UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HIMEAN:

A HYGENE APPROACH TO SEMANTIC ANALYSIS IN A MEDICAL DECISION-MAKING TASK

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

APRIL MARTIN
Norman, Oklahoma
2014

HIMEAN: A HYGENE APPROACH TO SEMANTIC ANALYSIS IN A MEDICAL DECISION-MAKING TASK

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Rick Thomas, Chair

_____
Dr. Scott Gronlund

_____
Dr. Lynn Devenport

_____
Dr. Michael Wenger

_____
Dr. Dean Hougen

Acknowledgements

"Acknowledgement" seems a word too far removed from what I need to express. I need to say "thank you" to both of my parents for their unwavering support: To my mother for her completely unrealistic and positive impression of my capabilities, and to my father for his words of wisdom. I need to express my gratitude to the pillars of knowledge that are my committee members for not only their guidance as pertains to this dissertation, but even more so for their many years of tutelage and the enriched education they provided me, which I would not be the person I am without. I need to share the depths of my gratitude for Rick Thomas, my advisor, my mentor, and my friend, especially. He was my inspiration for embarking on the journey of graduate school, and remains my inspiration for his dedication to integrity, knowledge, and his students. I am so grateful for all his effort and instruction these past few years and for his diligence in helping me succeed even in the face of sometimes great difficulty. There is no doubt this accomplishment would not have been possible without him. Finally, there is one more individual I need to mention and I think I understand at this moment why the term acknowledgement is used above. It is because it is not possible to adequately express the depth of heartfelt appreciation for certain of those in our lives. For me, that person is Zhanna Bagdasarov. Above and beyond all the support from those I have mentioned, she has been there to carry me through my lowest lows and the first there to celebrate the highs. She has been my constant companion, my reviewer, my cheerleader, and my (sometimes) unforgiving taskmaster. She has always been whatever I needed most, whenever I needed it, and I am truly grateful. She has been and will forever be my best friend, and I am so thankful for all that entails.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation makes an exploratory comparison between two semantics models, Latent Semantic Analysis (LSA) and a newly introduced HiMean model based on the HyGene architecture, in a medical decision-making context. Emphasis is placed on using real-world, human decipherable input to produce rational diagnoses. Base rate information is manipulated as a proxy to expertise or learning in different information environments, and outcomes on decision measures are examined. Model performance in terms of correct probe or query identification, alternative hypothesis generation, probe degradation resilience, probability judgments, and diagnostic capability is evaluated. Multidimensional scaling is also employed to investigate two-dimensional projections of the models' respective semantic spaces. Experimental outcomes reveal that both the LSA and HiMean models, as well as HiMean variants perform well in a variety of conditions. The models produce performance tradeoffs between each other in terms of accuracy, judgment calibration, and robustness to probe error, though not in diagnostic capability. The models are demonstrated to be capable of utilizing non-trained data and producing identification accuracies up to 80%. Generally, both LSA and HiMean prove to be capable decision architectures with a wide variety of potential applications. Some thought is given to future work dedicated to a multi-agent decision system which capitalizes on the strengths of both models.

Introduction

Models, by their definitions, are simplified versions of more complex systems (Rodgers, 2010). By necessity, their creators must make decisions regarding what to include and exclude from the models while still retaining their important explanatory features. In models of human memory, a rather important area of consideration is often sacrificed in the name of simplicity or in the pursuit of a tightly-defined scope. That is, it is commonplace to represent the mental information in memory that actually maps to reality using an abstract grouping of features (usually numbers). For example, traces or images in memory might be represented as integers indicating varying memory strengths, as in the Search of Associative Memory (SAM—Raaijmakers & Shiffrin, 1981), for particular items, or as vectors of concatenated feature values indicating the presence, absence, or lack of information about specific attributes for the memory item (MINERVA 2—Hintzman, 1984, 1986, 1988; MINERVA-DM—Dougherty, Gettys, & Ogden, 1999). In such instances it is presumed that items to be represented in memory can invariably be decomposed into signals which allow the various components of these cognitive models to operate. However, the act of explicitly employing these model representations in a real-world, everyday task is seldom accomplished. Thus, while it is possible, and perhaps even likely, that the assumptions of these models hold if real-world information is appropriately translated and deployed in a task, they remain empirically untested and, further, the question as to whether such conversion is even possible remains.

In opposition to the information abstract representation schemes often employed by cognitive memory models, computational semantic models have been touted for the

ability to reduce complex, multidimensional semantic spaces of real-world phenomena into representations manageable for various computational and analytical processes (Landauer & Dumais, 1997). It has even been proposed that the human memory system operates on a similar process of abstracting meaningful information networks, though not necessarily explicitly based on frequency of semantic associations (Landauer & Dumais, 1997). In theory, then, semantic models lend themselves to deployment within cognitive models purported to explain human memory processes. However, modelers concerned with computational semantics are not often interested in explicitly tying them to feasible memory models or the applications thereof. Thus, in one hand we have memory models which do not necessarily concern themselves with how their chosen representational systems reflect real-world information, and in the other we have semantic representation systems that are generally not concerned with their assimilation into models of cognitive processes. Additionally, even given the integration of semantic and cognitive psychological models within a memory-theoretic framework, little has been done in the way of demonstrating how they might contribute to decision-making processes in an applied setting.

The workable integration of these ideas in a functional, "real-world" application is the subject of this dissertation. After a brief overview of the models involved and their underlying mechanics, I explicate the rationale and method for directly translating a domain which has been semantically decomposed using semantic analysis techniques into a feasible memory representation operationally governed by the HyGene (**Hy**pothesis **Gene**ration—Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, Harbison, & Sprenger, 2008) cognitive model. I introduce the HiMean model, named

for both the fact that it investigates **Hi**gher order, latent relationships as well as operates on the semantic **Mean**ing of associated concepts. This amalgamated HiMean model is investigated by deploying it during a diagnostic decision-making task operating over a realistic information ecosystem. The performance of this model as it compares to a "decision model" based on traditional Latent Semantic Analysis (LSA) is discussed along with the effects of various model manipulations and concomitant implementation considerations.

HyGene Overview

HyGene is a cognitive process model that explains the dynamics of memory activation, memory retrieval, hypothesis generation, and information search and judgment (Dougherty et al., 2010; Thomas et al., 2008). Under HyGene, information in the external or internal (i.e., physical or mental) environment serves as a cue to the memory retrieval processes which are then responsible for furnishing the working memory construct with information requisite in rendering judgments or further testing the environment for additional information. Figure 1 serves as a visual illustration of HyGene's machinery, demonstrated as a series of iterated steps (though they are not considered to be necessarily serial in execution).

1. The experience of some information (Data observed, or D_{obs}) in the environment activates related traces in episodic memory. Episodic memory is defined as the long-term storage of an individual's experiences. The episodic memory in HyGene, as in real life, contains imperfect traces (records) of those experiences. That is, individuals may fail to properly encode into memory some features of the observed

event. Importantly for this work, episodic memory is also presumed to contain the base rate information for those traces. That is, there is some implicit encoding of the frequency of co-occurrences between traces in memory and the data in the environment (Gigerenzer, Hoffrage, & Kleinbölting, 1991). This enables the architecture to respond to observed data with those traces most frequently associated with similar observations in the past. Thus, the probability of any trace activating as a response given the D_{obs} is a function of the strength of the frequency relationship between the trace and the data, where higher frequencies (associations) lead to higher activation strengths.

2. When the activation value of a trace exceeds a threshold of activation, a probe representing the strongest (most frequent) trace hypotheses is generated. This probe is referred to as unspecified because it has not yet been linked to semantic information and its location and membership within the semantic memory space cannot be explicitly determined.

3. Semantic classification of the unspecified probe is accomplished by matching the probe to semantic memory. Semantic memory is also part of long-term memory and contains a record not only of the individual's semantic associations to past directly experienced information, but also information that is more general and abstract. Semantic classification of the unspecified probe allows for the identification of the most representative hypotheses in memory according to their similarity to the probe. In short, a hypothesis generated in response to D_{obs} is comprised of meaning from semantic memory and, due to its encoding of frequency information, relevance (likelihood) from episodic memory. However, because traces in episodic memory are imperfect and

because semantic memory is dependent upon the probe created from those traces, it is possible for the model to generate incorrect inferences just as humans do.

4. Hypotheses (whether correct or not) with activation strengths (A_s) that are sufficiently high ($>Act_{MinH}$) are considered by the individual as explanations for D_{obs} by gaining access to a construct labelled the Set of leading Contenders (SOC). The SOC is HyGene's working memory component in that it is a temporary activation of a subset of long-term memory, is capacity limited, and is the construct in which mental information can be manipulated and must be actively maintained in order for its contents to be considered. Here, candidates generated from semantic memory vie for limited cognitive resources and are retained according to their individual associative strengths. The minimum activation strength required for admittance to the SOC is set to be equal to the activation strength of the weakest contender currently in the SOC (Act_{MinH}). If the SOC has reached capacity and a contender stronger than the weakest candidate arises, the weakest candidate is displaced by the contender. In this way, the SOC contains only the strongest (most likely) explanations for the observed data. However, it is important to note that task characteristics such as limited time or dividing the individual's attention may prevent opportunities for the best possible hypotheses to enter the SOC by constraining the individual's ability to consider all relevant information. Further, as working memory capacity is an individual difference, it also potentially moderates an individual's ability to generate and consider ideal hypotheses.

5. Competition for consideration continues until the conditions of a stopping rule are met. This stopping rule can be external (e.g., a time limit) or internal (e.g., encountering a certain number of retrieval failures where hopeful contenders had

activation strengths insufficient to surpass those of the candidates already in the SOC). In Figure 1, the parameter T can be thought of as the unit of measure used in a stopping rule, and T_{MAX} as the condition satisfying the rule. Attempts are made to populate the SOC with better candidates until $T = T_{MAX.}$

6. The probability of any hypothesis in the SOC as the best explanation for D_{obs} is defined by its activation strength relative to the activation strengths of all the other SOC candidates. Once the SOC has been populated and the conditions of the stopping rule have been met, a posterior probability judgment conditional on the hypotheses in the SOC can be rendered, or a search for further external information that is contingent on the focal hypotheses (hypothesis-guided search) can be conducted. Thus, further information search is engaged in differentially based upon the contents of the SOC.

For the present work, it is especially important to understand how the memory retrieval processes of HyGene give rise to subsequent judgments of the probability of any hypothesis as the best explanation for the observation. Specifically, the likelihood of any hypothesis being generated as a potential response is a function of its memory strength which, in turn, is derived from the base rate frequency of occurrences of those traces in the past. Ultimately, this means that the more frequently a hypothesis co-occurs with a piece of data, the higher the probability that hypothesis will be generated in response to similar situations in the future. Bearing this principle of cohesive covariation in mind, I move to a discussion of semantic analysis.

Semantic Analysis Overview

Semantic analysis is a method for extracting the meaningful relationships between various elements in often complex domains. When examining such relationships in language, semantic analysis is frequently applied to textual data sources. The basic premise is that the statistical properties of word co-occurrences convey something about the meaning of and relationship between those words. Words that frequently appear together within specified contexts are presumably concerned with some of the same subject matter. For example, due to the frequency with which they appear together, dog and cat would seem to share a relationship. Conversely, words that rarely appear together may also be highly semantically related. For example, while Great Dane and Rottweiler are both large breeds of dog, they may be unlikely to be discussed together, as a text describing either of them would most likely be focused on one or the other. However, when they are discussed, there is a great deal of overlap between their contexts which suggests that the two are, in fact, highly semantically related.

While similarity between words of the first instance are perhaps easily assessed by comparing simple counts of their relative occurrences within the same contexts, recognizing the higher order relationships that exist between concepts, as in the second example, are a little less straightforward. Fortunately, a number of analytical models have been employed to accomplish this task (e.g., Vectorspace--Salton, Wong & Yang, 1975; Latent Semantic Analysis--Kintsch, McNamara, Dennis, & Landauer, 2006; Sparse Independent Components Analysis--Bronstein, Bronstein, Zibulevsky, & Zeevi,

2005; Topics--Griffiths & Steyvers, 2002; Sparse Nonnegative Matrix Factorization--Xu, Liu, & Gong, 2003; Constructed Semantics Model--Kwantes, 2005; and the Bound Encoding of the Aggregate Language Environment Model--Jones & Mewhort, 2007). Each of these models represents semantic spaces uniquely, with their varying qualities chosen according to different computational and theoretical motivations.

Perhaps the quintessential approach to these types of analyses is LSA described by Landauer and Dumais (1997). These authors demonstrated that the higher order, latent, relationships between words could be captured by LSA using the properties of matrix mathematics. In LSA, a matrix record of the frequencies with which words appear in certain contexts is first created. Here a "context" is defined as an individual corpus of text. For example, a corpus could be comprised of a paragraph, the text of an entire book, the contents of a particular website, or an article on a specific topic within an encyclopedia where each article is considered a separate corpus. A collection of these corpora (multiple paragraphs, books, websites, articles, etc.) capture the entire semantic space of interest and a list of the words appearing in the corpora is made. Usually, this list is preprocessed in some way to exclude words that do not contain much semantic information ("stop words") in order to reduce the statistical noise they would otherwise introduce to the analyses. For example, words that appear repeatedly in every context (a, and, the, etc.) do not tell us much about their meaning. Once the final word list is derived, the frequency of each word's appearance within each corpus (document) is recorded in an M x N matrix where the rows are the words and the columns are the documents. Following the example set in Landauer and Dumais (1997),

an example matrix consisting of 1,000 documents and 30,000 words is shown in Figure 2 where x represents the number of times a word appears in a document:

Depending on the specific analysis method being deployed, the values in the cells of the matrix are then sometimes transformed according to various techniques. In LSA, each cell's value is given by the formula

$$\frac{\ln(x+1)}{-\sum_1^d \left(\frac{\ln(x_i+1)}{\sum_1^d \ln(x_i+1)}\right) \ln\left(\frac{\ln(x_i+1)}{\sum_1^d \ln(x_i+1)}\right)}$$

where *d* is the total number of documents in the corpora. The theoretical motivation for this transformation was that the log function models growth in simple learning and that by dividing this log-transformed term by its entropy over the entire corpora, $-\sum_1^d \left(\frac{\ln(x_i+1)}{\sum_1^d \ln(x_i+1)}\right) \ln\left(\frac{\ln(x_i+1)}{\sum_1^d \ln(x_i+1)}\right)$, each cell is weighted by the amount of information the word conveys according to its context specificity (Landauer & Dumais, 1997). Following the formula application, the matrix is linearly decomposed into its principal components by way of singular value decomposition (SVD). The result is three derived matrices which can be multiplied together in order to reconstruct the original matrix. One of these three matrices is a condensed diagonal matrix of singular values representing the scaled strengths of all the intercorrelations of the words and documents. With the original *term* x *document* matrix thus decomposed, it becomes possible to remove singular values accounting for the smallest contributions from the diagonal matrix by replacing them with zeroes and reconstruct an approximation of the original matrix where individual elements are mathematical composites of the singular vectors and the newly reduced singular values.

The reconstructed matrix is now comprised of word (row) vectors that have "lost" information associated with the removed dimensions, which serves to increase the similarity of related word vectors and reduce the similarity of unrelated word vectors. This method essentially exposes which contexts are the best (most informative) representations of the word relative to all the other words. The number of dimensions in which to represent the space determines the similarity of the vectors. If too many dimensions are retained, the surface information is not diluted and the abstract relationships between items remain obscured. Conversely, because reducing dimensions reduces variance in the matrix, discarding too many dimensions results in a collapse of the similarity structure and distinguishing between vectors becomes meaningless. The choice of dimensionality is therefore an important consideration. In properly constrained space, the greater the similarity between different row vectors, the more likely they are to share semantic properties and the more dominant those vector co-relationships are relative to the other term vectors. Various methods for computing word vector similarity can be deployed. Cosine similarity is often chosen as a metric because of its suitability in determining the angular difference between vectors occupying high-dimensional spaces.

While LSA is certainly a powerful tool with a simple yet effective representation for textual analysis, its ability and versatility to robustly model psychological processes in a variety of domains is rather limited without modification. For example, in the case of information retrieval (where LSA becomes LSI, or latent semantic indexing), querying the covariance matrix (database) results in an isomorphic and static return structure. This is a desirable quality in variety of circumstances especially where

reliability is crucial, but it would not account for the more variable and errorful human memory retrieval processes without perturbing the characteristics of either the probe or the database itself, or entrenching it within a different operational structure. This is a significant limitation despite LSA's ability to otherwise rather elegantly uncover semantic relatedness.

The HiMean Model

Dimension reduction via SVD is not the only way to form meaningful representations of semantic word spaces, however. In 2005, Kwantes explicated a representational system called the (Constructed) Semantics Model that formed semantic spaces using some of the cognitive machinery behind the MINERVA-2 memory model. Here, MINERVA-2's memory traces, which are normally comprised of feature vectors of 1s, 0s, and -1s representing the presence, lack of information about, or absence of specific features within that trace, were replaced with word vectors derived from their tabulated occurrences within specific contexts, as in LSA. Kwantes (2005) had to modify the way that MINERVA-2 trace activation (similarity to probe word) weights are derived and applied because of differences in representation (the Semantics model trace vectors were not constrained to 1s, 0s, and -1s). In LSA, while it is dimensionality reduction that dilutes the effects of less informative vector elements and creates a more coherent similarity matrix, the Semantics model accomplishes this by eliminating from a composite trace vector the contributions of those memory traces that fail to reach a requisite threshold of activation. Despite the Constructed Semantics model's minor differences between both LSA's dimensionality reduction and MINERVA-2's

representation and activation calculations, it was nonetheless able to produce semantic memory composites that were useful in constructing a meaningful multidimensional space whose member vectors' mathematical distances from each other conveyed their latent semantic associations.

