and our concept based, automatically created, computational models is not accidental. Our best model achieves roughly 70% similarity. We also describe possible extensions of this work.

∗∗ Corresponding authors: {wzadrozn,hhematia,sgopala}@uncc.edu; Luciana.Garbayo@ucf.edu
Keywords: medical guidelines, natural language processing, epistemic models, conceptual similarity, near-peer model of expertise,

1. Introduction and Motivation

Research Objective: In this article we investigate the differences in medical guidelines in response to the same decision problem: whether to recommend a breast cancer screening for patients with same conditions. Our research objective is to create a computational model accurately representing medical guidelines disagreements; a model which is simple and general enough to be potentially applicable in other situations. We evaluate our approach using a case study, where we are asking whether differences in medical recommendations come from differences in knowledge brought to the problem by different medical societies.

This article should also be viewed as a case study in computational implementation of the near-peer epistemic model of expert disagreement proposed in several of our earlier work (Garbayo (2014), Garbayo et al. (2018), Garbayo et al. (2019), Garbayo (2019)). The near-peer model can be viewed as a refinement of the standard epistemic peer model (e.g. Lackey (2014)).

More specifically, we use natural language processing tools to build computational representations of a set of seven breast cancer screening guidelines. These computational representations are created from full texts of the guideline documents. Since we are using vectors to represent the documents, we can ask whether the distances between vectors (e.g. the cosine distance) are semantically significant; and in particular, whether the degrees of conceptual disagreement between the guidelines correspond to the differences in semantic distances automatically computed from full documents. We measure the conceptual disagreement using a CDC summary of the full documents (CDC (2017), which focuses precisely on the differences in screening recommendations. An

---

1 by knowledge we mean both the domain knowledge and the associate epistemic practices
annotated summary of the seven documents from [CDC (2017)] is reproduced in Fig. 1, with colors added for areas of agreement and disagreement. In addition, we plot these disagreements in a diagram (see Fig. 2).

![Table Image]

Figure 1: Note the contradictory recommendations in green and blue boxes. In [Zadrozny et al. (2017)], we annotated the table from CDC comparing recommendations of different medical societies [CDC (2017)] to explicitly show the types of disagreements in different guidelines. We also showed there that some of these contradictions can be computed automatically. Only part of the table is reproduced here.

In other words, we ask about whether concepts, that is single words or two consecutive words (bigrams), which are present in full documents determine the relations seen in the documents’ summary. That is, whether concepts used by experts affect their recommendations. (Note that the concepts that appear in summaries, such as mammography, are common in all documents, so the difference in recommendations must come from other knowledge).

Technically speaking, we compute semantic distances between several pairs of breast cancer screening guidelines, using different automated methods, and then we compare the resulting measurements to an expert opinion in Fig. 2.
As we explain later in the article these comparisons suggest the potential for computational tools to be used to estimate epistemic distances between the guidelines.

**Motivation:** Guidelines are complex products for medical decision making. Despite some convergence in the interpretation of medical evidence, different medical specialties might produce dissimilar guidelines for the same medical problem. Nevertheless, physicians who have similar training tend to generate similar guidelines based on their sharing of methods, objects and overall academic background typical of the specialty Garbayo et al. (2018). These specialties cover a specific region of theoretical and practical medical knowledge that may overlap with others. For example, oncology or radiology overlaps with general practice. Per our hypothesis, this variable similarity of backgrounds and specialization is likely to be reflected in the types recommendations for either screening, treatment or prevention of specific conditions, such as breast cancer.

Since medical guidelines are developed by different medical associations which count on experts with different specialties and sub-specialties, there is a high possibility that there may be disagreement in the guidelines. And indeed, as noted by CDC (2017) and discussed in Zadrozny & Garbayo (2018); Zadrozny et al. (2017), breast cancer screening guidelines contradict each other. Besides breast cancer screening disagreements, which we model in this article, we have witnessed controversies over PSA screening, hypertension and other treatment and prevention guidelines.

**The near-peer hypothesis:** Conceptually, this article presents a computational study in epistemic modeling of medical guidelines disagreement, as a model of *near-peer* disagreement. By that we mean the following: traditionally, domains experts, e.g. the authors of medical guidelines, have been viewed as *epistemic peers*, i.e. groups of professionals accessing the same medical knowledge in the process of creating the guidelines; and in the case of disagreements, they can evaluate opposing views, and have identical levels of competence.
As exemplified in Christensen et al. (2013), this has been the dominant paradigm in analyzing disagreements among experts (peers). However, a more realistic and fine grained model is possible, namely, where we see these groups as having partly overlapping knowledge, and therefore can be named near-peers. While, intuitively, such a model sounds reasonable, it raises the question, how should such near-peer models be created and analyzed. Clearly we do not know exactly what kind of knowledge the individuals involved in creating the guidelines bring to the table, and even with access to the discussions underlying the creation process, doing a manual conceptual analysis would be slow and tedious. Yet, given the progress achieved in building computational models of text documents [Zhou et al., 2020, in press], we can hypothesize that such computational models might be of some use. Hence this hypothesis:

**Hypothesis:** The epistemic near-peer disagreement between medical experts and their societies can be measured using natural language processing techniques to measure the conceptual distance between the produced guidelines.

**Contributions:** The main contribution of this article is in showing that automated, and relatively straightforward, methods of text analysis can compute conceptual differences between documents addressing the same topic (breast cancer screening); and these automated judgments have a moderate to high correlation with an expert judgment. Namely, we compute the similarity and the dissimilarity between the breast cancer guidelines provided by different medical societies, using a few standard methods of representing text and computing such metrics. We then correlate it with previously done conceptual analysis of the main recommendations of these guidelines. Thereby, we show the viability of the near-peer model.

Another contribution is the articulation of a very natural graph-clique based algorithm/method for comparing similarity of two *collections* of documents. Given two sets of documents, each of the same cardinality, and a mapping between nodes, we compute the percent distortion between the shapes of the
two cliques, and the chances that the mapping arose from a random process.  

We also document all steps of the process and provide the data and the code to facilitate both extensions of this work and its replication. Even though NLP methods have progressed enormously in the last decade [Zhou et al. (2020, in press)], they are far from perfect. In our experiments, we use some of the simplest semantic types—words and simple collocations—represented as vectors in high-dimensional spaces. However, this simplicity is helpful, as we can run several experiments, and compare the effects of using different representations and metrics. This gives us confidence that the correlations we are discovering tell us something interesting about guideline documents.