Given that HyGene's memory representation is based on MINERVA-2, it should, as with the Constructed Semantics model, be amenable to utilizing semantic spaces derived from real-world contexts. Unlike MINERVA-2 and the constructed semantics model, however, HyGene has two memory systems in operation. One of the strengths of this aspect of HyGene that make it particularly suitable for probability judgment during decision-making is the potential to benefit from the base rate frequency information stored in its episodic memory component. Encoding base rate information regarding document prevalence into HyGene's episodic memory can serve to change the activation strengths of the associated semantic traces rather than simply relying on word frequency counts across separate documents alone to convey the importance or relatedness of memory items as is done in the standard sematic models. From a scientific modeling perspective, this allows for memory trace (word or document vector) frequency manipulations to be carried out in a psychologically-principled manner in order to determine whether there is an advantage to cognitive models within a decision framework when compared to standard computational semantics. Integrating semantics and HyGene allows for further testing of the model's theoretical competence in decision tasks. It also becomes possible under these conditions to examine more closely the differential effects of more sophisticated

memory probes, the retrieval dynamics of specific pieces of information, and their potential impacts on hypothesis testing and diagnosis.

Comparison models. Because the various model implementation considerations of LSA and HiMean may have a differential impact on their performance under varying conditions, I aimed to conduct an exploratory study designed to examine this potential. Specifically, I wanted to compare the performance of the HiMean model both to variations on the HiMean model itself (i.e., a psychologically unconstrained "ideal" version) and to LSA (which does not have components specified according to psychological principles) on various measures. I expect that the more human-like psychological aspects to the HiMean model can be both a boon and a hindrance to its performance with respect to the psychologically indifferent mechanisms of LSA. More specifically, the stochastic retrieval processes in the HiMean models may be beneficial or detrimental to diagnostic performance when compared to the static query dynamics of LSA. Additionally, even after controlling for the undercurrents of the response generation processes, model divergence in information representation itself may lead to differential outcomes despite equivalent inputs.

Base rate information. I also manipulate the composition of the models' semantic spaces by changing the frequency of the semantic vectors comprising the models' memories. This does not seem to be a conventionally explored aspect of semantic analysis. By varying the base rate information associated with various diseases' memory traces, I am able to generate memories tailored to reflect specific information environments or experience. For example, even given the same set of symptoms, we might expect different diagnoses from a doctor operating under

conditions where those symptoms are rarely occurring and highly diagnostic of associated diseases than from a doctor operating in an environment where those symptoms are ubiquitous. Similarly, an expert doctor with much more experience with a particular set of diseases is likely to respond differently to a set of symptoms than a novice. Base rates of categories of disease can be manipulated such that the prevalence of disease categories (e.g., psychological disorders versus digestive disorders) may lead to differential diagnostic performance or probe sensitivities.

As previously mentioned, HiMean output is sensitive to changes in base rate information and this is expected to hold despite the specific contents of the memory traces that HiMean is operating over. Despite the importance of base rate information to psychologically-plausible cognitive models generally, and to HiMean in particular however, it has not been shown whether an LSA approach to semantic retrieval is influenced by base rate changes to the same degree, or even at all. Thus, the discovery of any differences in diagnostic performance yielded by the manipulation of base rate information would represent an important contribution to this area of research by providing an opportunity to examine the influences of base rate information on hypothesis generation and testing processes.

Probe diagnosticity and error. Another domain for examination regards the influences of the characteristics of the probes/cues themselves on decision outcomes. The memory probe ("D_{obs}" to HiMean, "query" to LSA) can be varied according to its diagnosticity within the semantic space and the amount of error it contains. Probes with high diagnosticity should be expected to lead to better diagnostic performance, while more ambiguous probes are expected to lead to relatively degraded performance. It

therefore becomes possible to use the manipulation of probe diagnosticity to evaluate the robustness of the models subject to these differences and again we may see inconsistencies in performance between variants of the HiMean model and the more traditional LSA. Relatedly, increasing probe error may have similar deleterious effects on performance. Probe error refers to the quality (or fidelity) of a probe with respect to the pristine form of that probe and/or the corresponding traces in memory, with more errorful (noisy) probes potentially leading to more errorful retrieval (with respect to the retrievals elicited by the error-free version of the probe). This manipulation serves to evaluate the various models' sensitivity to perturbations of probe information and their effect on outcome measures. It would be important to learn if psychologically derived semantic spaces demonstrate a lower sensitivity to such perturbations as compared to standard computational semantics (or vice versa) and where and why these differences might exist. Therefore I manipulate probe diagnosticity and error to allow for their influences in diagnostic decision-making to be explored in depth.

Model Performance

Relative choice. I evaluate the influence of these manipulations on model performance according to a number of measures. The first is relative choice in a diagnosis task. Given the input of a probe, how will a model respond? This measure assesses the models according to their ability to generate the correct hypotheses in response to a probe. By having knowledge of the actual disease from which a symptom probe is extracted, the ideally appropriate responses are known *a priori*, and the degree to which a model's output concords with those responses is a metric for the optimality

of the model's decision processes. Cosine distance serves as a convenient measure of model performance in terms of evaluating diagnostic choice. In LSA, the vector with the highest cosine distance with respect to a probe can be seen as having the greatest semantic similarity to the probe. In the HiMean model, the retrieved memory trace (subject to the constraints of the particular instantiation of HiMean which produced it) with the highest activation strength in response to the probe is the diagnosis. Trace-probe cosine distance is also employed as a similarity measure in the HyGene models. Using this metric, it becomes possible to compare the performance of these models in terms of their diagnostic capabilities.

Consideration set and probability judgments. I also measure model performance by the set of alternative considerations they generate. While there is certainly something to be learned from the models' primary response to a probe, there is also important information to be gleaned from the entire set of likely responses, especially from a psychological perspective. By examining the top few responses to a probe, the models' relative performances over a larger set of circumstances can be determined. For example, while the top choice generated by LSA may have a greater cosine similarity than the top choice generated by a HiMean model, it might be that LSA's second and third options are relatively poor responses to a probe in comparison to HiMean's second and third choices. This would demonstrate that LSA may actually be a poorer option from a decision support standpoint because having viable alternatives to the primary choice is important under this framework. Another possibility is that the constrained version of HiMean may fail to generate alternatives altogether and this cannot be appreciated without considering the full set of hypotheses. Finally, by taking

into account the entire consideration sets generated by the models, the posterior probability judgments for each of the items in the sets can be calculated. The judgment as to the probability of a given response to a probe is contingent upon the strength and availability of alternatives in the comparison set. Again, from a decision support standpoint, it is important to understand whether the primary hypothesis offered by a model is considered nearly as probable as the alternatives in its set or if the alternatives have extremely low probabilities with respect to the focal hypothesis.

Semantic space evaluation. Beyond the properties of the model outputs, I also aim to investigate the qualities of the semantic spaces themselves. Using cosine similarities, it is possible to explore how semantic clusters are arranged in the different representations (e.g., dense vs sparse clusters, total number of clusters) and to make comparisons between distances associated with various words or semantic concepts. The process of removing all but the strongest dimensions during singular value decomposition in LSA intentionally decreases the distance between similar constructs while retaining the most informative features of the space, but in doing so may change the multidimensional structure and shape of the entire space differently than the pruning process used to cultivate the semantic memory in the HiMean model. Conversely, the multidimensional space of a global match memory model and LSA vector representation may share a great deal of features despite being constructed in dissimilar ways.

Method

Materials

In order to assess the respective decision-making capabilities of HiMean and LSA, I deployed them in a medical diagnosis task. An online medical information database consisting of 514 web documents was used as the source for both the LSA corpora and HiMean's constructed semantic memory. Each web document corresponded to a different disease, condition, or ailment. The words (rows) in these contexts were largely comprised of disease definitions, symptoms, causes, affected biological systems and structures, treatments, and diagnoses associated with the different disease documents (columns). The average number of words in each document was 495.98 (*SD* = 77.98). All experiments were conducted on a PC utilizing an Intel Core i5-4670K 4.2 GHz (3.4 GHz overclocked) processor with 32 GB of RAM and running a 64-bit version of Windows 8.1 Pro. The models were programmed using Wolfram Mathematica 9.0.1 64-bit and analyses were conducted using a combination of Mathematica and R (version 3.0.3 x64).

Design

Three models were constructed for comparison. The first model was a standard implementation of LSA. The second was a basic version of HiMean that incorporated a customized semantic memory but which still adhered to human-like psychological capabilities. The third model was an "ideal observer" version of HiMean which, though based on the same semantic memory as in the second model, was not subject to the

same “psychological” constraints and processes as the common HiMean model. The construction of these models is detailed in the procedures section. Each of the models was compared on their performance of a diagnosis task under varying probe conditions across varying base rate and disease cluster conditions, with relative choice, consideration set, and probability judgment as the dependent measures. Semantic spaces are also examined.

Procedure

Preprocessing. Each document was pre-processed by removing all punctuation, special characters, and numbers, and by converting all text to lowercase characters. A list of all words appearing across all documents was compiled. From this word list, all words with less than two letters were also removed. This was done because the structure of the medical text included many roman numerals and abbreviations (e.g., cc, mm, iv, im, etc.) which do not contribute much information to the semantic space. This basic word list was then further processed by different techniques in order to make it suitable to the model representation it was deployed in.

In order to allow for base rate manipulations, each disease document was classified according to its location within the Medical Subject Headings (MeSH) information structure found on the National Library of Medicine, National Institutes of Health web site (http://www.nlm.nih.gov/mesh/). The MeSH index is a hierarchical description of medical vocabulary and can be used to illustrate the conceptual relationships between diseases. Classification according to this structure resulted in each of the diseases being categorized into a total of 619 concepts (some diseases

belong to multiple categories) subsumed under 27 major disease groups. The major disease groups largely corresponded to physiological systems (e.g., musculoskeletal diseases, respiratory tract diseases) and disease etiologies (e.g., chemically-induced diseases, parasitic diseases, virus diseases). These categories were further grouped together according to cluster analysis on the number of diseases in each category. A full listing of the major disease categories, the number of diseases in each category, and cluster assignment can be found in Appendix A.

LSA processing. The basic word list of all words consisting of more than two letters as described above was used. The total number of words used in the LSA model analysis was 9,595 resulting in a 9,595 (term) x *N* (document) matrix, where *N* was dependent on the base rate manipulation of the disease document frequency. Once the word counts in each document were tabulated, the vector elements of the LSA term x document matrix were subjected to the same log transform function and entropy weighting technique as used in Landauer and Dumais (1999) (described previously). Singular value decomposition was then performed on the entropy weighted matrix. The number of dimensions to retain was chosen such that approximately 80% of the variance accounted for by the full dimensional matrix was intact. This method generally indicated an optimal dimensionality between 250 and 300. Figure 3 depicts the average within cluster similarity of the different disease groups for the unmodified base rate matrix (i.e., all disease documents appeared only once in the corpora) as a function of decreasing number of dimensions retained.

After the SVD, a new matrix corresponding to the retained dimensions was constructed and used for analysis. The cosine distance between vectors was calculated as the similarity metric used for judging model performance.

HiMean processing. Constructing the episodic memory in HiMean, though similar in purpose to the term x document matrix used in LSA, required techniques that were different from those used in the LSA model. In order to create the memory traces, I followed the steps described by Kwantes (2005) in his discourse on the semantics model. The same 9,595 word list as used for the LSA corpora was used as a starting point. This list was further trimmed to remove words appearing in more than 90% of the 514 disease documents. Because they occur in almost every document ("Promiscuity"; Kwates, 2005), these words convey little meaning about the individual documents. Similarly, words that appeared multiple times but only in the same document, and words that appeared less than two times across all documents ("Monogamy" and "Celibacy", respectively; Kwantes, 2005) were also removed because some overlap of contextual information is necessary to understand a word's meaning. The final result was a word list containing 5,767 words.

The number of times each word appeared in each document was calculated and used to form the environmental context vector for each disease document. Kwantes (2005) used the same logarithmic transform as in LSA to adjust the vector elements (though he did not use entropy weighting), however, for HiMean no transform or weighting was applied and the original word counts themselves were used as vector elements. These vectors represented the information structure of the external world (i.e., the information environment the model operated in). Each context vector was encoded

as an imperfect trace into the model's memory according to a learning parameter *L*, where $L \in [0,1]$. Elements (word counts) in the context vector were randomly replaced with zeroes with a probability equal to 1-*L* before being recorded into episodic memory. Because the vectors were each 5,767 elements long, *L* was set to 0.99. This episodic memory served as the foundation for the retrieval dynamics in HiMean. In accordance with HiMean operating principles, probe/query (D_{obs}) items were matched against episodic memory and those traces responding with the highest activation to the probe were then compared to traces in semantic memory in order to identify the best traces. The activation calculations and model semantic memories differed between the ideal observer and common version of HiMean.

Ideal Observer HiMean. In the constructed semantics model (Kwantes, 2005), the similarity (or resonance) between a probe item and a memory trace was calculated as the cosine between the two vectors, $Similarity = \frac{\sum_{i=1}^{N} Probe_i \times Trace_i}{\sqrt{\sum_{i=1}^{N} Probe_i^2} \times \sqrt{\sum_{i=1}^{N} Trace_i^2}}$, where *i* represents the vector elements. As discussed briefly earlier, this departs from the MINERVA 2 similarity calculation ($S = \sum_{i=1}^{N} \frac{Probe_i \times Trace_i}{N}$, where *N* is the number of element pairs not equal to zero; Hintzman, 1984) because the vector representations of MINERVA 2 and Kwantes' (2005) semantics model are different. As HiMean' trace structure is the same as that used by Kwantes (2005), I used a nearly identical angular similarity calculation,

$$S = 1 - \frac{2\cos^{-1}\left(\frac{\sum_{i=1}^{N} Probe_i \times Trace_i}{\sqrt{\sum_{i=1}^{N} Probe_i^2} \times \sqrt{\sum_{i=1}^{N} Trace_i^2}}\right)}{\pi}$$

because the vector coefficients are always positive and this creates a normalized similarity metric bounded between [0,1].

Typically, this similarity is then used to compute a memory trace activation strength. In MINERVA 2, this is simply the cube of the similarity, $A = S^3$, which serves to increase the separation between relatively good and relatively poor matching vectors, thereby allowing the better matching traces responding to the probe to dominate the system (Hintzman, 1984). Kwantes (2005), did not cube the similarity, instead setting the activation equal to the similarity ($A = S$) and opting to follow the example of MINERVA -DM (Dougherty, Gettys, & Ogden) by imposing a minimal threshold of activation which trace activation must exceed in order for those traces to contribute to the model output. IO HiMean uses a combination of these techniques by both cubing the calculated similarity to represent trace activation and utilizing a threshold of activation (A_c) which traces must exceed. Kwantes (2005) set the activation threshold to 0.1 and chose to implement this threshold both due to computational considerations and because the large number of traces (>86,000) involved meant that even exponential weighting of the similarity was unlikely to remove enough noise from the output to ensure coherency. In IO HiMean, rather than selecting a convenient cutoff, the A_c for responses to each probe is computed dynamically over a parameter space [0,1] such that the chosen A_c minimizes the ratio of false responses to correct responses. Optimal A_c is given by $Min\left[\frac{\frac{N_{\neg Dobs}}{N}}{1-\frac{N_{\neg Dobs}}{N}} \times \frac{N_{>Ac} - N_{Dobs>Ac}}{N_{Dobs>Ac}}\right]$ where N is the total number of traces in memory, $N_{\neg Dobs}$ is the number of traces that don't correspond to the probe, $N_{>Ac}$ is the number of traces whose activation exceeds A_c, and $N_{Dobs>Ac}$ is number of traces correctly corresponding to the probe whose activations exceed A_c (Thomas et al., 2008). This parameter optimization is possible because the true identity of the probe item is known.

Note that humans do not have access to this pristine knowledge when searching memory under real-world circumstances.

In both Kwantes (2005) and IO HiMean, elements of traces whose activations exceed A_c are then weighted by their activation strengths and summed to form a composite (unspecified) probe representative of their semantic meaning. In this way, a semantic trace vector can be comprised of both relevant and irrelevant episodic traces as long as the constituent traces have an activation level greater than A_c. The ideal observer HiMean model attempts to mitigate the influence of the irrelevant traces by using an adaptive threshold set to minimize the number of false alarms contributing to the makeup of the unspecified probe. In IO HiMean, this unspecified probe is then matched against semantic memory, returning the semantic traces most closely resembling the probe. In order to create a "perfect" semantic memory in IO HiMean, the unspecified probe that was a composite of only episodic traces actually belonging to D_{obs} was stored as the semantic memory trace for that D_{obs}. Despite the number of traces comprising episodic memory (which is dependent upon the base rate of the disease documents), only one composite trace for each possible disease was recorded into semantic memory. Therefore, while episodic memory may contain many thousands of traces, semantic memory only contained 514 (one for each unique context in this dataset). With the semantic memory thus constructed, the full machinery of IO HiMean could be queried using various probe items (D_{obs}). The cosine distance between unspecified probe vectors and semantic memory traces was calculated as the similarity metric used for judging model performance. Results were computed across the entire semantic memory space,

returning a cosine similarity for every probe-trace pair. The best matching pairs were considered to be the model's best choice(s).

Standard HiMean. The basic operation of the standard version of the HiMean model is the same as in the IO model with a few exceptions. The word list used was the same. Similarity and activation calculations were accomplished in the same manner as well. The activation threshold (A_c) was not set optimally as in the IO model, however. Instead, A_c was set to .04, the average threshold derived by the optimized approach. This enabled unspecified probes extracted from episodic memory to be comprised of more irrelevant memory traces than in the IO model. Similarly, semantic memory was not comprised only of pristine traces in the standard model. Semantic memory in the standard model consisted of L parameter (here L = .99) degraded versions of the original environmental (context) vectors corresponding to each disease document. This semantic memory base did not change according the base rate manipulations in episodic memory and was meant to represent the gist information extracted from the information environment (Reyna & Brainerd, 1991). Moreover, the standard HiMean model put the retrieval processes and working memory constraints of HyGene to work. While the IO model generated similarity responses over the entire semantic memory and chose the best possible candidate, the standard model sampled from memory stochastically, where retrieval attempt failures and capacity limitations could lead to suboptimal outcomes. In order for a memory trace to be initially considered as a likely candidate response to a probe, its activation must exceed the minimal activation threshold upon being sampled, subsequent attempts at generation must have activations that exceed the activation of the lowest threshold item in the SOC. Additionally, too many failed attempts at

retrieving a sufficiently activated memory item resulted in the model's failure to generate responses to its full capability. For these simulations the working memory capacity of the model was set to four and the maximum number of retrieval failures (T_{max}) was set to three.

Base rate manipulations. Disease context base rate manipulations were accomplished by changing the number of documents in the corpora corresponding to those diseases. For example, to increase the base rate of skin diseases to five, each disease document classified as a skin disease would be written to the corpora five times. All diseases not classified as a skin disease would be represented by only one document in the corpora. Because some diseases could belong to multiple categories (e.g., Pneumonia can be a bacterial disease but is also a respiratory tract disease), sometimes a disjoint in the requested base rate arose. In cases where the base rate manipulation resulted in such a discrepancy (e.g., the base rate of bacterial diseases was set to five, but the base rate of respiratory diseases was set to one), the disease at issue was recorded into the corpora at the highest requested base rate. The additional documents were added according to their base rates to the LSA corpora before the log transform entropy weighting and SVD. In the HiMean models, additional memory traces corresponding to the disease documents were added to episodic memory before memory activation calculations took place. Note here that while the documents added to the LSA model corpora were perfect copies, in the HiMean models they underwent the same encoding degradation that all other episodic memory traces were subjected to. That is, the additional traces were imperfect copies of each other and only similar to each other within the degree of probability set by the encoding parameter (L).

Probe quality manipulations. Disease context vectors were used as the probe items for the LSA and HiMean models. In order to manipulate to the quality (amount of error) of these probes, each of the probe's non-zero original vector elements had a chance to be replaced by a zero with a probability established by a probe noise parameter set between 0 and 1. For example, if the noise parameter was set to 0.1, each non-zero probe vector element had a one in ten chance of being replaced by a zero. Every disease vector was exposed to this degradation process and each was used as a probe to query the models. I originally considered assigning an equal probability of replacing each vector element with the same element from another randomly chosen disease vector, but opted for the current method for two reasons: 1) the sparseness of each vector made it so that simply replacing an element with some probability led almost inevitably to an already zero element being replaced with a zero from another vector and thereby failing to accomplish the goal of the manipulation (i.e., degrading the probe) and 2) the current approach allows for more conclusive findings vis-à-vis the robustness of the models to degradation because it is possible to state exactly which probe was used with more certainty than if its elements had been comprised of the elements of other vectors (which could result in a vector resembling the originally intended vector in name only). The degradation was applied to the unmodified context vectors (i.e., before activation and similarity had been calculated) in the IO HiMean and standard models.

Because the LSA and HiMean models used a differing number of words, their respective disease context vectors were of different lengths. HiMean vectors contain only a subset of the context vector elements used in the LSA model. In order to make

the noisy probe items equivalent across models, a query translator was used to interpret the output from the HiMean noisy probe generator and convert it into an equivalent probe formatted for use in the LSA model. LSA vectors were first subjected to the same degradation process as in the HiMean models to ensure the extra elements in the longer LSA vectors were affected with the same probability. Then, the query translator worked by replacing the LSA context vector elements corresponding to the same words in the noisy HiMean probe context vector (since all words in the HiMean context vectors could also be found in the larger word set constituting the LSA context vectors) with the appropriate elements defined in the HiMean noisy probe.

In the cases using the HiMean models, the resultant noisy probes were then directly used as queries because their representations were immediately amenable to the format necessary to operate within the models' memory structures. In the case of LSA, the translated noisy probes had to be further translated into "concept space" before they could be used as queries (Rajaraman & Ullman, 2011). This process involved multiplying the noisy probe vector by the right singular vector matrix (i.e., the non-transposed V matrix) that resulted from the SVD (Rajaraman & Ullman, 2011). This concept vector was then translated back into "disease space" by multiplying the vector by the conjugate transpose of V and used to query the LSA model. Note that this process mitigates to some extent the effectiveness of the probe degradation because the mathematics force some of the previously degraded vector elements (i.e., zeroes) to take on approximate values according to the particular vectorspace they're operating within. This may be considered an advantage of LSA with respect to performance in this

particular domain. Normalized angular cosine distances between query vectors and reconstructed LSA vectors were again used as measures of similarity.

Diagnostic capability evaluation. The diagnostic capability of each model was tested by constructing customized query vectors that contained information specifically pertaining to user-selected symptoms and features. When this portion of the evaluation was initiated, an open text box appeared which prompted the user to "Describe the symptoms". The user could type anything in this box and a query containing the elements of the input text was generated. For these experiments, portions of the "signs and symptoms" sections of Wikipedia articles dedicated to ten diseases in total were chosen from the corpora disease list and used as input to the text boxes. One disease was chosen at random from two different categories for each of the large disease clusters (clusters 1, 2, and 3), and one disease was chosen from each of the two remaining clusters (

Appendix C: Disease Clusters). Care was taken to make sure the input text did not contain the name of the disease itself, though input text may involve a portion of the disease name. For example, a query about Rheumatic fever, would not contain the word "rheumatic", but may contain the word "fever" in its description of symptoms.

A complete, full-length query vector was always generated. Unlike the probe quality manipulations, a translator was not used to convert HiMean probes into the longer LSA probes. Instead, the same text was entered in the text boxes for both model types. This allowed for the LSA model to capitalize on the involvement of additional words in the text box that might not be represented in the shorter HiMean vectors, though the text entered into the boxes was the same for both model types. Any elements of the query vector that corresponded to words the user did not input were left at zero. Any words the user input that were not part of the word lists used to generate the context vectors for each model were not used in the generation of the query vector. Words that the user did input and that were also a part of the word list had their corresponding vector elements set to one. The output of the query generators was thus a context vector of 1s and 0s corresponding to words that were either present or absent from the user's input, respectively. The HiMean models used these query vectors as probes directly, while the LSA model used the VV^T matrix multiplication process described in the previous section to translate the query into LSA disease space first. The best cosine matches between the query vectors and the other context vectors or memory traces were used as a proxy to indicate the models' disease diagnosing capabilities given the input text information.

Consideration set and probability judgments determination. In the LSA model, the five diseases with context vectors having the greatest cosine similarities to the probe (excluding the probe disease context vector itself) comprised the model's consideration set. The probability of the selection of any alternative from among that set was computed as the cosine similarity of that alternative divided by the total sum of all the cosine similarities of all items in the consideration set. In the IO HiMean model, consideration set was defined as the five diseases with context vectors having the highest similarities to the probe item and the probability of selection for any given item was calculated as the activation of any item divided by the sum of the activations of all other memory items. In the standard HiMean model, the number of diseases in the consideration set was defined by the number of diseases generated into the SOC, and the probability of any of those alternatives as the best explanation for the D_{obs} was calculated as the activation of that item divided by the sum of the activations of the other items in memory.

Semantic space construction. Non-metric multidimensional scaling was used to generate two-dimensional representations of the LSA and HiMean models semantic spaces under different manipulations of base rate information. The multidimensional scaling was performed using the isoMDS function as part of the MASS package in R for the HiMean models' semantic memories and for LSA. The calculations were performed over the symmetric cosine distance matrices generated from all pairwise disease context vector comparisons in semantic memory or in the term-by-document matrix.

Results

The results section is divided roughly into four sections. Each is dedicated to formally reporting the outcomes of the procedures set forth in the previous section. The reporting of outcomes and probability judgments is spread between the base rate, probe quality, and diagnostic capabilities information, however as it was not possible to talk about them removed from the context of the other investigations.

Base rate results

To fully explain the impact of the base rate manipulations in these experiments, it is necessary to first examine their impact on the model semantic spaces themselves before discussing how they come to bear on the model output.

LSA base rate effect. The effect of manipulating the number of times the same document appeared throughout the corpus was examined. Using the LSA model, I found that increasing the number of documents belonging to particular disease categories resulted in a higher average cosine similarity between diseases belonging to those categories relative to the average similarity found in disease clusters corresponding to documents that appeared only once. Figure 3 depicts the average within disease cluster cosine dissimilarity of the disease groups as dimensionality was reduced in the LSA framework when all diseases had a base rate of one (i.e., each disease document appeared only once). Figure 4 shows the same graph but where diseases classified as being cardiovascular diseases had their base rate increased to five. That is, all cardiovascular diseases contexts appeared five times in the corpora. In this

graphic it is plain to see that cardiovascular diseases were much more similar to each other than other disease clusters were to themselves, even at the full dimensional space. Similarity between cardiovascular diseases also increased more rapidly as dimensions were discarded, eventually collapsing into near identicalness with a higher number of dimensions retained than did the other disease clusters. When the base rate for cardiovascular disease was increased to ten, the within cluster similarity again increased with respect to the unmodified disease cluster base rates, and the collapse in the similarity structure occurred at an even higher dimensional space (Figure 5). For clarity, Figure 6 and Figure 7 depict these same manipulations, but the plots only contain those disease groups from the same size cluster as the cardiovascular disease group (i.e., the cluster whose group members contained 37-50 diseases each).

In order to demonstrate that this effect was not singular to the diseases contained in the cardiovascular disease group or due to the number of diseases in the cardiovascular size cluster, the same base rate manipulations were performed for diseases classified as psychological diseases. Figure 8 and Figure 10 show a psychological disease base rate of five and Figure 9 and Figure 11 show a base rate manipulation of ten for psychological diseases. Although the effect of reducing the within group disease dissimilarity is not as pronounced for the psychological diseases in the base rate five condition (Figure 8) as it was for the same base rate in the cardiovascular disease condition (Figure 4), the effect is quite clear when the base rate for psychological diseases was increased to ten (Figure 9). Here again, the ability to distinguish between diseases in the psychological disease grouping becomes difficult

due to their increased similarity even with approximately $1/5^{th}$ (100) of the available dimensions retained (Figure 9, Figure 11).

The cosine distance between every possible disease context vector pair was calculated and then the distance between each disease of one group and each disease of every other group in the same size category was averaged to give the average between cluster distances for each disease group in the same size category. For example, the cardiovascular disease group belongs to the 37-50 size category (Cluster 3) along with the digestive, male urogenital, neoplasms, and respiratory disease groups. The cosine distance between each disease in each of those groups was calculated and averaged according to group membership. Figure 12 and Figure 13 show the average between cluster dissimilarities computed for the size categories containing the manipulated psychological and cardiovascular disease groups according to base rate manipulation and with 300 dimensions retained.

To investigate the effects of simultaneously changing the base rates of multiple disease categories, the base rates of cardiovascular and psychological diseases were both set to five and then ten. Figure 14 and Figure 15 show the impact of these manipulations on the cosine distance relationships across diseases as a function of dimensionality reduction. These figures demonstrate that the within cluster similarity of the manipulated disease groups increases relative to the un-manipulated groups and with increases in base rate.

HiMean base rate effect. The IO HiMean model used a perfect semantic memory consisting of composite memory traces made of only and all memory traces

correctly associated with the probe. This feature of the model made the semantic memory relatively static despite episodic memory trace base rate manipulations. That is, the semantic memory was designed to feature memory traces that were the ideal prototype representing all relevant episodic trace exemplars and thus was essentially immune to fluctuations in the quantity of exemplars introduced by disease context vector base rate changes. Note that context vectors were encoded into episodic memory with the encoding parameter, *L*, set to .99. This made it so that the episodic memory traces comprising the semantic traces were 99% faithful with respect to the original context vectors and introduced little variability into the IO HiMean memory system. Furthermore, although there was little variation between semantic memories constructed from varying episodic memory trace base rates, the process of creating a composite trace from relevant activated traces does lend a slight advantage to those traces that were recorded into episodic memory more than once because the composite of multiple traces would result in a higher fidelity prototype with respect to the original disease context vector and make it more robust to changes introduced by *L* that singly recorded traces have no way of overcoming.

Average between group cosine dissimilarity for semantic memory traces constructed from an episodic memory where all diseases documents were recorded with a base rate of five ($L = .99$) demonstrates that, while not quite as strongly associated with each other as occurred in the LSA model, similarity is greater for diseases belonging to the same category than for diseases belonging to different categories (Figure 16, Figure 17, Figure 18, Figure 19, Figure 20). Keep in mind that IO HiMean simulations compute a new semantic memory based on the individual characteristics of

the episodic memory for each run, so while each semantic memory is nearly the same, they are not identical. These figures therefore represent an approximation of the relationships between the IO HiMean semantic memory traces for any given simulation run. In the standard HiMean model, recall that the semantic memory is produced once using an *L*-degraded episodic memory with a base rate of one for all diseases and this same semantic memory is used for all simulations, rather than being computed differently for each episodic memory base rate configuration. The between and within cosine similarity measures generated by the two models were essentially identical ($r(298) = .999$, $p < .0001$) and thus the graphs are not duplicated for the standard HiMean model.

Base rate manipulations on model output. Each model was probed with all 514 of the possible context vectors and output the most likely candidates given the probe for each memory system (where applicable, the cosine similarity of each response to the probe, and the probabilities associated with each response. Under the control base rate condition (i.e., all diseases set to have a base rate of one), the IO HiMean model correctly identified the probe in semantic memory on every trial, the standard HiMean model without a working memory capacity limitation in place also correctly identified the probe in semantic memory on every trial, and the standard HiMean model employing the working memory construct correctly generated the probe as the most probable item in the SOC on 90% of trials (Table 1). The LSA model also returned the correct item vector as the highest match in every case. This is not surprising however, given that in the LSA model, the document vectors belonging to the correct return were always identical to the probe. The average normalized cosine matches of the LSA

model's best predictions are also higher (M = .77, SD = .16) than in the other models (IO HiMean: M = .44, SD = .29, t(5138) = 51.19, p < .001, Standard HiMean: M = .44, SD = .28, t(5138) = 51.56, p < .001) though the disease category membership predictions of the LSA model (M = .56, SD = .26) are less accurate than the others (IO HiMean: M = .79, SD = .25, t(1026) = 14.73, p < .001, Standard HiMean: M = .79, SD = .25, t(1026) = 14.89, p < .001).

The base rates of diseases belonging to the cardiovascular and psychological categories were manipulated in order to examine base rate effects. Table 2 shows the same information as in Table 1, but for model output derived from memories where the base rate for cardiovascular diseases was set to five rather than all disease context vectors being written to memory only once. As can be seen from a comparison of these tables, the models performed in nearly the same way with regard to overall accuracy and average similarities. When examining the category membership of model selection however, LSA clearly demonstrates selection biased in favor of cardiovascular diseases. That is, the proportion of its highest ranked query vector responses belonging to the cardiovascular diseases category increased from .069 to .19 overall while, for instance, its selection of candidate vectors from the psychological category did not change (.014 to .013). Conversely, within the HiMean models, the proportion of cardiovascular diseases selected remained relatively unchanged despite the increase in cardiovascular disease base rates contributing to the models' memory systems. Table 3 shows a breakdown of proportionate disease selection for all the base rate manipulations across models. Under the conditions where the base rates of diseases categorized as psychological were manipulated, the same pattern of findings was revealed where LSA

displayed an increase in category congruent response selection relative to the manipulation while the other models did not demonstrate this same sensitivity (Table 3). Where both cardiovascular and psychological disease category base rates were manipulated at the same time, LSA again demonstrated correlated selection, though the influence of cardiovascular disease manipulations seemed to have a larger impact on its behavior (reference also Figure 14 and Figure 15).

Base rate manipulations on probability judgments. To assess the accuracy of the probability judgments rendered by the models, Brier scores were calculated for each model across base rate conditions. A Brier score demonstrates how well calibrated a prediction system is with respect to the probabilities assigned by the system to particular outcomes, as well as the actual outcomes of the predicted events (Brier, 1950). The formula for this calculation is $BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(p_{ti} - o_{ti})^2$, where p represents the probability assigned to each of the forecasts for an inquiry/probe (here, usually the top five strongest associates to the probe, making $R = 5$, or the number of items in the SOC) and o is the outcome for each prediction (1 if the probe was correctly identified, 0 otherwise). N is the total number of events predicted by the system, here 514, the total number of disease probes used to query the models. The higher a probability assigned to a correct prediction, the lower the mean squared difference between them, and the better the calibration of the system. Thus, a lower Brier score indicates a more accurate or better calibrated system. Table 4 shows the Brier scores and average probability of top choices for each model's performance across all disease probes in the control base rate condition (i.e., base rate = 1 for all diseases).