Although the article merely establishes this correspondence in one case, nevertheless it might be a good starting point for analysis of other medical guidelines, and perhaps other areas of expert disagreement. In addition, fast progress in automated document analysis using text mining and deep learning techniques can perhaps make such analyses more accurate and deeper.

**Organization of the article:** In Section 2 we provide both, an overview of applications of natural language processing to texts of medical guidelines, and introduce the near-peer model of epistemic disagreement. Section 3 explains our data sources: a CDC summary table of breast cancer screening guidelines and the corresponding full text documents. There, we also discuss the steps in the conceptual analysis of the table, resulting in a graph of conceptual distances between the columns of the table (i.e. summaries of the full documents). We then proceed to the analysis of full documents using a two standard vectorization procedures in Section 4. After observing roughly 70% correlation between the

---

2Given the naturalness of the method, it’s likely that we are reinventing something, but we couldn’t find anything similar in literature. We would appreciate pointers to related prior art.

3The Github link will be provided here before publication.
distances in the summaries and the distances in the full documents, we prove in Section 5 that this correlation is not accidental. We conclude in Sections 5 and 6 that this case study shows that NLP methods are capable of approximate conceptual analysis, in agreement with the near-peer model. This opens the possibility of deepening such research using more sophisticated tools such as relationship extraction and automated formal analysis.

2. Discussion of prior art

We are not aware of any work directly addressing the issues we are tackling in this article; namely, the automated conceptual analysis of medical screening recommendations, and connecting such analysis to broader problems of epistemic peers or near-peers and their disagreements. However, there is a body of knowledge addressing similar issues individually, which we summarize in this section.

2.1. Text analysis of medical guidelines

An overview article Peck et al. (2015), from a few years ago, states that different types of analysis of medical guidelines are both a central theme in applications of artificial intelligence to medicine, and a domain of research with many challenges. The latter includes building formal, computational representations of guidelines and a wider application of natural language processing. From this perspective, our work is relevant to these central but general themes.

To switch to more recent and more technical work, Bowles et al. (2019) focuses on finding and resolving conflicting recommendations using a formal model and automated proof systems – it relies on a manual translation into a formal language, Labelled Event Structure. This is a very interesting work, somewhat in the spirit of our own attempts to do it, Zadrozny et al. (2017), using a combination of NLP and information retrieval tools. In another article dealing with contradictory recommendations, Tsopra et al. (2018) focus on the semi-automatic detection of inconsistencies in guidelines and apply their tools.
to antibiotherapy in primary care. In an application of Natural Language Processing, Lee et al. (2020) show that one can accurately measure adherence to best practice guidelines in a context of palliative care.

More broadly, modern NLP methods have been applied to clinical decision support, e.g. Seneviratne et al. (2019), with ontologies and semantic web for concept representation; and to automatic extraction of adverse drug events and drug related entities, e.g. Ju et al. (2020) using a neural networks model. For document processing, we have e.g. Benedetti et al. (2019) proposing a knowledge-based technique for inter-document similarity computation, and Rospocher et al. (2019) successfully applying conceptual representations to document retrieval.

All of these show that the state-of-the-art systems are capable both of performing statistical analysis of large sets of documents, and a semantic analysis fitting the need of a particular application.

**Extending the limits of current practice:** This work extends the state-of-the-art computational analysis of medical guidelines. Namely, instead of semi-automated conceptual analysis, we prove the feasibility of automated conceptual analysis. That is, in our study, we use a representation derived from a (relatively shallow) neural network (BioASQ embeddings Tsatsaronis et al. (2015)), and knowledge-based annotations derived from MetaMap. Our results, in Section 5, show that both are useful as representations of our set of guidelines, and show similar performance in modeling conceptual similarities. From the point of view of methodology of analyzing medical guidelines, this article contains the first computational implementation of the near-peer model.

2.2. Conceptual analysis of disagreement in medical guidelines

The significance of medical guidelines disagreement, such as in the case of breast cancer screening disagreement, is expressed in the projected preventable harms of unwarranted clinical variation in health care, Sutherland & Levesque (2019). The cited article proposes a theoretical framework for the examination of

---

https://metamap.nlm.nih.gov/
unwarranted clinical variation, and suggests we can identify such unwarranted clinical variation in three dimensions: assessing variation across geographical areas or across providers;” through criteria for assessment, measuring absolute variation against a standard, or relative variation within a comparator group”; and as object of analysis, using process structure/resource, or outcome metrics.”

It is important to notice that such unwarranted clinical variation is not about precision medicine and customization of care; rather, it represents a disagreement that includes knowledge management and translation issues, as well as a variation of epistemic practices. For instance, Solomon (2015), provides a compelling social epistemology study of consensus conferences and their epistemic pitfalls.

In this article we focus more narrowly on the epistemic analysis of medical guidelines disagreement. As argued in our earlier work (Garbayo et al. (2019)), the expectation of a epistemic agreement and consensus across different medical societies of specialties rests in a historic habit of idealization. That is, it lies in viewing domain experts as having identical knowledge and identical reasoning capabilities, and whose reasoning processes can – in principle – be expressed formally and mechanically verified.

Departing from a broader interpretation of epistemic peers Lackey (2014), to allow for variation, Garbayo et al. (2019a) and Garbayo et al. (2019) explore the de-idealization of medical consensus and disagreement, and propose a category of near-peers, to express more accurately the mismatches in knowledge domain and variability among multiple guidelines developers as epistemic agents. In the broader interpretation, they are all epistemic peers, but, in the study of variations, such micro-analyses provide us conceptual distances between their perspectives in a measurable way.

2.3. The gap between theory and engineering analysis of contradictory guidelines

Based on the above discussion, there is a clear gap between theoretical analyses of the creation and content of medical guidelines and the computational analysis of the same content, e.g. from the epistemic point of view. In partic-
ular, as far as we know, there has been no computational investigation of contradictory guidelines incorporating the epistemic point of view of either expert peers nor near-peers. We demonstrate in the subsequent sections the feasibility of such computational models.

3. The approach: data sources of guidelines and conceptual analysis

As mentioned earlier, we are comparing the result of a manual conceptual analysis of several breast cancer screening guidelines with an automated analysis. In this section, we describe the process of this manual analysis of the summary document. In particular we produce numerical representations of the differences in the guidelines (per CDC (2017)), which later in Section 4 will be compared with the results of an automated analysis of full guideline documents.

The guidelines documents: In this article, we are using both the CDC summary (CDC (2017), reproduced and annotated in Fig. 1), and the full text of the guidelines used by the CDC. The detailed information about these guidelines is shown in Table 1. The focus of this section is on Fig. 1.

As shown in Fig. 1, reproduced from our earlier work Hematialam & Zadrozny (2016) and Zadrozny & Garbayo (2018), there are several clear disagreements in the recommendations.

Conceptual Analysis: Figure 2 shows a manually generated graph showing the differences between the guidelines, also presented in Garbayo et al. (2019a). There are two sides to the circle. The yellow side indicates the scenario where patients will likely decide when breast cancer screening should be done, and the purple color side specifies the situation where breast cancer guideline providers most likely will demand screening interventions. Black color indicates the different societies boundaries. The red color marks indicate the physician decides. Green color marks indicate patients’ decisions. Since the consideration of ultrasound (blue) appears only in radiology...
| Guideline Abbreviation | Full Name of the Organization | URL Reference | Document Citation |
|------------------------|-------------------------------|---------------|-------------------|
| ACOG                   | The American College of Obstetrics and Gynecology | [http://msrads.web.unc.edu/files/2019/05/ACOGBreastCAScreening2014.pdf](http://msrads.web.unc.edu/files/2019/05/ACOGBreastCAScreening2014.pdf) | American College of Obstetricians and Gynecologists (2017) |
| AAFP                   | American Academy of Family Physicians | [https://www.aafp.org/dam/AAFP/documents/patient_care/clinical_recommendations/cps-recommendations.pdf](https://www.aafp.org/dam/AAFP/documents/patient_care/clinical_recommendations/cps-recommendations.pdf) | Action (2017) |
| ACP                    | American College of Physicians | [https://annals.org/aim/fullarticle/2294149/screening-cancer-advice-high-value-care-from-american-college-physicians](https://annals.org/aim/fullarticle/2294149/screening-cancer-advice-high-value-care-from-american-college-physicians) | Wilt et al. (2015) |
| ACR                    | American college of Radiology | [https://www.sciencedirect.com/science/article/pii/S1546144009004803](https://www.sciencedirect.com/science/article/pii/S1546144009004803) | Lee et al. (2010) |
| ACS                    | American Cancer Society | [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831582/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831582/) | Oeffinger et al. (2015) |
| IARC                   | International Agency for Research on Cancer | [https://www.nejm.org/doi/full/10.1056/NEJMoa1508733](https://www.nejm.org/doi/full/10.1056/NEJMoa1508733) | Jørgensen & Bev. (2015) |
| USPSTF                 | United States Preventive services Task Force | [https://annals.org/aim/fullarticle/2480757/screening-breast-cancer-us-preventive-services-task-force-recommendation](https://annals.org/aim/fullarticle/2480757/screening-breast-cancer-us-preventive-services-task-force-recommendation) | Siu (2016) |

Table 1: Guidelines with references

guidelines, we decided to abstract it out.

The observed differences seem to support a near-peer model: we see partially overlapping agreements. If we look into the above diagram (Fig. 2) we can infer that ACS is the nearest guideline to IARC on the purple side and USPTF is the nearest one on the yellow side. USPSTF is the nearest guideline to the AAFP. ACR is the farthest node to USPSTF. ACR, ACOG are close to each other. Notice that all we did was to observe the differences in recommendations. We are not trying to judge if some of these differences are more important than the others. Obviously, this categorization process is informed by medical knowledge, but the lists of differences are clear from reading the document reproduced

11
Figure 2: Similarities between guidelines using expert annotations. The yellow coloring shows patient making decisions, the purple coloring shows explicit screening recommendations. The concentric circles refer show different age groups. Red marks – physician decides, green marks – patient decides.

So the question is whether we can reproduce these similarities and differences using a fully automated process, and without access to the summary document in Fig[I]
Table 2: The table shows recommendations as follows: N — no recommendation; b — both, patient and doctor, shared decision; r — recommending mammography.

Table 2 represents the content of this analysis as a collection of features. Table 3 shows the distances between the guidelines derived from Tables 2 and 3 using the Jaccard distance (the percentage of different elements in two sets): \( d_j(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \)

| Guideline | 40-49 | 50-74 | 75+ | Dense Breast | Higher than average risk |
|-----------|-------|-------|-----|--------------|-------------------------|
| AAFP      | b     | r     | b   | b            | N                       |
| ACOG      | r     | r     | b   | b            | r                       |
| ACP       | b     | r     | r   | N            | N                       |
| ACR       | r     | r     | r   | b            | r                       |
| ACS       | b     | r     | r   | b            | b                       |
| IARC      | b     | r     | N   | b            | r                       |
| USPSTF    | b     | r     | b   | b            | r                       |

Table 3: This table shows the number of different feature values for each pair of guidelines, based on Table 2.
|       | AAFP | ACOG  | ACP | ACR | ACS | IARC | USPSTF |
|-------|------|-------|-----|-----|-----|------|-------|
| AAFP  | 0    | 0.0238| 0.0357 | 0.0357 | 0.0238 | 0.0238 | 0.0119 |
| ACOG  | 0.0238 | 0    | 0.0476 | 0.0119 | 0.0238 | 0.0238 | 0.0119 |
| ACP   | 0.0357 | 0.0476 | 0    | 0.0357 | 0.0238 | 0.0357 | 0.0357 |
| ACR   | 0.0357 | 0.0119 | 0.0357 | 0    | 0.0119 | 0.0238 | 0.0238 |
| ACS   | 0.0238 | 0.0238 | 0.0238 | 0.0119 | 0    | 0.0119 | 0.0119 |
| IARC  | 0.0238 | 0.0238 | 0.0357 | 0.0238 | 0.0119 | 0    | 0.0119 |
| USPSTF| 0.0119 | 0.0119 | 0.0357 | 0.0238 | 0.0119 | 0.0119 | 0    |

Table 4: Distances between the summarized guidelines computed using Jaccard distance from Tables 3 and 2.

Tables 2, 3, and 4 represent the process of converting the information in Fig. 2 into a set of distances. These distances are depicted graphically in Fig. 3, where we depict both Jaccard distances between the annotated guidelines, and the number of differing features as per Table 3.