Table 5 displays the average of the probabilities generated by each model for only those probes categorized as either a cardiovascular or psychological disease. While the average probabilities output by the first three models (LSA, IO HiMean, and Standard HiMean without a working memory limitation) are relatively constrained due to both their underlying retrieval mechanisms and by being required to generate no more or less than five candidates in response to each probe, the standard HiMean model does not share this restriction. The standard model demonstrated a slight increase in average rendered probability judgments for those hypotheses generated in response to probes that were congruent with the increased base rate manipulation (Table 5). In the cardiovascular base rate 10 condition, for example, the average probability judgment for diseases generated in response to cardiovascular probes was higher than the same average in the base rate 5 condition which was, in turn, higher than in in the control condition.

The Brier scores for the models across base rate conditions can be viewed in Table 6 and Table 7. The general pattern of findings indicates that, via a tendency toward lower Brier scores, HiMean models produce output that is better calibrated to the frequency distribution of the diseases within the corpora than LSA. With few exceptions, the IO HiMean model produced the best Brier scores. Indeed, the model seemed to be so well calibrated that its scores were practically invariant with respect to the base rate manipulations. It may therefore be exhibiting a floor effect given the nature of its operating characteristics. LSA produced Brier scores that actually increased (i.e., performed worse) with increases in base rate, while the other models generated mixed results with a tendency toward decreased (i.e., better calibrated) scores in

response to increased base rates. A further trend that can clearly be seen is an increase in the accuracy of probability judgments rendered by the semantic memory systems of models in comparison to the episodic probability judgments made by those same models. It should also be noted that the Standard HiMean model with the SOC in place demonstrated the most calibrated behavior under at least some of the conditions. Because this model is not forced to consider suboptimal alternatives (as the other models may), it is capable of extremely high performance under the right conditions. For example, it is the only model that can ever state with 100 percent certainty that it believes a single candidate hypothesis to be true. Unfortunately, the random aspect of its retrieval mechanisms can also lead to worse performance under some conditions. Taken together, these characteristics contributed to a mediocre performance overall, though one that still performed excellently, especially in comparison to LSA on this particular metric.

Probe quality results

Only the models' performances in response to perfect probes have been investigated to this point. Experiments were also conducted to investigate the impact of degraded probe information on model output. The proportion of trials in which the models' top choice was correct is shown in Table 8. The HiMean models' performance in the semantic choice category was basically immune to the degradation of the probes. This was only true to a point, however. Once probe quality was degraded sufficiently (~0.6 of the non-zero feature elements for each vector were replaced by zeroes), the models failed to perform at all. In contrast, LSA was able to continue to operate, displaying a graceful degradation in correct choice as probe quality decreased. Episodic

memory choices for IO HiMean demonstrated a similar degradation in performance to LSA that was not demonstrated in the semantic condition, but again was unable to perform once the fidelity of the probes had been sufficiently compromised. The standard implementations of HiMean produced more mixed results in episodic memory for this task. HiMean without working memory actually improved at correctly identifying the probe as probe quality decreased, and the basic model performed at the same level across probe quality changes (Table 8).

The model's performance as determined by the presence of the correct choice among the each model's top *N* alternatives is shown in Table 9. Unsurprisingly, LSA model performance improved, though it still declined in tandem with probe integrity. Considering a larger alternative set likewise increased HiMean model performance, but the models were still unable to produce output once probe fidelity reached a lower threshold of approximately 0.6. Table 10 contains the Brier scores for the models as a function of changing probe quality. Semantic memory probability judgments are seemingly stable for LSA across quality conditions and even with respect to the scores attained in the other experiments. This finding is an artefact of the similarly nearly invariant probability judgments rendered by LSA. In the case of IO HiMean semantic probability judgments, however and as was seen with its episodic choices, performance declines, albeit minutely, as probe reliability is reduced. The same behavior seems to be demonstrated by HiMean without working memory, but the random element to the standard model with WM intact makes the same conclusion difficult to reach. As concerns the episodic probability judgments, All HiMean models are more poorly calibrated than in semantic memory, but Standard HiMean without WM paradoxically

improves with probe feature decline, while IO HiMean demonstrates the expected trend. Overall, all models still perform very well, and the LSA model's capabilities are edged out by HiMean.

Diagnostic capability results

Each model's responses to the ten selected Wikipedia disease signs and symptoms descriptions were recorded. Performance was assessed as the proportion of correctly diagnosed diseases and the probabilities associated with the top alternatives generated by each model were also documented and brier scores calculated. Note that the Standard HiMean model still had its threshold of activation set to be equal to the average of thresholds generated by the adaptive IO HiMean model. This resulted in the activation threshold for the Standard HiMean model being lowered from .04 by an order of magnitude to .003. Table 11 displays each model's performance. The models compared were LSA, IO HiMean, and Standard HiMean without working memory. The working memory model was excluded from this analysis in the interest of clean data presentation and as the Standard model without working memory essentially represents the same output with the random sampling removed.

The results reveal that the models performed at a relatively even level, bot with respect to each other and to their past performance, with the exception of a considerable reduction in their ability to choose the top choice correctly. The LSA and IO model both performed well in terms of at least generating the correct choice among the top five alternatives. Even the Standard model performed above chance and even when making episodic evaluations. LSA's calibration is to be expected given the manner in which it

generates probability judgments. The IO HiMean model again demonstrated the best calibration, and the standard model, while not as high-performing, did not do worse than LSA. As can also be seen in Table 11, while the best models had the correct choice among the top alternatives on 80% of the trials, nearly half of their incorrect hypotheses were at least in the correct disease category. Even the standard model performed slightly above chance by this measure. The standalone diagnostic capabilities of these models do seem to be operating correctly, albeit at an understandably diminished capacity with respect to the corpora they were trained on.

Semantic space results

Two dimensional graphs of the cosine distance matrices generated from the SVD reconstructed LSA log transformed and entropy weighted context matrices at varying levels of dimension reduction are shown in Figure 21, Figure 22, and Figure 23. Multidimensional scaling was used to reduce the semantic spaces to two dimensions for the purpose of graphing. Because these diagrams involve such a large number of plot points and disease categories, the labels for most of the diseases have been removed and the points selectively colored to indicate category membership. Red text describes cardiovascular diseases and blue text labels psychological diseases. Green text represents virus diseases and brown text digestive diseases. The black dots are the locations of the remaining diseases.

Again, these graphs demonstrate that while the diseases are relatively spread out in semantic space under full dimensionality, they become more similar to diseases within their category as dimensions are reduced, until too many dimensions are

removed and all categories essentially collapse to the center of graphs as they retain little of their uniqueness. Psychological diseases are again shown to be the most resilient to dimensional reduction as their relatively far starting distance from each other demonstrates the dissimilarity between these semantically disparate disease concepts.

The IO HiMean and Standard HiMean semantic memory spaces for the control base rate condition are shown in Figure 24 and Figure 25, respectively. These two graphs demonstrate that starting semantic spaces are structured quite differently from the LSA model semantic space, but similar in many ways to each other. For example, within concept grouping is apparent in both the IO and Standard HiMean spaces. The psychological category diseases are also not as widely spread in the HiMean spaces as in the LSA spaces.

Figure 26 and Figure 27 show the LSA and IO HiMean graphs for the semantic spaces based on a cardiovascular disease base rate of 10. Remember that the Semantic memory in the Standard HiMean model did not change with base rate manipulations and therefore is not graphed again. Also note that the base rates for virus and digestive disease categories (green and brown text) were not manipulated but are shown to illustrate how their relationships may change as a result of manipulations in other disease's base rates. The next two figures (Figure 28 and Figure 29) show the same graphs but for the psychological disease base rate 10 condition, and Figure 30 and Figure 31 show the condition where the base rates of cardiovascular and psychological diseases were both set to 10.

The effects of the base rate manipulations on the LSA semantic spaces are quite clear from the figures. The same effects can be seen in the figures featuring the IO HiMean model, though in a less dramatic fashion. The cardiovascular disease grouping in Figure 27 is more condensed than in Figure 29, for example. It is also clear that the non-manipulated groups in the LSA model seem to retain their position irrespective of the manipulated groups. It is more difficult to see if the HiMean model exhibits this same characteristic.

Discussion

The purpose of this dissertation was to examine the operating characteristics of tested (LSA) and novel (HiMean) semantic models in relation to their performance in a medical decision-making context using a real-world information environment. It is clear from these experiments that while closely related, LSA and HiMean have quite different capabilities in the realm of decision-making stemming from semantic processing. Both types of model demonstrate utility. LSA appears to be quite sensitive to base rate information in terms of inter-item vector similarity, requiring far fewer dimensions to be dropped in order to recognize semantic similarity between items. This was carried out in a larger context of additional stable (i.e., base rate controlled) corpus information, however, and the effects would not exist in an environment where all base rates of all contexts were changed to the same degree. The base rate sensitivity of LSA is therefore contingent upon an information environment which allows the model to learn the relationships of dynamic information relative to static (or at least differently accelerating) information. Further, because the semantic relationship between different items is contingent on their cohesive covariation rather than on any “real” semantic

similarity in LSA, this also makes the system susceptible to false conclusions. For example, the disease category assignment could have been conducted at random and LSA would learn to strongly associate completely unrelated diseases as long as there was at least some overlap in their vector elements. This is perhaps only a concern in artificial environments however, as it is likely that in real-world operation, covariance does tend to indicate some association between items even if the linking variables are not always uncovered. Additionally, the susceptibility to base rate manipulations may be seen as a negative attribute in situations where very high discriminability between items is desirable, as the increase frequency leads LSA to infer increased similarity. This may be counterproductive to the purpose of LSA, however, which is to find latent similarity between items rather than preserve distinction. Indeed, the very act of SVD and dimension reduction is intended to reduce dissimilarity.

Perhaps you have heard the adage that everything looks like a nail to a person holding a hammer. Increasing the base rate for a particular group of items is like giving LSA a hammer. As can clearly be seen in Table 3, LSA tends to much more frequently posit guesses that are category congruent with the groups featuring increased base rates. Its overall performance in terms of correctly identifying the (pristine) probe disease, however, is not degraded due to this manipulation (Table 1 and Table 2) nor is its ability to generate probe-category relevant alternatives. The latter outcome seems likely to be due to the information environment structure where single diseases can belong to multiple categories, however. This would account for both the increased number of base rate manipulation relevant hypotheses posited and the paradoxically stable average probe-category relevant hypotheses proposed (Table 1 and Table 2). Alternatively, this

outcome could be explained by LSA always finding the same strongest competitors for each probe regardless of base rate and only after having included these same items does it add in the less-similar-to-the-probe base rate relevant postulates in order to reach its quota of five.

The IO and Standard HiMean models were also influenced by base rate information, though this influence was demonstrated somewhat differently than in LSA. Similar to LSA, the performance of the HiMean models in terms of their ability to correctly identify the pristine probe was not affected by base rate manipulations (Table 1 and Table 2). Contrary to LSA's behavior of increasing selection related to base rate manipulations, however, the HiMean models' selections seemed to operate independently of the base rates. In fact, essentially the only fluctuation in the performance of the models' selection behavior is seen in the case of the Standard HiMean model. The resultant changes are most likely best explained by that model's stochastic retrieval dynamics rather than attributing them to any changes in the information structure introduced by changing base rates. The performance of all the models in terms of selection was very good. Both LSA and the IO HiMean model performed perfectly in terms of correctly identifying the pristine probes, with the Standard HiMean model without working memory performing practically equivalently, and the working memory constrained Standard HiMean model lagging slightly behind (again most likely due to its retrieval characteristics).

It is also worth mentioning that the HiMean models were much better at offering alternatives that belonged to the same category as the probe across all probes and across base rate conditions (Table 1 and Table 2). This likely explains why these models'

selection behavior was invariant with respect to the base rate manipulations: they were doing a better job than LSA, in general, in identifying relevant alternatives. A more important conclusion can be drawn based on this finding , however, and that is while all the models were good at selecting the most correct answer, the quality of the alternative selections differed between models. Using only the proportion of probe-category congruent alternatives selected as a measure, the alternatives suggested by the HiMean models would at first glance seem to be more rational than those proposed by LSA, at least insofar as within category similarity between diseases serves as an indicator of a quality choice. To the extent that the non-category congruent alternatives chosen by the models are rational, however, this could indicate a particular model's bias toward diagnostic choices. In other words, if all the generated diseases were very similar, this could be viewed as a form of confirmatory (i.e., intracluster or exploitative) search, whereas the generation of highly semantically related, but category incongruent alternatives can be seen as a more diagnostic (intercluster or explorative) approach. In still other words, one might make a decision as to which model to use based on whether one was interested in identifying several closely related items or seeking a more broadly defined (or creative) set of alternatives. If tasked with identifying a flying object, for instance, the former might suggest the object is one among many missiles of a particular class, whereas the latter might suggest a set of possible missile classes it could belong to. Given that the models exhibit similarly correct selection overall, the preference for either approach is dependent on the specific demands of the task being performed.

Upon examining the probabilities assigned by the models to probe responses, some tradeoffs between the HiMean and LSA approaches again emerge. In base rate

control condition, the Brier scores indicate that the HiMean models have a clear advantage over LSA. The most confident of the four models (in aggregate) is the Standard HiMean model with working memory due to its propensity to terminate search for alternatives when the best (or an excellent) choice has been found and its four item capacity limitation (Table 4). Unfortunately, it's also the second worst performing model in terms of calibration, though still markedly better than LSA. It is difficult to imagine an environment where a model with the characteristic performance of Standard HiMean evinced in Table 1 would be preferred, except to say that this captures human abilities to satisfice and perform non-exhaustive search. This could have benefits where search takes place over an extremely large decision space, computational or time limitations are present, or a satisfactory rather than perfect solution is acceptable. Additionally, the Standard HiMean model with working memory is the only model with human-like limitations and the only model which returns suboptimal alternatives.

Table 5 shows LSA to be the only model sensitive to base rates in terms of average probability judgments rendered for affected diseases, and only for the psychological diseases, and even then the difference is small and in the opposite direction that might be expected. That is, in the conditions where the base rate for psychological diseases was increased, the average probability judgments associated with hypotheses belonging to the psychological disease category actually decreased relative to the control and cardiovascular only conditions. I speculate that this occurs because the increase in base rate reduces the discriminability of these items relative to each other, thereby making the probability judgments rendered about them regress toward the mean. The fact that LSA does not display this effect for cardiovascular

diseases might be explained by the idea that cardiovascular diseases are more inherently similar to each other relative to psychological disorders (which can have vastly different etiologies, symptoms, and treatments, for example). Thus, the increased base rate for cardiovascular diseases does little to reduce the already lowered discriminability between cardiovascular diseases and thereby does little to further lower their respective probabilities.

When examining the Brier scores associated with each model's hypothesized responses, a reverse pattern in base rate sensitivity is displayed. Where HiMean models take advantage of base rate information when rendering probability judgments, the LSA model does not, instead exhibiting a slight decrease in calibration (Table 6, Table 7). Again, this is likely due to the fact that base rate increases seem to have a weakening effect in terms of discriminability for LSA, whereas they strengthen discriminability in the HiMean memory architectures. Also of note is the increase in performance of semantic memory as compared to episodic memory in the HiMean models. This occurs because the probability judgments associated with episodic memory are based solely on the activations of individual traces in response to the probe, whereas the semantic probability judgments are based on the semantic memory's response to an activation weighted composite of all relevant episodic traces. Standard HiMean is outperformed by Standard HiMean without working memory, which is outperformed by the IO HiMean model.

Each of the four models demonstrate some degree of sensitivity to base rate manipulations with performance generally increasing for the HiMean models and decreasing for LSA. The floor effect has dampened the impact of the base rate

influences in the case of IO HiMean and to a lesser extent Standard HiMean without working memory, but they are still discernible (Table 6 and Table 7). The basic Standard HiMean model, however, greatly demonstrates the benefits to calibration brought on by changes in base rate (except, peculiarly, in the psych 5 condition, a phenomenon for which I cannot give an account), seen easily by comparing its performance in the base rate condition to that of its performances in manipulated and congruent conditions. If highly attuned probability judgment is being sought, the IO HiMean model seems to be the best performing of the models overall when the query information is completely intact.

Although the probe quality manipulations did not seem to bear very dramatic results, the patterns displayed in the model output are still somewhat informative. The lack of fluctuations in outcome as a result of probe quality could be due to the relatively stable semantic spaces rendered by the corpora preprocessing or a good degree of semantic separation between concepts in the corpora itself. That is, probes (contexts) may have been dissimilar enough to begin with, that changing the features with even a moderate probability fails to make them look alike, until they looked so much alike that the models could not perform (as in the case of HiMean). LSA does not require a threshold of activation to be met in order to posit a hypothesis. Thus, its generation process is insensitive to probe quality, but its selection process is not. The only impact that decreasing probe quality can have on the choices made by the LSA model is to cause it to select the wrong alternative because of increased confusability (i.e., decreased dissimilarity between contexts). Given that the accuracy of LSA best choices did decline greatly with deficient probes (Table 8), it is plain that LSA is not simply

immune to such manipulations. Despite the HiMean models producing more calibrated judgments, however, the lack of a required threshold of activation for the LSA machinery to function does present a distinct advantage under the right circumstances, such as in this case.

The fact that the models scored in a qualitatively similar manner regardless of probe quality can also be seen as a relative strength of the models. Moreover, even though the pattern of findings are generally in the direction of showing a decrease in performance with poorer quality probes, the decline is slow. This indicates that when these models are operating over complex spaces with large representations, model performance can still be excellent despite severely degraded input.