In Section 4, we investigate how close automated tools can replicate this analysis, using document distances typically used in information retrieval and natural language processing.
Figure 3: A pictorial representation of the distances between recommendations, and numbers of differing features, as per Tables 3 and 4. Can we replicate this geometric structure using automated tools? See Section 4 for an answer.
4. Automated analysis of conceptual distances between document guidelines

In the last 10 years, we have witnessed a new era in automated semantic analysis of textual documents [Zhou et al. (2020, in press)]. While no system can claim to really understand natural language, in several domains such as data extraction, classification and question answering, automated systems dramatically improved their performance, and in some cases perform better than humans, due to the unmatched pattern recognition and memorization capabilities of deep neural networks (see e.g. Smith (2020) for an overview).

Some the simplest, easiest to use and effective of the new methods are different types of word and concept embeddings (Mikolov et al. (2013), Pennington et al. (2014), Shalaby et al. (2019), Kalyan & Sangeetha (2020)). Embeddings represent words and concepts as dense vectors (i.e. all values are non-zero), and are a preferred tool to make similarity judgments on the level of words, phrases, sentences and whole documents.

Word embeddings have been widely used to compare documents, and in particular to compute their degree of similarity [Nguyen et al. (2019); Tien et al. (2019)]. Other methods proposed to compute documents similarity are based on using background knowledge [Benedetti et al. (2019)]. This works uses both methods, human knowledge encoded the analysis of the CDC table, and embeddings.

However, before we create our document embeddings Section 4.2 and show the similarity between the representations of full documents and their summaries in Section 4.3, a few words about data preparation are in order.

4.1. Data preparation for all experiments

From the breast cancer screening guidelines listed in the CDC summary document [CDC (2017), USPSTF, ACS, ACP, and ACR guidelines are available in the HTML format, from which we extracted the texts of these guidelines. We used Adobe Acrobat Reader to obtain the text from the pdf format of AAFP,
ACOG, and IARC guidelines. Since the AAFP documents also included preventive service recommendations for other diseases (such as other types of cancers), we added a preprocess step to remove those recommendations, and leaving the parts matching breast cancer.

As mentioned earlier, the manually annotated distances were obtained from the CDC table (referenced above in Fig. 1), which summarized all the breast cancer guidelines. However, the automated computation of conceptual distances was performed on the full guideline documents.

Additional Experiments: We also performed additional experiments with modified views of the full guidelines documents, as enumerated below. This was driven by the fact that the levels of agreement may change if we compute the similarities/distances between selected sentences, which are explicitly related to the statements from the CDC table in Fig. 1. For these additional experiments we split each guideline document into two different subsets:

1. Related: containing sentences that are related to CDC table, by having common concepts, as represented by UMLS concepts. This was done in multiple ways, giving us 6 possible experiments:
   (a) All the sentences in the CDC guideline table were considered as a single document. If a sentence had a number of mutual concepts with that document, that sentence was considered as related sentence.
   (b) If a sentence had minimum number of mutual concepts with at least one statement from CDC table, that sentence was considered as related sentence.
   Different minimum numbers of mutual concept(s) were examined in our experiment, that is the minimum was set at 1, 2, and 3.

2. Unrelated: the other sentences.

Unrelated sentences were not used for these additional experiments.

For full text guidelines (as per Table 1), the result of the experiments are shown in Table 6 are discussed in Sections 5.1 and 6. For full text guidelines
minus **Unrelated** sentences, the result of this experiment in Tables 7 and 8 discussed in Sections 5.3 and 6.

**Concept extraction:** For all experiments, we used MetaMap to extract UMLS concepts and semantic types in sentences. We only considered concepts with informative, in our opinion, semantic types. This meant using concepts related to diagnosis and prevention, for example 'findings,' and not using ones related e.g. to genomics. Our final list had the following: [[diap], [hlca], [dsyn], [neop], [qaco], [qlco], [tmco], [fneg], [geoa], [topp], [lbpr]].

### 4.2. Vector representations and similarity measurements used in the experiments

After data preparation, our approach consists of using a vector representation of each document (guideline), and measuring similarities (or, equivalently, distances) between each pair of the vectors representing the documents. We use two standard measures: cosine similarity and word mover’s distance (WMD, WM distance) we use Gensim as a tool for our experiments.

We experimented with three language models of medical guidelines disagreement: "no concept," conceptualized and BioASQ (see Tables 6, 7 and 8). The first two were trained using the PubMed articles as the training data. The third one used pre-trained BioASQ word embeddings created for the BioASQ competitions.

Our first model, trained on PubMed includes only words (no additional con-
ceptual analysis with MeSH was done). In the second, more complex model, MeSH terms are replaced with n-grams. For example, if breast and cancer appeared next to each other in the text, they are replaced with breast-neoplasms and treated as a concept.

The details of our experiments, with computation steps and algorithms to get numerical values, are shown and discussed in Sections 5.1, 5.2 and 5.3. Because of the large number of experiments we performed it might be best to discuss our best model, before going into the gory details of the experiments.

4.3. Our best model: Using BioASQ embeddings and word mover’s distance

Table 5 shows (unnormalized) WM distances between the seven guidelines using BioASQ embeddings. Fig. 4 shows side by side the geometries of the two graphs: one generated based on human comparisons of the abstracted guidelines, and the second one based on the machine generated representations of the full guideline documents. The similarity is visible in a visual inspection, and will be quantified in the next Section to be about 70%. There, we will also answer two questions:

- How do we measure the distortion between the two graphs?

- Could this similarity of shapes be accidental?

To create Fig. 4 for each metric, a diagram representing the distance between the nodes (guidelines) and a diagram with the labelled edges were drawn, using Python networkx library. All the values were normalized to the same scale to allow visual comparison.

---

1. https://www.nlm.nih.gov/mesh/meshhome.html
2. https://networkx.github.io/
Figure 4: Visual comparison of a the similarity/distance graphs based on human analysis in panel (a), and computer generated from Table 5 in panel (b) suggests similar geometry. As we show in Section 5 this 69% similarity is not accidental; the distortion is about 31% (see Table 6).
Table 5: This table shows the words mover distances between guidelines using BiaAsq embeddings. This is our best model.

5. Details of experiments and analysis of results

With any of the simple semantic metrics discussed above, we obtained higher than random alignment between the expert and the machines judgments of conceptual similarity of the guidelines documents. Our program did not perform the same kind of analysis as the expert; namely, there was no relationship extraction, no focus on who makes the decision etc. (see Section 3 above). The automated analysis used latent factors, that is, the statistical distribution of concepts in the documents was used to establish degrees of similarity.