In terms of the model's diagnostic capabilities and their ability to function in a novel, but related domain, they performed well. Overall accuracy of around 30% is admirable considering the sheer number of feature values that could have been represented, but were not in the short descriptions used to query the models. Additionally, their well-calibrated probability judgments and capacity for at least suggesting the correct choice among the alternatives is testament to the power of these models to find semantic associations. It is a little surprising that LSA did not perform better as it did have the advantage of having its query probe "translated" into concept space which should have imparted additional information to it that may not have been present in the original query vector. This makes the finding that IO HiMean performed at least as well and, by some metrics, better than the LSA model even more impressive. Given that model accuracy was approximately equivalent, I would have to give the advantage to the IO HiMean model for its more informative probability judgments.

More importantly, however, is the demonstration of the HiMean model's capability to operate in an untrained environment using real-world input.

The plots of the multidimensional scaling of the models' semantic spaces revealed that the 2D projections of LSA and HiMean are very different. LSA has a more ordered structure and demonstrates more space between clusters than do the HiMean models. The graphs also make the increase in similarity between concepts as their base rates increase easy to see and may provide a more intuitive understanding of how probability assessment for any particular concept may suffer under such conditions. On the other hand, while the changes in the space structure for HiMean are more subtle, the shift that the concepts do experience seem to be enough to benefit the model's probability judgments to a good degree. Finally, the depictions of the semantic spaces seem to suggest that the semantic spaces that LSA forms are more insulated against perturbations that do not directly impact individual categories. In other words, changes in the semantic structure of one part of the space, do not seem to have a large impact on other parts of the space. In the IO HiMean model, the effects of tampering with one aspect of the space seem to have a more diffuse influence on the rest of it. Both characteristics can be advantageous depending on their application.

Limitations

There are a number of limitations to this work that should be considered. The first is that a number of implementation decisions had to be made in order to get the models to perform. These decisions obviously lead to inequalities that make it difficult to render unskewed comparisons between them. Perhaps of even greater concern is the

unintended consequences these decisions may have on model outcomes. Preprocessing text is, in itself, a way of pruning the information so that the leftover information is associated and meaningful. Here, pre-processing decisions were made according to previously established work, but in this domain, these decisions are key. Another limitation of this study was the amount of overlap between the disease categories. It is difficult to get a sense of the accuracy of a model when it proposes an answer that could simultaneously be classified in three or four different ways. Not being a medical expert, it would have been difficult for me to assign definite category membership without concern for unduly biasing the results. At the same time, working in an information environment where all borders between topics were clear cut would have defeated the purpose of the experimentation as correct classification would have been neatly defined. Perhaps future researchers will work on identifying optimal domains according to this characteristic.

In contrast to the limitation just mentioned, it is also the case that corpora used as the subject matter for this work was conveniently demarcated. Model performance was exceedingly good on a number of experimental trials and much of this owes to the nature of the data used itself. It would be interesting to examine model performance in more complex environments to see if any semblance of model rationality or usefulness could be achieved. Another point to consider is that there were two models compared here, but there are certainly other models out there designed to operate in similar environments. It is difficult to draw objective conclusion about the overall performance of a model with such a small base for comparison. While these models performed similarly in many ways, for example, there is no telling whether their output is even

remarkable on a grander scale without further investigation. Even given these limitations, however, it was still valuable exercise to investigate base rate effects as they relate to semantic modeling, to compare two distinctly different architectures from a new perspective, and to put both models to work in a believable decision-making task using real-world and human-interpretable information.

Future Work

Aside from the opportunities for improvement in future research alluded to in the previous section, there are other theoretically compelling areas of research opened by the discoveries of this dissertation. The present work focused on a single domain. It would be interesting in the future to provide models with overlapping and distinctly different domains of knowledge and gauge their performance on a more well-rounded battery of tasks. Additionally, in this dissertation base rates could be viewed as a type of expertise, but the way in which base rates were manipulate was rudimentary. Future work examining information tailored according to experience could be fruitful in examining learning as well as expertise in a variety of contexts. I would also like to further refine the present work to the point where concepts are not simply demonstrated but the models produce a reliable outcome that can actually be a tool. For example, while priapism and erectile dysfunction are certainly semantically related (and confusable by these models), I would suggest that anyone with a physician who confuses the two should consider a switching practitioners. One way of achieving a more practical performance could be through the use of a combined model.

Multi-agent modeling. Given the findings of this investigation of these models' performances under the parameters described above, it would be extremely useful to develop a multi-agent framework which can capitalize on their strengths and avoid their weaknesses (Alanyali, Venkatesh, Savas, & Aeron, 2004; Ren, Beard, & Atkins, 2005; Yang, Wu, & Bonissone, 2013). The goal could be to create variations of simple consensus models constructed from both HiMean and LSA frameworks in order to 1) assess the feasibility of their integration, and 2) determine if performance can be increased over either of the constituent models alone according to the same measures of diagnostic choice, consideration set, and posterior probability judgment (Rauhut & Lorenz, 2011). For example, LSA demonstrates a great ability to accurate select the correct probe response, and is more robust to decreases in probe quality than the HiMean models as tested here. The HiMean models however, produce more well-calibrated probability judgments as well as allow for both episodic and semantic assessments. Either model may have an advantage dependent upon the relatedness of alternatives desired. Thus finding a working amalgamation of these models may prove extremely useful. This represents a novel effort on a number of fronts. First, as discussed, models of semantic analysis are not generally constructed with performing ecologically-based diagnosis tasks in mind. Second, where their performances are compared, they are evaluated as standalone models and no effort is made to combine them. Third, this represents a new domain for decision-based multi-agent semantic models as, to my knowledge, no such effort has been academically pursued or published in research.

Using the information gleaned from the previously outlined work, it should be possible to construct a multi-agent decision support system that exploits the positive capabilities of the HiMean and LSA models while mitigating their weaknesses. By constructing multiple decision layers where the constituent agents operate according to individualized parameters (e.g., IO HiMean in tandem with LSA using adaptive dimensionality, etc.) optimized for particular circumstances (e.g., base rate information known important, probe error known, etc.) and by using a final decision-making component that has access to information allowing it to adaptively handle input from these layers (e.g., using a method of evidence evaluation such as Dempster-Shafer theory), it may be possible to improve performance (Klopotek & Wierzchoń, 2002).

The integration of these models could be carried out in a number of ways (e.g., using intersection regions, voting methods, top choice combinations, etc.) according to methods described in previous literature on multiple classifiers (Ho, Hull, & Srihari, 1994; Xu, Krzyzak, & Suen, 1992). Can a multi-agent system lead to better performance? How does the composition (number/type of agents and layers) of this system affect performance? What are the optimized weights for each agent/layer's inputs for producing best performance? It is possible to evaluate the optimality of systems resulting in identical decision outcomes despite those outcomes having arisen from systems with varying underlying processes, strategies, and complexity (Glöckner, 2009). These experiments represent a novel integration of various lines of research and would provide the opportunity to answer questions in a new domain, gain new insights into the practical applications of real-world semantics-based cognitive models, and open the door to new directions in research.

Conclusion

Although models of semantic analysis have been around for decades and alternative methods for accomplishing related work have been proposed, it seems their usefulness has not expired, nor has their domain been completely explored. This work represents an important first step toward applying a semantically-based HyGene framework in real-world decision support frameworks. Moreover, LSA has been comparatively explored in a novel light, assessing its suitability in a specific and highly constrained application relative to a never before tested memory-based semantics model. Base rate information is important in human decision-making and it has been demonstrated here to be important to cognitive models as well. Though there are tradeoffs involved in selecting any one of the models discussed here, each model has something important to offer in this domain. One must carefully consider the constraints of their operating environment as well as their desired outcomes if one hopes to maximize model selection. There are many interesting opportunities for future research stemming from this work, perhaps chief among them the potential to deploy a multi-agent framework which powerfully combines the qualities of LSA and HiMean. Computationally-augmented decision making is already a vital part of our lives and will only continue to increase in necessity in the future. It will be essential to understand the dynamics involved in these complex systems in order to maximize our potential to benefit humankind. I hope this work and others like it will serve as at least a small step in that progressively important direction.

References

Alanyali, M., Venkatesh, S., Savas, O., & Aeron, S. (2004, June). Distributed Bayesian hypothesis testing in sensor networks. In *American Control Conference, 2004. IEEE Proceedings of the 2004* (Vol. 6, pp. 5369–5374).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Bronstein, A. M., Bronstein, M. M., Zibulevsky, M., & Zeevi, Y. Y. (2005). Sparse ica for blind separation of transmitted and reflected images. *International Journal of Imaging Science and Technology*, *15*, 84–91.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106*, 180-209.

Dougherty, M. R., Thomas, R. P., & Lange, N. (2010). Toward an Integrative Theory of Hypothesis Generation, Probability Judgment , and Hypothesis Testing. *Psychology of Learning and Motivation, 52*, 299–342.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making*, *4*(3), 186–199.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive society* (pp. 381–386). Mahwah, NJ: Lawrence Erlbaum Associates.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96–101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace model. *Psychological Review, 93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition in a multiple-trace memory model. *Psychological Review, 95*, 528–551.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 16*(1), 66–75.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37. doi:10.1037/0033-295X.114.1.1

Kintsch, W., McNamara, D., Dennis, S., & Landauer, T. (2006). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Klopotek, M. A., & Wierzchoń, S. T. (2002). Empirical models for the Dempster-Shafer-Theory. *Belief Functions in Business Decisions, 88*, 62.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin and Review, 12*, 703–710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Raaijmakers, J. G. W., Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge, England: Cambridge University Press.

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of mathematical Psychology*, *55*(2), 191–197.

Ren, W., Beard, R. W., & Atkins, E. M. (2005, June). A survey of consensus problems in multi-agent coordination. In *American Control Conference, 2005. IEEE Proceedings of the 2005* (pp. 1859–1864).

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, *4*(4), 249–262.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*, 1–12.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18,* 613–620.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*(1), 155–185.

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on, 22*(3), 418–435.

Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*(pp. 267–273). ACM.

Yang, T., Wu, L., & Bonissone, P. P. (2013). A Directed Inference Approach towards Multi-class Multi-model Fusion. In *Multiple Classifier Systems* (pp. 352–363). Springer Berlin Heidelberg.

Table 1

Model Choices in Base Rate Control Condition

Model	Episodic Proportion of Trials with Correct Top Choice	Semantic Proportion of Trials with Correct Top Choice	Episodic Proportion of Trials with Correct Option Among Best Guesses	Semantic Proportion of Trials with Correct Option Among Best Guesses	Episodic Average Proportion of Best Guesses in Same Category as Probe	Semantic Average Proportion of Best Guesses in Same Category as Probe	Semantic Proportion of Top Guesses in Psychological Category	Semantic Proportion of Top Guesses in Cardiovascular Category
LSA		1.		1.		.56	.014	.069
IO HiMean	.97	1.	1.	1.	.79	.79	.055	.095
Standard HiMean No WMC	.76	1.	.99	1.	.79	.79	.055	.093
Standard HiMean	.86	.88	.88	.88	.77	.77	.025	.045

Table 2

Model Choices in Cardiovascular Disease Base Rate Five Condition

Model	Episodic Proportion of Trials with Correct Top Choice	Semantic Proportion of Trials with Correct Top Choice	Episodic Proportion of Trials with Correct Option Among Best Guesses	Semantic Proportion of Trials with Correct Option Among Best Guesses	Episodic Average Proportion of Best Guesses in Same Category as Probe	Semantic Average Proportion of Best Guesses in Same Category as Probe	Semantic Proportion of Top Guesses in Psychological Category	Semantic Proportion of Top Guesses in Cardiovascular Category
LSA		1.		1.		.56	.013	.19
IO HiMean	.98	1.	1.	1.	.79	.79	.054	.095
Standard HiMean No WMC	.75	1.	.99	1.	.79	.79	.054	.097
Standard HiMean	.88	.91	.91	.91	.77	.77	.022	.043

Table 3

Model Best Predictions Proportional Disease Category Membership by Base Rate Condition

	Base Rate 1		Base Rate Cardio 5		Base Rate Cardio 10		Base Rate Psych 5		Base Rate Psych 10		Base Rate Cardio 5 Psych 5		Base Rate Cardio 10 Psych 10	
	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych	Prop. Cardio	Prop. Psych
LSA	**.069**	.014	**.196**	.013	**.205**	.013	**.067**	.034	**.061**	.088	**.221**	.027	**.227**	.063
IO HiMean	**.095**	.055	**.095**	.054	**.096**	.053	**.095**	.054	**.095**	.054	**.094**	.054	**.095**	.054
HiMean no WM	**.093**	.055	**.097**	.054	**.096**	.053	**.094**	.053	**.093**	.053	**.095**	.054	**.095**	.054
Standard HiMean	**.045**	.025	**.043**	.022	**.042**	.024	**.040**	.021	**.044**	.023	**.039**	.025	**.046**	.022

Note. Numbers bolded for clarity.

Table 4

Probabilities and Brier Scores for Base Rate Control Condition

	Avg. Episodic Probabilities	Overall Episodic Brier Score	Avg. Semantic Probabilities	Overall Semantic Brier Score
LSA			.20	**.136**
IO HiMean	.20	**.008**	.20	**.004**
HiMean no WM	.20	**.053**	.20	**.008**
Standard HiMean	.56	**.068**	.56	**.054**

Note. Numbers bolded for clarity.

Table 5

Average Probabilities Associated with All Model Guesses across Base Rate Conditions

	Base Rate 1		Base Rate Cardio 5		Base Rate Cardio 10		Base Rate Psych 5		Base Rate Psych 10		Base Rate Cardio 5 Psych 5		Base Rate Cardio 10 Psych 10	
	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio	Avg. Prob. Psych	Avg. Prob. Cardio
LSA	**.28**	.20	**.29**	.19	**.29**	.19	**.21**	.20	**.20**	.20	**.22**	.19	**.20**	.19
IO HiMean	**.22**	.16	**.22**	.16	**.22**	.16	**.22**	.16	**.22**	.16	**.22**	.16	**.22**	.16
HiMean no WM	**.22**	.16	**.22**	.16	**.23**	.17	**.23**	.16	**.23**	.16	**.22**	.17	**.23**	.17
Standard HiMean	**.68**	.48	**.67**	.46	**.66**	.52	**.73**	.57	**.76**	.50	**.69**	.54	**.76**	.50

Note. Only semantic probabilities are shown because there were no differences found between episodic and semantic average probabilities.

Table 6

Brier Scores for Predictions Responding to Cardiovascular and Psychological Probes across Base Rate Conditions

	Base Rate 1				Base Rate Cardio 5				Base Rate Cardio 10				Base Rate Psych 5				Base Rate Psych 10			
	Psych		Cardio		Psych		Cardio		Psych		Cardio		Psych		Cardio		Psych		Cardio	
	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem
LSA		.123		.139		.122		.154		.122		.157		.141		.139		.151		.138
IO HiMean	.001	.001	.016	.008	.001	.001	.008	.006	.001	.001	.008	.006	.001	.001	.016	.008	.0015	.001	.021	.008
HiMean no WMC	.024	.002	.100	.017	.024	.002	.090	.017	.028	.002	.088	.017	.027	.002	.103	.018	.026	.002	.097	.017
Standard HiMean	.011	.011	.153	.117	.020	.020	.118	.087	.020	.020	.051	.027	.041	.041	.115	.079	.0001	.00003	.094	.054

Table 7

Brier Scores for Predictions Responding to Cardiovascular and Psychological Probes across Base Rate Conditions

	Base Rate Cardio 5 Psych 5				Base Rate Cardio 10 Psych 10			
	Psych		Cardio		Psych		Cardio	
	Ep	Sem	Ep	Sem	Ep	Sem	Ep	Sem
LSA		.139		.155		.151		.157
IO HiMean	.0015	.001	.008	.006	.0015	.001	.008	.006
HiMean no WM	.027	.002	.089	.017	.026	.002	.091	.017
Standard HiMean	.003	.0001	.117	.087	.050	.040	.096	.074

Table 8

Proportion of trials with correct top choices rendered according to probe quality

Semantic Memory					
Probe Integrity	1.0	0.9	0.8	0.7	[…] 0.1
LSA	1	.97	.97	.95	.63
IO HiMean	1	1	1	1	*
HiMean no WMC	1	1	1	1	*
Standard HiMean	.88	.87	.88	.88	*
Episodic Memory					
LSA	*	*	*	*	*
IO HiMean	.97	.93	.89	.85	*
HiMean no WMC	.76	.81	.85	.89	*
Standard HiMean	.86	.85	.85	.86	*

**No output could be produced by the model*

Table 9

Proportion of trials with correct option among top choices according to probe quality

Semantic Memory					
Probe Integrity	1.0	0.9	0.8	0.7	[…] 0.1
LSA	1	.99	.99	.98	.76
IO HiMean	1	1	1	1	*
HiMean no WMC	1	1	1	1	*
Standard HiMean	.88	.87	.88	.88	*
Episodic Memory					
LSA	*	*	*	*	*
IO HiMean	1	1	1	.998	*
HiMean no WMC	.99	.99	.99	.99	*
Standard HiMean	.88	.87	.87	.88	*

**No output could be produced by the model*

Table 10

Brier scores according to probe quality

Semantic Memory					
Probe Integrity	1.0	0.9	0.8	0.7	[…] 0.1
LSA	.136	.150	.150	.150	.125
IO HiMean	.004	.006	.008	.010	*
HiMean no WMC	.008	.009	.010	.010	*
Standard HiMean	.054	.058	.057	.054	*
Episodic Memory					
LSA	*	*	*	*	*
IO HiMean	.008	.016	.024	.033	*
HiMean no WMC	.053	.042	.034	.027	*
Standard HiMean	.068	.066	.068	.063	*

**No output could be produced by the model*

Table 11

Model performance on diagnosis task

	Brier Score	Proportion Correct Top Choice	Proportion Correct Choice Among Alternatives	Proportion of Alternatives in Correct Category
Semantic Memory				
LSA	.134	.3	.8	.58
IO HiMean	.010	.3	.8	.46
HiMean no WMC	.133	.3	.6	.34
Episodic Memory				
LSA	*	*	*	*
IO HiMean	.095	.3	.8	.54
HiMean no WMC	.128	.1	.5	.34

**No output could be produced by the model*

Appendix B: Figures

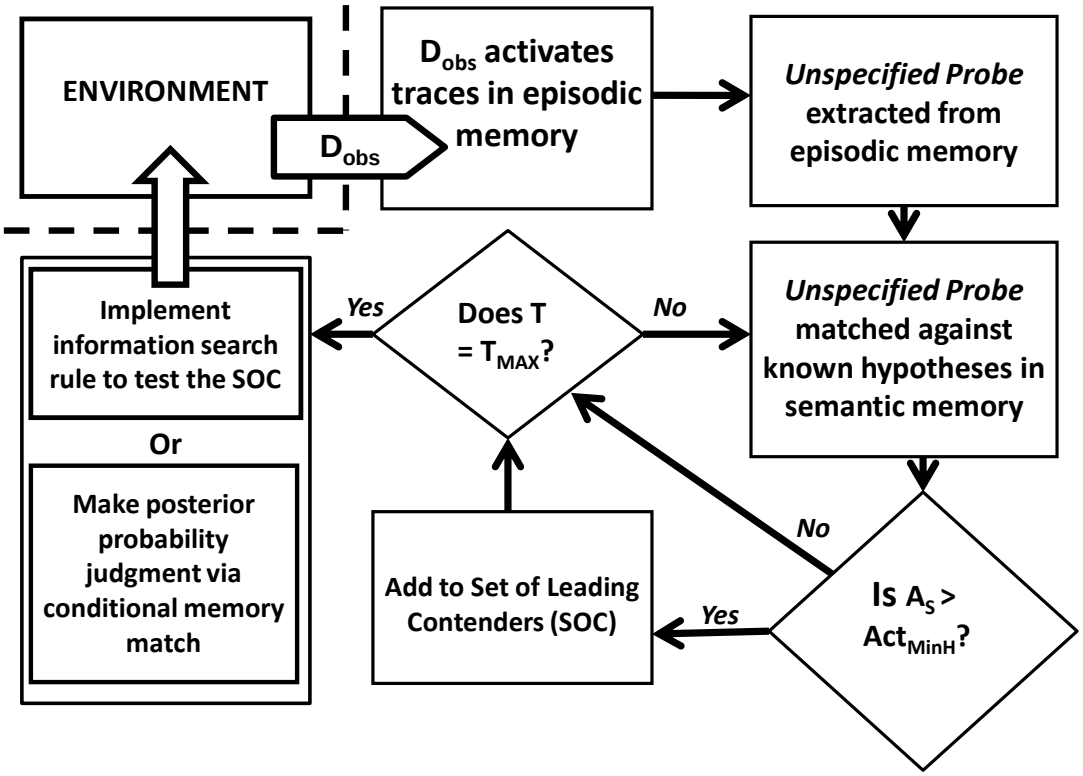

Figure 1. HyGene Architecture.

		Document				
		1	2	.	.	1,000
Word	1	x	x	.	.	x
	2	x	x	.	.	x
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	30,000	x	x	.	.	x

Figure 2. Example term x document matrix.

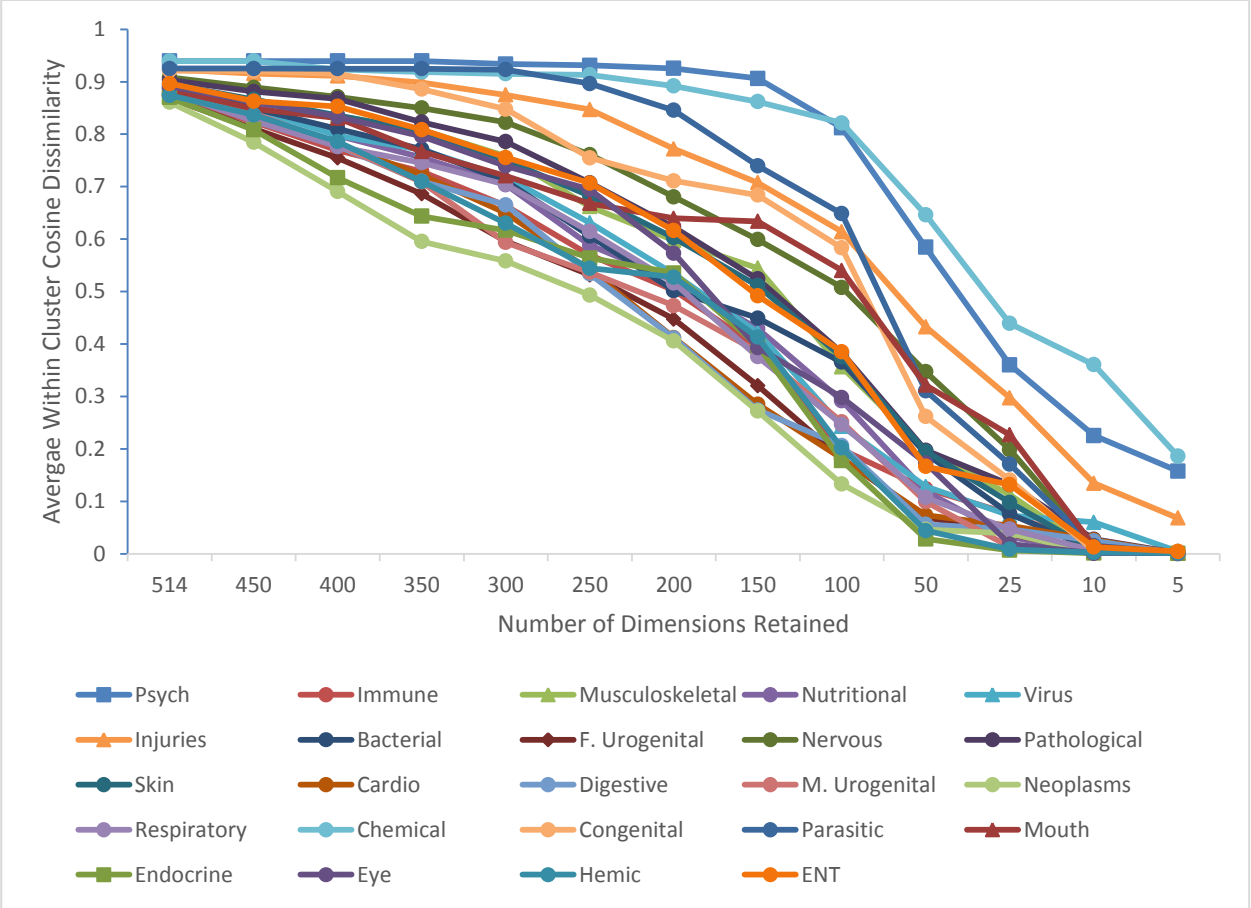

Figure 3. Average within disease cluster dissimilarity as a function of dimension reduction with disease base rate of one.

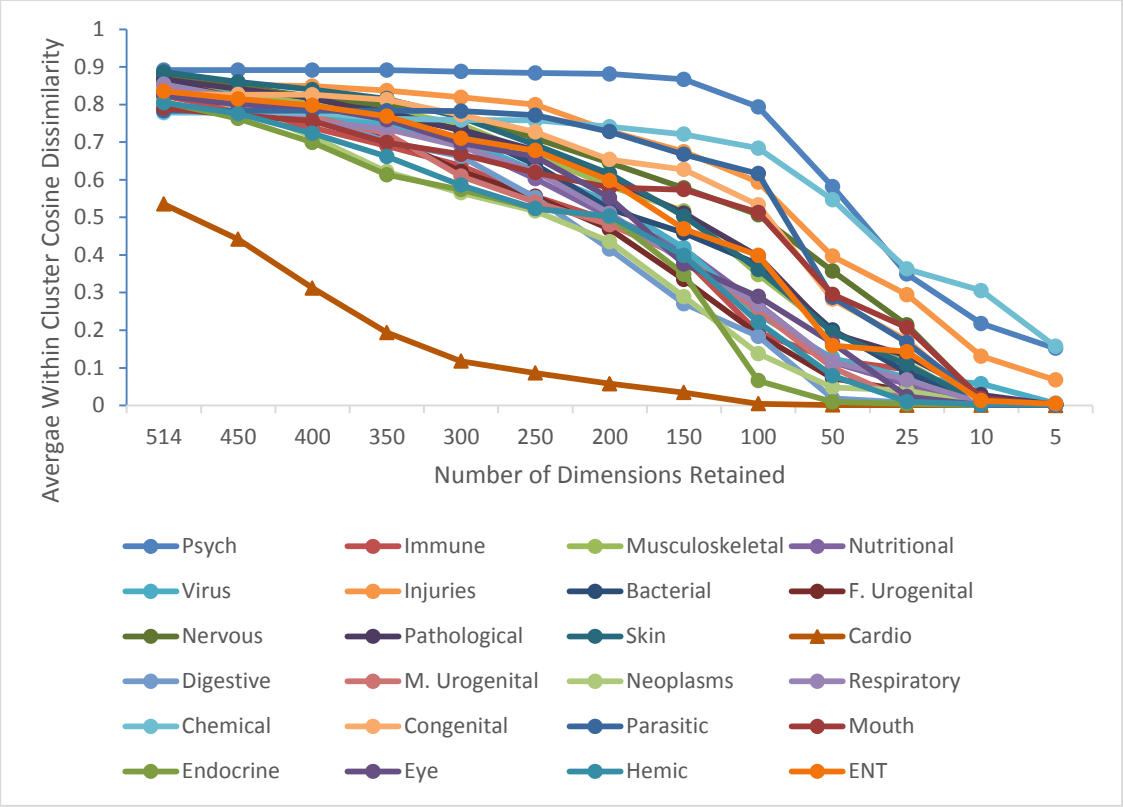

Figure 4. Average within disease cluster dissimilarity as a function of dimension reduction with cardiovascular disease base rate of five and all other disease clusters with a base rate of one.

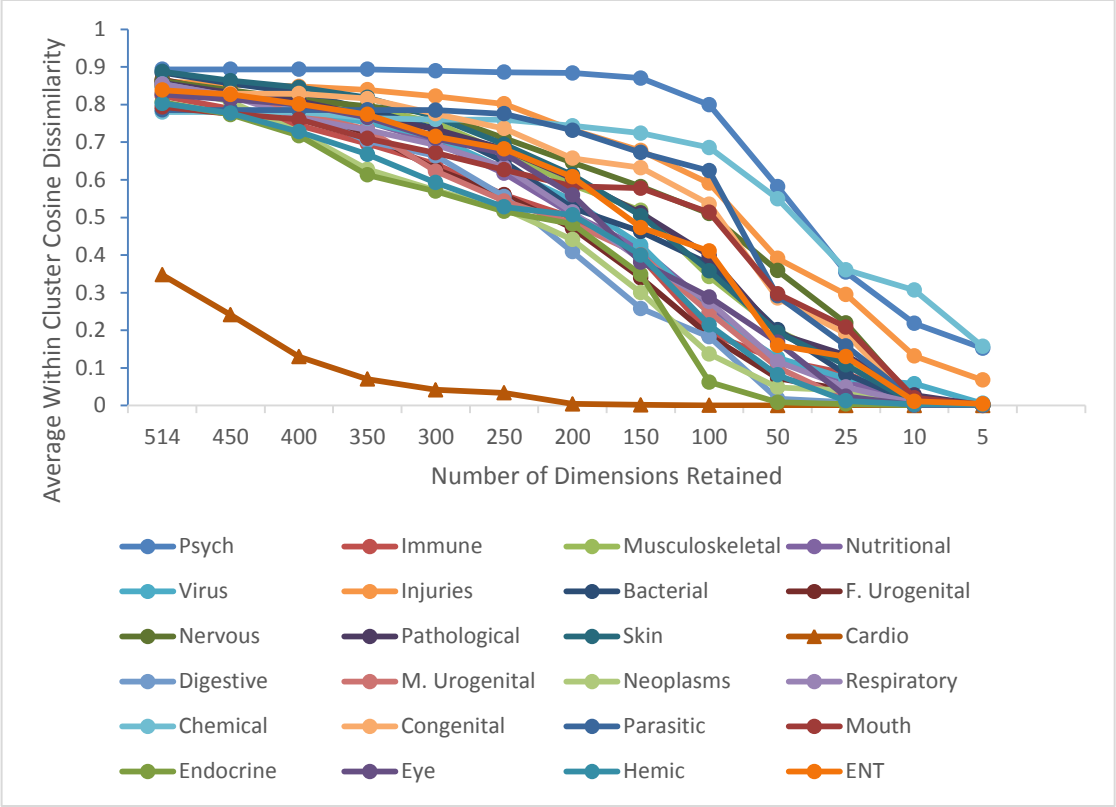

Figure 5. Average within disease cluster dissimilarity as a function of dimension reduction with cardiovascular disease base rate of ten and all other disease clusters with a base rate of one.

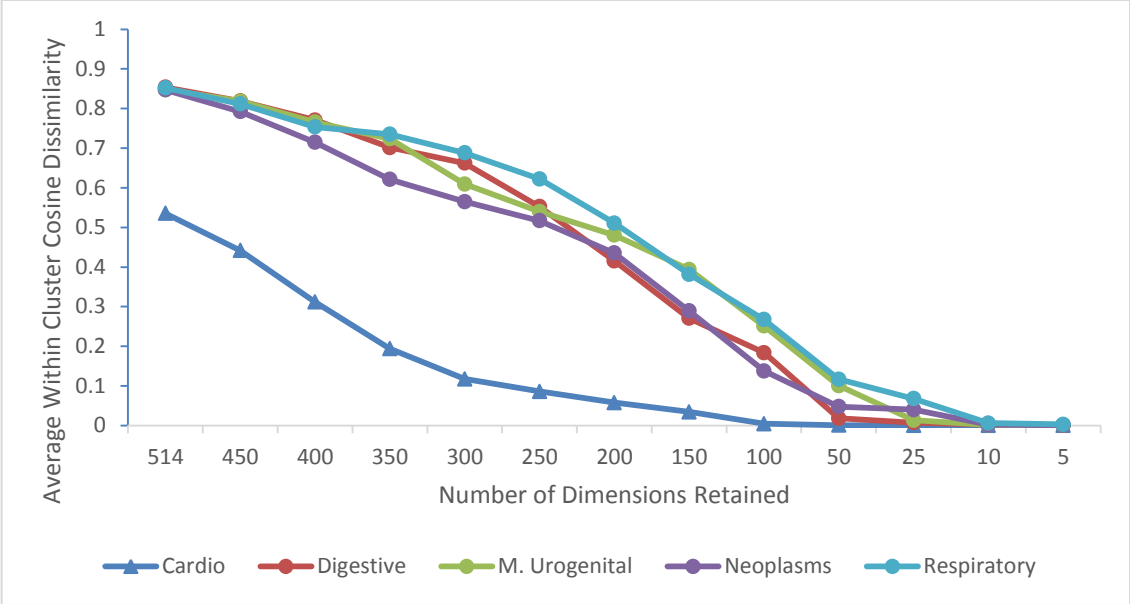

Figure 6. Average within disease cluster dissimilarity as a function of dimension reduction with cardiovascular disease base rate of five and all other disease clusters with a base rate of one for all disease clusters in size cluster three.

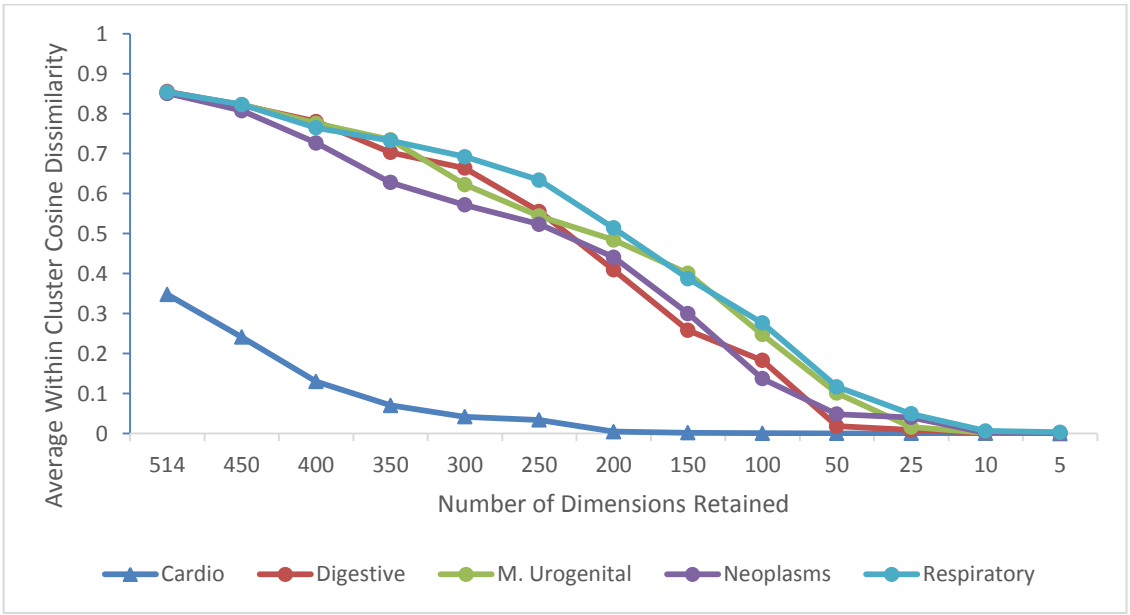

Figure 7. Average within disease cluster dissimilarity as a function of dimension reduction with cardiovascular disease base rate of ten and all other disease clusters with a base rate of one for all disease clusters in size cluster three.