In this section we first discuss the statistical properties of the experiments to show our models capture statistically significant geometric correspondences. Then we provide the details of the steps we used to obtain the geometry and the statistics. In the last subsections we show results of additional experiments where the Unrelated sentences were removed from full guidelines (per Section 4.1).

5.1. Automated judgements significantly align with the expert judgement

Table 6 shows the results of the experiments with full text of the guidelines. Given seven documents, and the similarities/distances from the features established by the expert (Table 4), the average distortion value, computed in over ten thousand simulations, is 0.523% where we always assume distance of 0 between a document and its copy. (See the diagonal of Table 4).
For our best model, BioASQ_WMD, we have a 69% similarity. As we can see in Table 6, the distortion of this model is about 31%, the average distortion of permutations (using the distances produced by BioASQ_WMD) is 38%, however the standard deviation of the distortions is less than 1%. So, the distance between the our model and the mean is about 7 standard deviations. Therefore, we conclude the correlation between the shapes of the two graphs is extremely unlikely to be coincidental. Hence the model represents a non-trivial similarity.

Moreover, we performed the same kind of analysis using different models, i.e. different embeddings and different measures. And while the distances and distortions change, the chances of similarities arising by accident are always smaller than 1/1000 (four standard deviations from the mean of distortions). By this standard statistical criterion, no matter what measures of distance we use, the similarity between two graphs, one from human annotations and the

| Model                  | Distortion | Distortion of permutations | STD       |
|------------------------|------------|----------------------------|-----------|
| BioASQ_WMD             | 0.31393661 | 0.381378177                | 0.009017982 |
| Conceptualized_WMD     | 0.335044003| 0.391185128                | 0.009293257 |
| NoConcept_WMD          | 0.344571557| 0.388227188                | 0.009099648 |
| BioASQ_CosineSim       | 0.417871068| 0.595697672                | 0.015729293 |
| Conceptualized_CosineSim | 0.334525231| 0.613500756                | 0.016206786 |
| NoConcept_CosineSim    | 0.51399564 | 0.590931627                | 0.015386531 |
| Search                 | 0.543674172| 0.619579945                | 0.017539478 |

Table 6: This table shows the values obtained in multiple experiments. Column 2, Distortion, shows the distortions of graphs produced using corresponding models from from Column 1. Average distortion of per permutation is shown in Column 3. STD is standard deviation of the distortion per permutation. Note that the distortion is somewhat depended on the how we measure distances; however, the the shapes of the distributions are very similar.
other from automated concept modeling, is non-trivial and not accidental. We conclude that vector based representation are capable of detecting conceptual differences, i.e. the types and densities of concepts brought to the writing of medical recommendations (at least in our case study).

5.2. Graph-based method for comparing collections of documents

We use a very natural, graph-clique based method for comparing similarity of two collections of documents. Given two sets of documents, each of the same cardinality, and a mapping between nodes, we compute the percent distortion between the shapes of the two cliques, and the chances that the mapping arose from a random process. In our case the nodes of both graphs have the same names (the names of the medical guidelines), but the shapes of the graphs are different, one coming from human comparisons (Fig. 1) and the other from machine produced similarities/distances. The details of the method, which was used to produce results of the previous subsection (5.1) are below, as a list of steps with references to three simple algorithms listed after the steps.

Specific steps to establish the conclusion that automated judgements significantly align with the expert judgments:

1. We work with the full text guideline documents, prepared as described in Section 4.1
2. We start with building a vectorial representations for each text document, based on one of the word/concept embeddings described in Section 4.1
3. Using WM distance (or cosine similarity) we compute the distances between each of the vectors from the previous step.
4. We put the labels and distances into an adjacency matrix $A_G$ (using Algorithm 1)
5. Using the procedure of Algorithm 2 we compute the distance/distortion between the two labeled graphs, using the matrix obtained in the previous step, and the matrix in Tab. 1. For our best model it is 0.31.
6. We ask the question: could this distortion be accidental? I.e. could another permutation of the graph nodes produce a similar result, that is,
match to a large degree the geometry of the graph created from human annotations, Fig. 4a.

7. To answer this question, we compute the average distortion and the standard deviation, based on all possible permutation of nodes (5040 = 7! permutations). The pseudo-code for this computation is in Algorithm 3.

8. Based on the fact that, per Table 6, the difference between our results and average distortion is seven (or more) standard deviations, we conclude the that the matching of the two geometries is not accidental and is highly significant.

Algorithm 1 Computing Graph of Distances Between Guideline Documents. The output of Algorithm 1 is shown in Fig. 3

Input: Guidelines: a set of guideline documents in textual format.
Model: a model to compute distances between two documents.
Output: $A_G$ — Adjacency matrix of distances between document guidelines.

1: for each pair of documents in Guidelines do
2: Compute the distance between the documents according to Model
3: Put the distance in $A_G$
4: end for
5: return $A_G$
Algorithm 2 Distance or Percentage Distortion between Two Complete Graphs (cliques of the same size).

Note. For example, the distance between the two graphs in Fig.4 is 0.31, equivalent to 31% distortion

Input: Adjacency Matrices $A_1, A_2$ of equal dimensions

Output: Graph distance/distortion $D(A_1, A_2)$, as a value between 0 and 1.

1: Normalize the distances in $A_1$ (by dividing each distance by the sum of distances in the graph) to produce a new adjacency matrix $A_{N1}$
2: Normalize the distances in $A_2$ to produce a new adjacency matrix $A_{N2}$
3: Set the value of graph_distance to 0.
4: for each edge in $A_{N1}$ do
5: Add the absolute value of the difference between the edge length and its counterpart in $A_{N2}$ to the graph_distance
6: end for
7: return $D(A_1, A_2) = graph\_distance$
Algorithm 3 Computing Graph Distortion Statistics.

We are computing the average distortion, and the standard deviation of distortions, under permutation of nodes. The input is two cliques of the same cardinality, with a mapping from one set of nodes to another.

**Input:** Normalized Adjacency Matrices $N_1, N_2$ of equal dimensions

**Output:** Baseline for the graph distance, standard deviation of graph distances under permutations of computed distances.

1: Set the value of $graph\_distances$ to an empty list.

   *We are permuting the labels of graph, leaving the lengths of the edges intact.*

2: for each permutation $N_{2p}$ of the nodes of $N_2$ do

3:    Compute $d = D(N_1, N_{2p})$ using Algorithm 2

4:    Append $d$ to $graph\_distances$

5: end for

6: Set

   \[
   graph\_distance\_baseline = Mean(graph\_distances)
   \]

   \[
   std = StandardDeviation(graph\_distances)
   \]

7: return $graph\_distance\_baseline$, $std$
5.3. Additional experiments

In the previous section we established that the relatively high similarity between conceptual distances in summary guidelines and full guideline documents was not accidental.