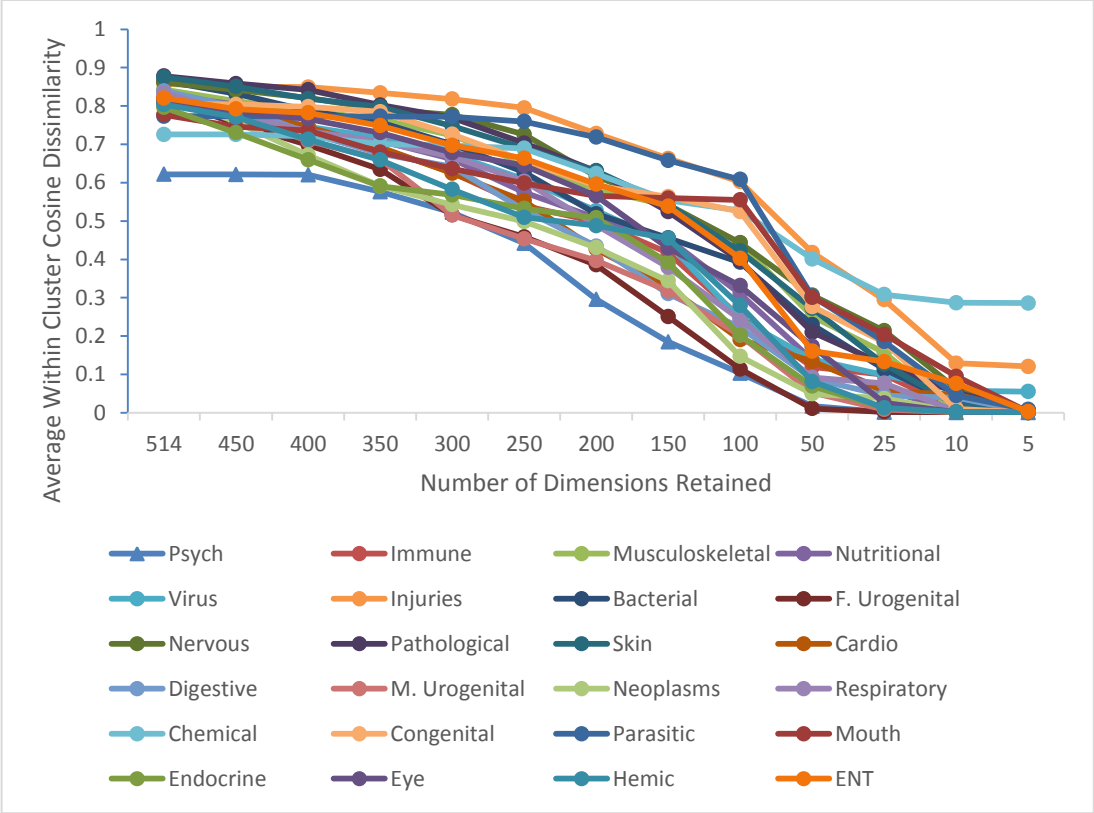

Figure 8. Average within disease cluster dissimilarity as a function of dimension reduction with psychological disease base rate of five and all other disease clusters with a base rate of one.

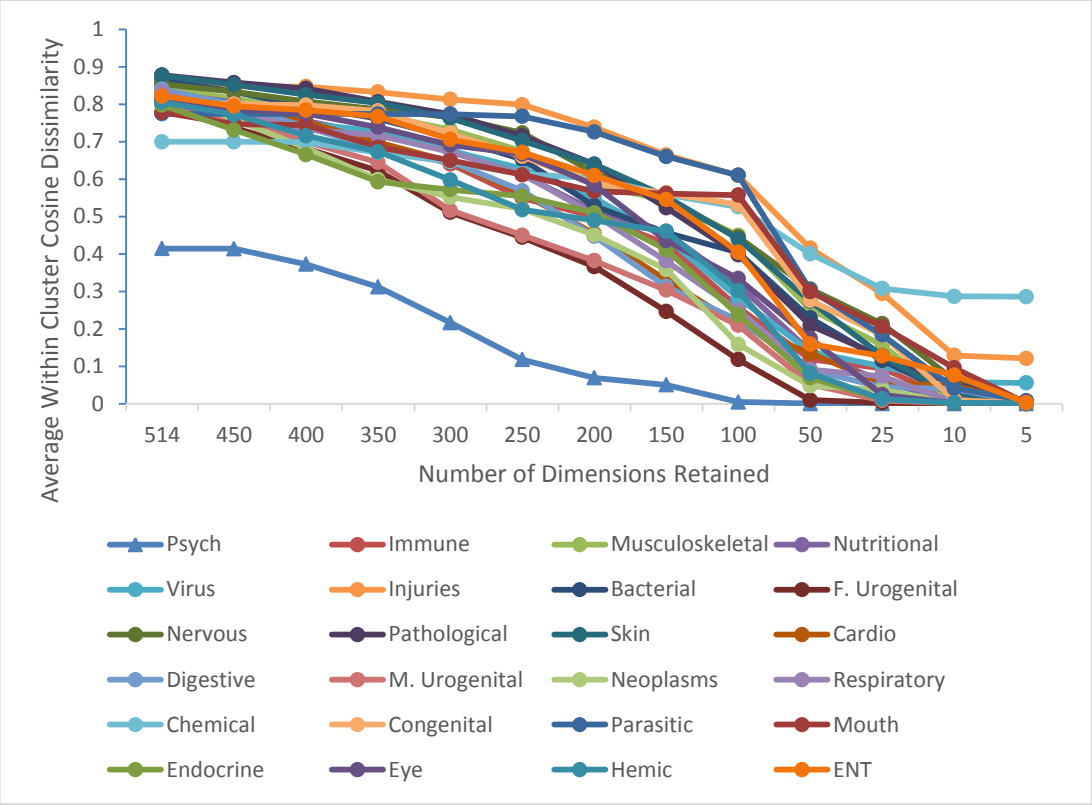

Figure 9. Average within disease cluster dissimilarity as a function of dimension reduction with psychological disease base rate of ten and all other disease clusters with a base rate of one.

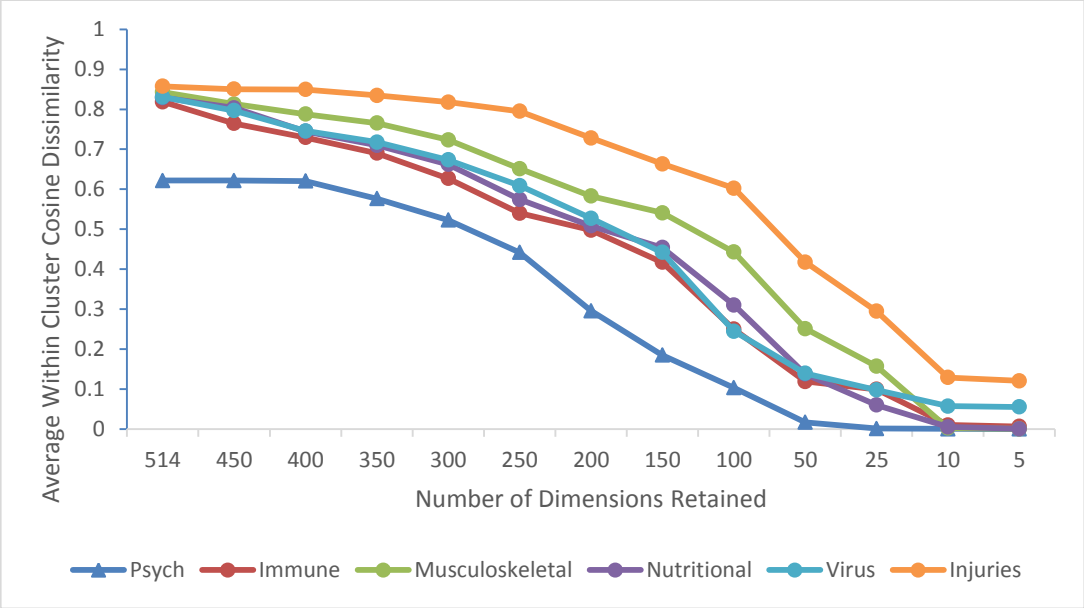

Figure 10. Average within disease cluster dissimilarity as a function of dimension reduction with psychological disease base rate of five and all other disease clusters with a base rate of one for all disease clusters in size cluster one.

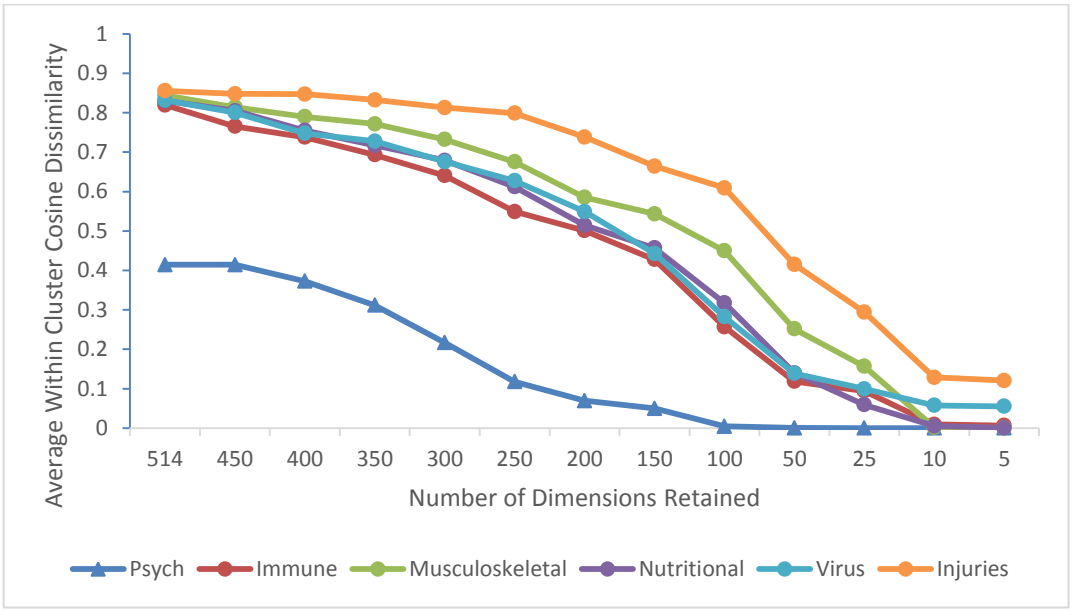

Figure 11. Average within disease cluster dissimilarity as a function of dimension reduction with psychological disease base rate of ten and all other disease clusters with a base rate of one for all disease clusters in size cluster one.

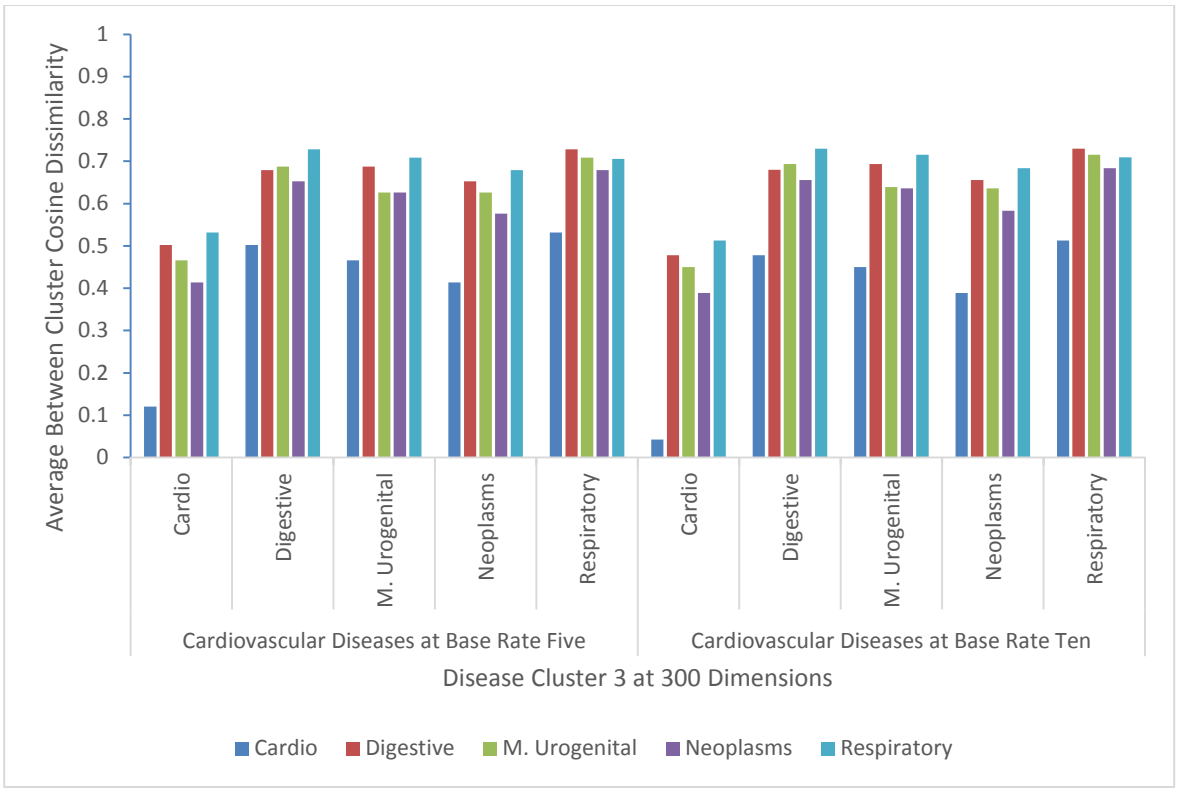

Figure 12. Average between cluster dissimilarities for all disease clusters in size cluster three at 300 dimensions retained and according to cardiovascular disease base rate.

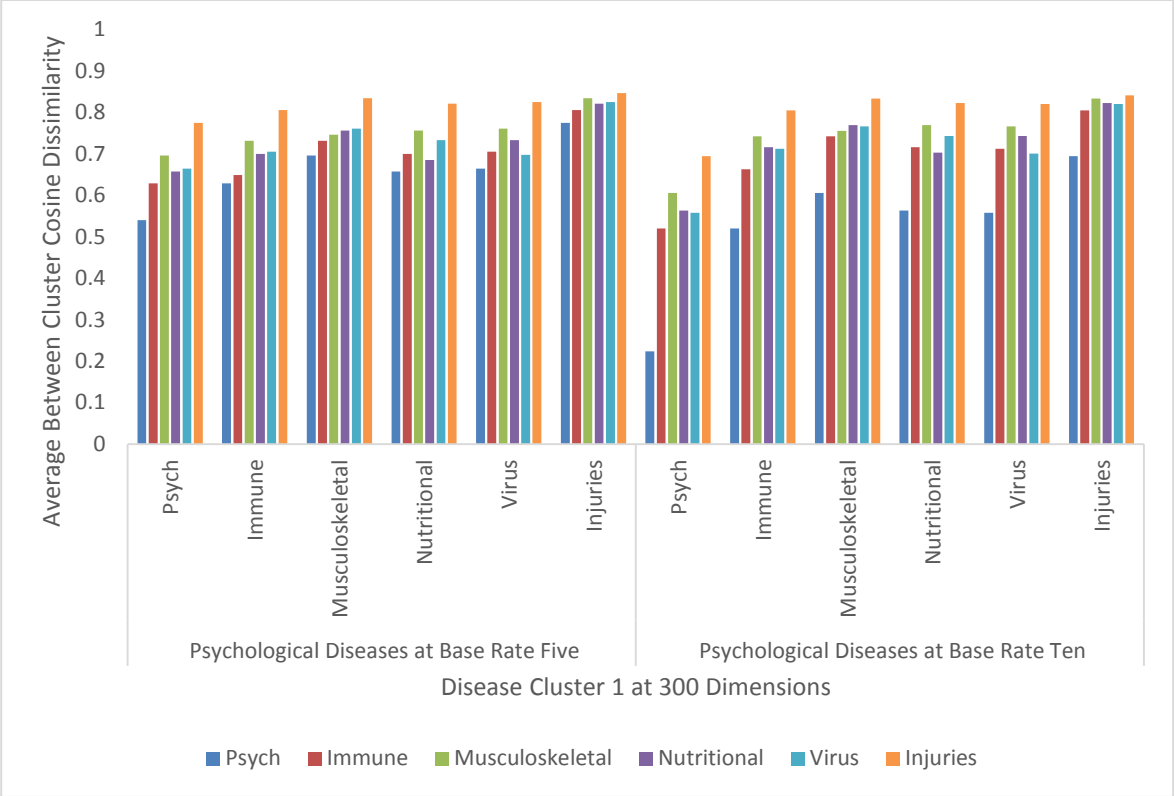

Figure 13. Average between cluster dissimilarities for all disease clusters in size cluster three at 300 dimensions retained and according to psychological disease base rate.