Tables 7 and 8 are based on the same type of comparison except based on full guidelines minus Unrelated sentences as described in Section 4. Again we observe that the similarity is not accidental, and that BioASQ embeddings with WM distance seem on average give the best performance.

| Comparing | Min mutual concepts | Model          | Distortion | Distortion of permutations | STD    |
|-----------|---------------------|----------------|------------|---------------------------|--------|
| Sentence  | BioASQ_CosineSim    | 0.526380991    | 0.602890558| 0.011755664               |
|           | Conceptualized_CosineSim | 0.635564030 | 0.646721788| 0.011417206               |
|           | NoConcept_CosineSim | 0.620957319    | 0.646906954| 0.011131221               |
|           | NoConcept_WMD       | 0.352402031    | 0.383526047| 0.00650777                |
|           | Conceptualized_WMD  | 0.39526888     | 0.390059373| 0.00626225                |
|           | BioASQ_WMD          | 0.336903254    | 0.38475148  | 0.00649344                |

Table 7: This table shows the values obtained in additional experiments, where full document guidelines were modified by attending to concepts in sentences (see Section 4.1). Column 2, refers to the number of concepts overlapping with summaries. Distortion, shows the distortions of graphs produced using corresponding models from Column 1. As before, in Tab. 6 the distortion depends on the how we measure the distances; however, the shapes of the distributions are very similar.

Note the potentially important observation about Tables 6, 7 and 8: they jointly show that the property we investigate, i.e. the conceptual distances between guidelines, is indeed geometric, and therefore the word ‘distances’ is not merely a metaphor. The correspondence between the two graphs is preserved no matter how we set up the experiments. That is, as with geometric properties
## 6. Discussion, future work and conclusions

Notice that unlike our earlier work [Zadrozny et al., 2017], in this article we have not performed any logical analysis of the guidelines. Thus the similarities and differences in distances are conceptual, they reflect the conceptual knowledge of the writers (through the word and concept embeddings of their texts), and not the style or formalization of recommendations. These non-accidental...
and substantial similarities support our thesis that automated methods can be used for conceptual analysis of guidelines, and in particular can capture some the near-peer epistemic relations discussed earlier.

We will start our discussion of the results by listing the assumptions that underlie our work; we then argue that the results presented here support the near-peer epistemic model motivating large parts of this research. We conclude with a discussion of some of the limitations of this work and its possible extensions.

6.1. Our assumptions about modeling epistemic stances using NLP tools

Before we summarize what we did, and discuss gaps in this work (to be followed up by further research), let’s review our assumptions, and put the work in a larger context.

This article provides support for the hypothesis that epistemic stances among medical societies can be to a substantial degree analyzed using natural language processing and machine learning tools. This hypothesis can be decomposed into three ideas, which we discuss briefly below:

1. Knowledge (i.e. knowledge claims) brought by the authors of guidelines to the writing process is reflected in the concepts they use.
   — This is intuitively plausible.

2. This knowledge to a large degree determines the types of recommendations that will be produced.
   — Again, we can intuitively hypothesize such relation. However, in principle, if the different groups of experts had been epistemic peers (Lackey (2014)), i.e. possessed the same knowledge of fact and methods, then, as epistemic peers, they would have produced similar recommendations (if the recommendations were to be inferred from their knowledge, only). As analyzed by the CDC, and shown here in Fig. 1, the recommendations differ. Therefore we either have to reject the complete epistemic peerhood,
or assume other influences. The results presented in the sections above strongly suggest some type of near-peerhood.

3. We assume that vector representations are appropriate for the texts of the guidelines.

A few things might be worth keeping in mind, about the above points. Especially, before we discuss near-peerhood and limitations of this work:

— We prove the near-peer relationship by observing the geometric similarity between the graph of recommendations and the distances between the vectors representing the full guideline documents (Fig. 4).

— As discussed earlier, in Section 3, given the partial overlaps of the recommendations, the near-peer hypothesis makes sense. In the subsequent sections we have shown these partial overlaps can be recovered from the concepts used in each guideline document.

— Note, our vector models cannot account for other factors, e.g. values, different moral utilities, risk perception [Lie et al. (2017)], or economic incentives.

6.2. Modeling epistemic near-peers

A theoretical model of expert disagreement was proposed in Christensen et al. (2013) and Lackey (2014), and analyzed in Garbayo (2014) and Garbayo et al. (2018). Earlier, the epistemology of agreement/disagreement and expertise was discussed in Goldman (2001), where the "ideal" model of expertise (e.g. unlimited logical competence) is replaced with a study of situations with epistemic constraints. More recently, Grim (2019) discusses computational models of epistemology, but does not focus on specifically on disagreement.

The complexity of normative standards in considering truth in the context of disagreement is discussed in Grim et al. (2017). In the data science setting, medical expert disagreement and an adjudication process, in analyzing time series data, is described in Schaekermann et al. (2019a) and Schaekermann et al. (2019b); there, the authors observe that this process does not eliminate the
disagreement, although it reduces its magnitude. Interestingly, the differences in experts backgrounds increase degree of disagreement. This looks to us as another argument for a near-peer model.

The epistemic model of multi-criteria expert disagreement Garbayo et al. (2018) can be applied to disagreements among medical guidelines, especially using the near-peer paradigm, which suggests that medical experts may disagree as a result of being not quite the ideal epistemic peers among themselves.

Accepting the assumptions discussed above in Section 6.1 as reasonable, we have shown that epistemic near-peerhood can be modeled using computational tools. That is, that conceptual stances and disagreements among near-peers regarding medical guidelines can be to a large degree observed in the distances between the guideline documents, which are represented as vectors in high dimensional spaces.

Thus, the near-peers share comparable (claims to) knowledge, but also hold relevant differences in their baseline assumptions and sub-areas of expertise and resources. These differences were once idealized in the literature via the epistemic peers model (Lackey (2014)), but now as argued in this work – they can be studied computationally using natural language processing and machine learning tools.

The importance of this new approach lies in its contribution to computational approaches to epistemology (Grim et al. (2017)), which could provide a complementary representation to the standard formal analysis, represented by Bayesian and formal epistemology.

This might be particularly important for healthcare, intelligence operations, disaster preparedness and others, where there are very tangible consequences of accepting ‘wrong’ epistemic stances. While what is wrong can only be seen in retrospect, disagreements and differences in epistemic stances can be modeled with deeper (formal) and faster (computational) analysis, as a part of decision

\[1\) https://plato.stanford.edu/entries/epistemology-bayesian/
\[2\) https://plato.stanford.edu/entries/formal-epistemology/}
The experiments reported here show the potential of computational tools to provide a different kind of analysis, and their power to make distinctions not possible before, increasing the resolution of our analysis of disagreements – just like the microscope allowed human eye to see previously unseen details.

While the intuitive case for positing the near-peer hypothesis is clear, this case would be strengthened by additional experiments. The most obvious ones would be trying replicate our findings for other guidelines and other types of expert opinions, e.g. in public policy.

Even though our experiments have shown strong influence of concepts on recommendations, the correlation is not perfect, and more accurate models should be possible, for example, using embeddings from larger models such as the universal encoder (Cer et al. (2018)) or the GPT family (Radford et al. (2019), Brown et al. (2020)). Another option could be in extending the epistemic model by connecting the authors of the guidelines to their other publications, and measuring not the distances between the guideline documents, but explicitly between the groups of authors, represented e.g. by a vector average of their relevant publications. We could then see whether such distances between groups can be translated (in the geometric sense of the word, say as an affine transformation) into distances between the guidelines, and produce a more accurate model.

6.3. Limitations of this work and future directions

As with any approximate NLP tool, the usefulness of the model depends on its accuracy. In our case, using a very elementary set of tools, we have shown we can approximate the differences in recommendation with about 70% accuracy. We are sure this accuracy can be improved, and we are actively working towards that goal. In particular, we (or someone else) should be able to use an architecture modeled after our previous system, Zadrozny et al. (2017), to add explicit detection of contradictions to the model. Please note such effort would be highly non-trivial, since the search space of the CDC summary (as in
Fig. 1 is much smaller than the search space of the guideline texts in Tab. 1 and additionally the syntax of the actual documents is much more complex than the syntax of the tables.

An obvious extension of this work would be to compare groups of guidelines, e.g. European medical societies vs. US medical societies. We know that for years their recommendations, e.g. on management of blood cholesterol, differed.

We used word and concept embeddings as a representation of such conceptual stances, but we have not experimented with other representations such as more complex word and document embeddings (Devlin et al. 2018; Peters et al. 2018; Cer et al. 2018). Neither have we tried to create more subtle semantic representations based on entity and relationship extraction (e.g. check Zhu et al. 2017), and on formal modeling of contradictions, like the ones discussed in Zadrozny et al. 2017; Zadrozny & Garbayo 2018; Garbayo 2019. Such extensions require new methods that would operate, e.g. perform inference, on the document level and not only on the sentence level. Both of these are our active areas of research.

Another potential limitation of our work has to do with our using only one expert for judgment of conceptual differences between different documents, and leveraging the work of the CDC to tabulate the differences in the recommendations. As discussed in Section 3, given the simplicity of the CDC table in Fig. 1 and with only binary and obvious differences, having only one expert does not seem to be a problem. However, clearly tabulated comparisons of guidelines, like the ones we started with [CDC 2017] are not common. Thus, translating this work to other set of guidelines will not be trivial. Also, because when working with full text, longer documents, there is more potential for disagreement about building summaries of recommendations.

There are other ways of extending the current work. In this article we have dealt with a very simple model of near-peer disagreement: we computed semantic distances between several pairs of breast cancer prevention guidelines, using different automated methods, and compared the results an expert opinion.

However, we have not incorporated at any of the logical properties of dis-
agreement in the distance measures, for example unlike our previous work. Zadrozny & Garbayo (2018) Zadrozny et al. (2017) Garbayo (2019) we have not made a distinction between contradictions and disagreements. Thus, some of our ongoing work is about incorporating logical structures of documents and finding metrics reflecting them.

**Description of individual contributions to this research:** HH performed the majority of experiments discussed in this article, and substantially contributed to writing. The idea to investigate computationally the concept of epistemic near-peers comes from LG, as well as the annotations allowing us to use distances, and not just qualitative measures; she also contributed with graphing and overall analysis and writing. SG performed several experiments allowing us to progress with the evaluation. WZ provided the overall supervision, suggested the clique based method for computing distortions, and did the majority of the writing. The overall progress was achieved during multiple team discussions, with equal contributions.

**References**

Action, A. P. (2017). Summary of recommendations for clinical preventive services. *American Academy of Family Physicians*, .

American College of Obstetricians-Gynecologists (2011). Practice bulletin no. 122: Breast cancer screening. *Obstetrics & Gynecology, 118*. URL: [https://journals.lww.com/greenjournal/Fulltext/2011/08000/Practice_Bulletin_No__122__Breast_Cancer_Screening_40.aspx](https://journals.lww.com/greenjournal/Fulltext/2011/08000/Practice_Bulletin_No__122__Breast_Cancer_Screening_40.aspx)

Benedetti, F., Beneventano, D., Bergamaschi, S., & Simonini, G. (2019). Computing inter-document similarity with context semantic analysis. *Information Systems, 80*, 136–147.

Bowles, J., Caminati, M., Cha, S., & Mendoza, J. (2019). A framework for automated conflict detection and resolution in medical guidelines. *Science of Computer Programming, 182*, 42–63.
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

CDC (2017). *Breast Cancer Screening Guidelines for Women*. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Christensen, D., Lackey, J., & Kelly, T. (2013). *The Epistemology of Disagreement: New Essays*. Oxford University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garbayo, L. (2014). Epistemic considerations on expert disagreement, normative justification, and inconsistency regarding multi-criteria decision making. In M. Ceberio, & V. Kreinovich (Eds.), *Constraint Programming and Decision Making* (pp. 35-45). Cham: Springer International Publishing. URL: https://doi.org/10.1007/978-3-319-04280-0_5.

Garbayo, L. (2019). Dependence logic & medical guidelines disagreement: an informational (in) dependence analysis. In *Logic Colloquium 2019* (p. 112).

Garbayo, L., Ceberio, M., Bistarelli, S., & Henderson, J. (2018). On modeling multi-experts multi-criteria decision-making argumentation and disagreement: Philosophical and computational approaches reconsidered. In M. Ceberio, & V. Kreinovich (Eds.), *Constraint Programming and Decision Making: Theory and Applications* (pp. 67-75). Springer International Publishing. URL: https://doi.org/10.1007/978-3-319-61753-4_10.

Garbayo, L., Zadrozny, W., & Hematialam, H. (2019). Measurable epistemological computational distances in medical guidelines. *16th International Congress on Logic, Methodology and Philosophy of Science and Technology*, 35.
Garbayo, L., Zadrozny, W., & Hematialam, H. (2019a). Converging in breast cancer
diagnostic screening: A computational model proposal. _Diagnosis, 6_, eA60.

Goldman, A. I. (2001). Experts: Which ones should you trust? _Philosophy and
phenomenological research, 63_, 85–110.

Grim, P. (2019). Modeling epistemology: examples and analysis in computational
philosophy of science. In _2019 Spring Simulation Conference (SpringSim) (pp. 1–
12). IEEE._

Grim, P., Modell, A., Breslin, N., Mcnenny, J., Mondescu, I., Finnegan, K., Olsen,
R., An, C., & Fedder, A. (2017). Coherence and correspondence in the network
dynamics of belief suites. _Episteme, 14_, 233–253.

Hematialam, H., & Zadrozny, W. (2016). Text mining of medical guidelines. In _Proc. of
the Twenty-Ninth Intern. Florida Artificial Intelligence Res. Soc. Conf.; FLAIRS-29
Poster Abstracts_. AAAI.

Jørgensen, K. J., & Bewley, S. (2015). Breast-cancer screening viewpoint of the iarc
working group. _N Engl J Med, 373_, 1478.

Ju, M., Nguyen, N. T., Miwa, M., & Ananiadou, S. (2020). An ensemble of neural
models for nested adverse drug events and medication extraction with subwords.
_Journal of the American Medical Informatics Association, 27_, 22–30.

Kalyan, K. S., & Sangeetha, S. (2020). Secnlp: A survey of embeddings in clinical
natural language processing. _Journal of biomedical informatics, 101_, 103323.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to
document distances. In _International conference on machine learning_ (pp. 957–966).

Lackey, J. (2014). Taking religious disagreement seriously. In L. Callahan, &
T. O’Connor (Eds.), _Religious Faith and Intellectual Virtue_ (pp. 299–316). Oxford
University Press. doi:\[10.1093/acprof:oso/9780199672158.003.0014\]

Lee, C. H., Dershaw, D. D., Kopans, D., Evans, P., Monsees, B., Monticciolo, D.,
Brenner, R. J., Bassett, L., Berg, W., Feig, S. et al. (2010). Breast cancer screening
with imaging: recommendations from the society of breast imaging and the acr on
the use of mammography, breast mri, breast ultrasound, and other technologies for
the detection of clinically occult breast cancer. *Journal of the American college of radiology*, 7, 18–27.

Lee, K. C., Udelsman, B. V., Streid, J., Chang, D. C., Salim, A., Livingston, D. H., Lindvall, C., & Cooper, Z. (2020). Natural language processing accurately measures adherence to best practice guidelines for palliative care in trauma. *Journal of Pain and Symptom Management*, 59, 225–232.

Lie, R. K., Chan, F. K., Grady, C., Ng, V. H., & Wendler, D. (2017). Comparative effectiveness research: what to do when experts disagree about risks. *BMC medical ethics*, 18, 42.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, 182, 104842.

Oeffinger, K. C., Fontham, E. T., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y.-C. T., Walter, L. C., Church, T. R., Flowers, C. R., LaMonte, S. J. et al. (2015). Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama*, 314, 1599–1614.

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (aime) conferences: A review of research themes. *Artificial intelligence in medicine*, 65, 61–73.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1, 9.
Rehurek, R., & Sojka, P. (2011). Gensimstatistical semantics in Python. *Retrieved from gensim.org.*

Rospocher, M., Corcoglioniti, F., & Dragoni, M. (2019). Boosting document retrieval with knowledge extraction and linked data. *Semantic Web, 10,* 753–778.

Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., & Law, E. (2019a). Capturing expert arguments from medical adjudication discussions in a machine-readable format. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 1131–1137).

Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., & Law, E. (2019b). Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction, 3,* 1–23.

Seneviratne, O., Das, A. K., Chari, S., Agu, N. N., Rashid, S. M., Chen, C.-H., McCusker, J. P., Hendler, J. A., & McGuinness, D. L. (2019). Enabling trust in clinical decision support recommendations through semantics.

Shalaby, W., Zadrozny, W., & Jin, H. (2019). Beyond word embeddings: learning entity and concept representations from large scale knowledge bases. *Information Retrieval Journal, 22,* 525–542.

Siu, A. L. (2016). Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of internal medicine, 164,* 279–296.

Smith, N. A. (2020). Contextual word representations: putting words into computers. *Communications of the ACM, 63,* 66–74.

Solomon, M. (2015). *Making medical knowledge.* Oxford University Press, USA.

Sutherland, K., & Levesque, J.-F. (2019). Unwarranted clinical variation in health care: Definitions and proposal of an analytic framework. *Journal of evaluation in clinical practice.*

Tien, N. H., Le, N. M., Tomohiro, Y., & Tatsuya, I. (2019). Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. *Information Processing & Management, 56,* 102090.
Tsatsaronis, G., Balikas, G., Malakasiotis, P., Pastalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D. et al. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics, 16*, 138.

Tsopra, R., Lamy, J.-B., & Sedki, K. (2018). Using preference learning for detecting inconsistencies in clinical practice guidelines: Methods and application to antibiotherapy. *Artificial intelligence in medicine, 89*, 24–33.

Wilt, T. J., Harris, R. P., & Qaseem, A. (2015). Screening for cancer: advice for high-value care from the american college of physicians. *Annals of internal medicine, 162*, 718–725.

Zadrozny, W., & Garbayo, L. (2018). A sheaf model of contradictions and disagreements. preliminary report and discussion. *arXiv preprint arXiv:1801.09036*, .

Zadrozny, W., Hematialam, H., & Garbayo, L. (2017). Towards semantic modeling of contradictions and disagreements: A case study of medical guidelines. *Proc. 12th International Conference on Computational Semantics (IWCS); arXiv preprint arXiv:1708.00850*, .

Zhou, M., Duan, N., Liu, S., & Shum, H.-Y. (2020, in press). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering, .*

Zhu, Q., Li, X., Conesa, A., & Pereira, C. (2017). GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics, 34*, 1547–1554. doi:10.1093/bioinformatics/btx815