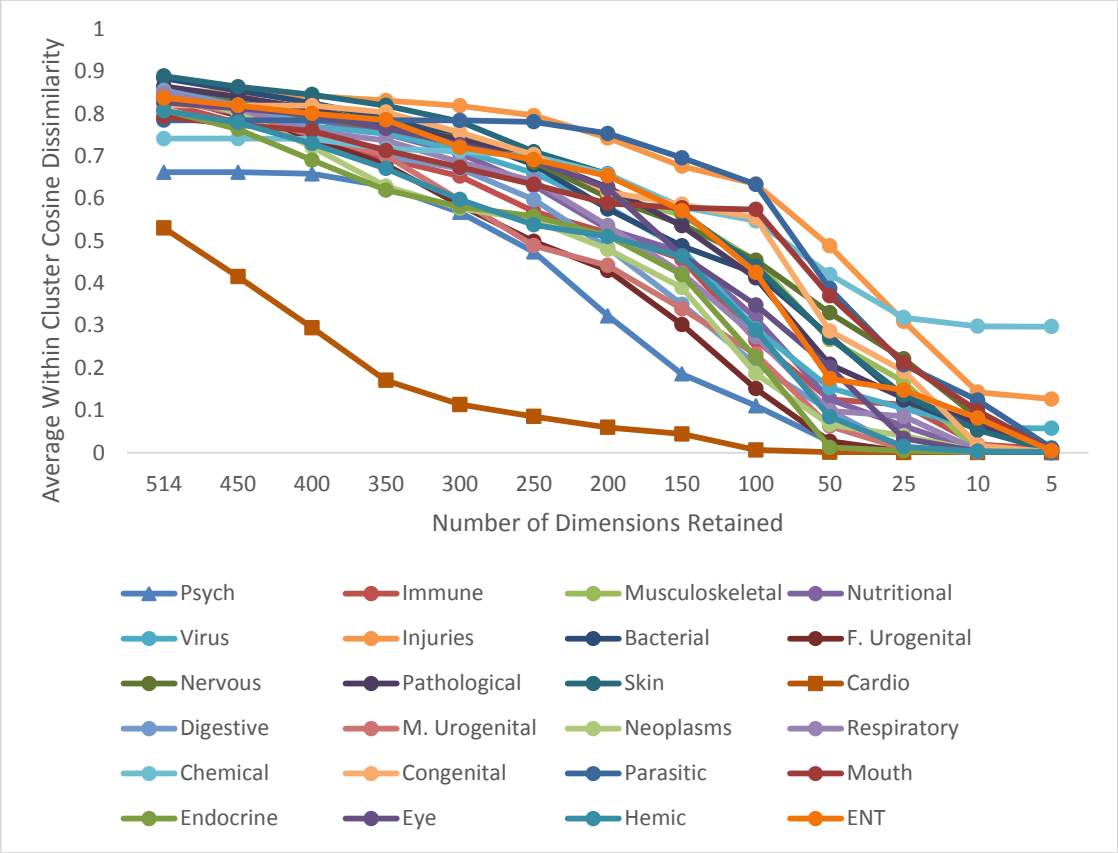

Figure 14. Average within disease cluster dissimilarity as a function of dimension reduction with psychological and cardiovascular disease base rates of five and all other disease clusters with a base rate of one.

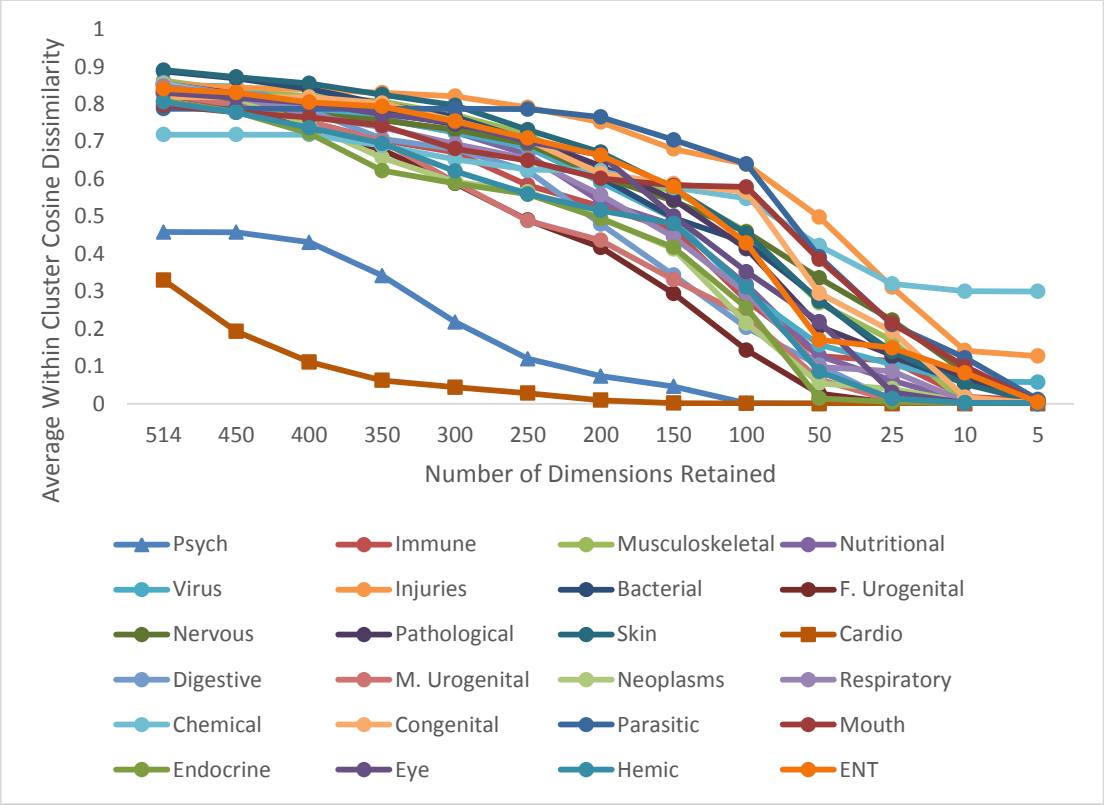

Figure 15. Average within disease cluster dissimilarity as a function of dimension reduction with psychological and cardiovascular disease base rates of ten and all other disease clusters with a base rate of one.

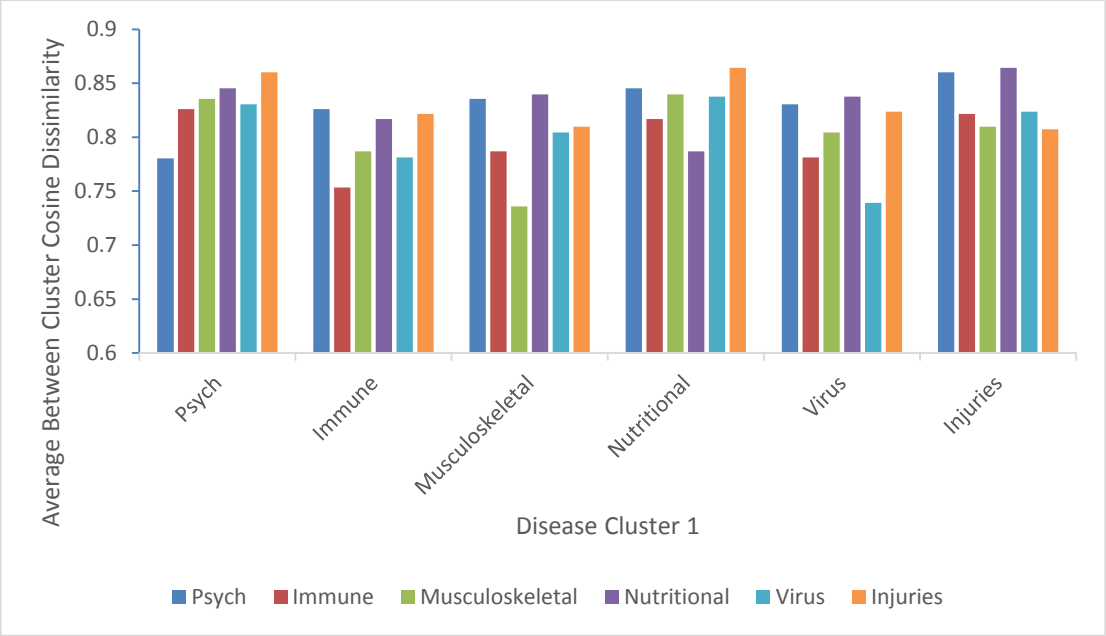

Figure 16. Ideal Observer HiMean model semantic memory between group cosine dissimilarity for diseases in cluster one.

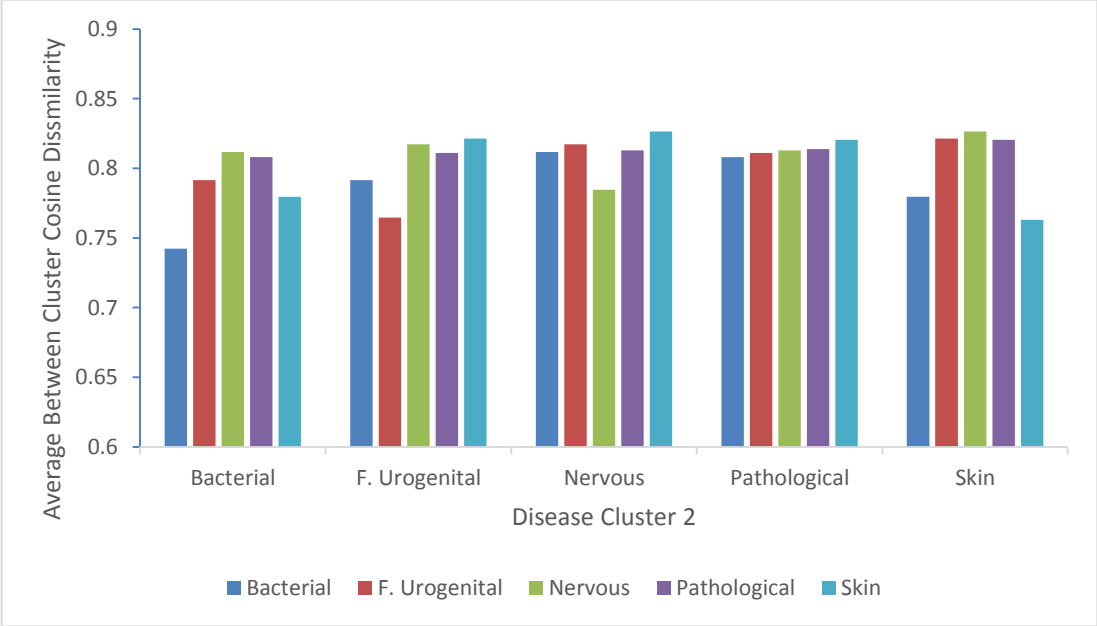

Figure 17. Ideal Observer HiMean model semantic memory between group cosine dissimilarity for diseases in cluster two.

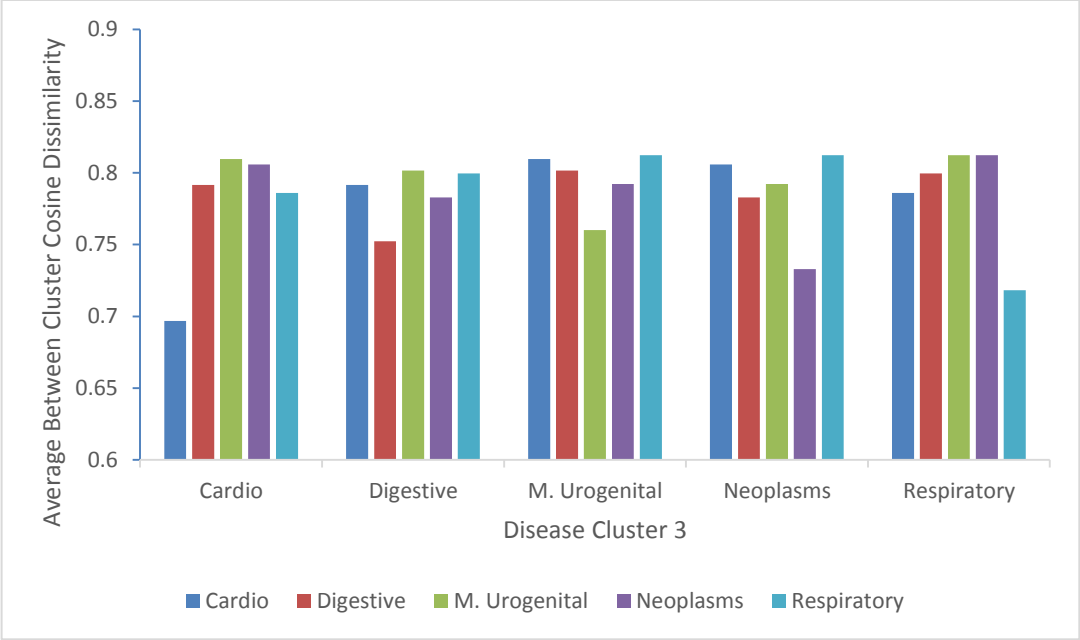

Figure 18. Ideal Observer HiMean model semantic memory between group cosine dissimilarity for diseases in cluster three.

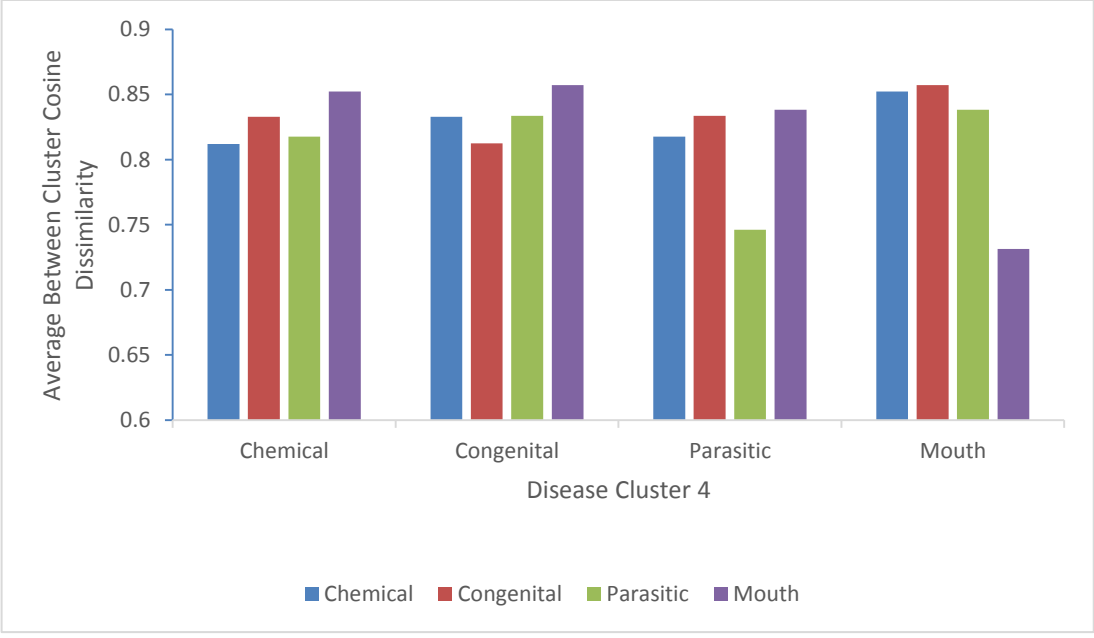

Figure 19. Ideal Observer HiMean model semantic memory between group cosine dissimilarity for diseases in cluster four.

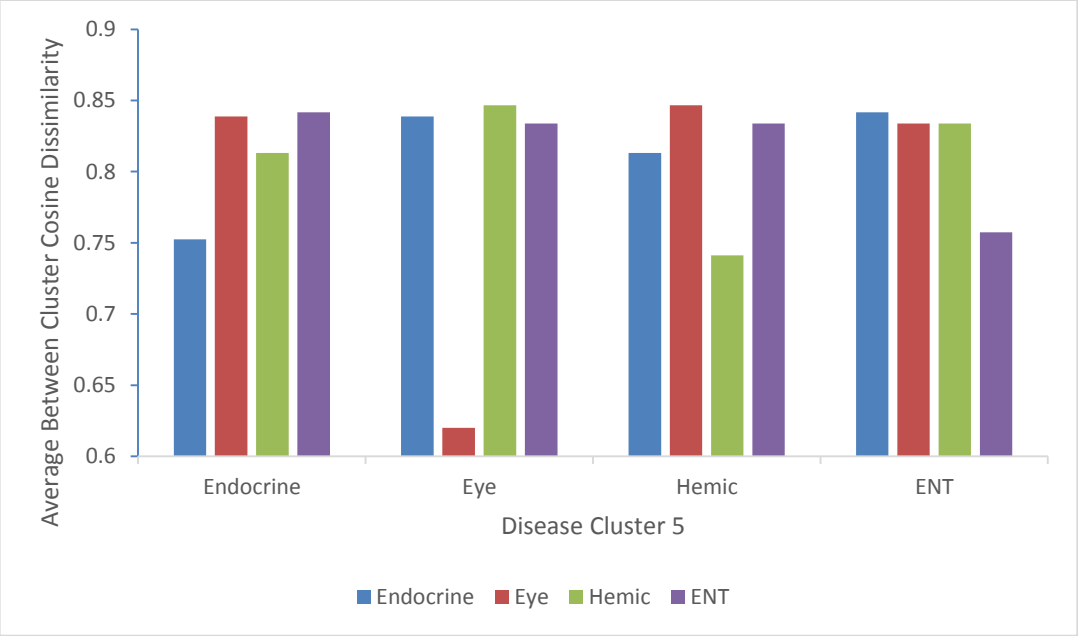

Figure 20. Ideal Observer HiMean model semantic memory between group cosine dissimilarity for diseases in cluster five.

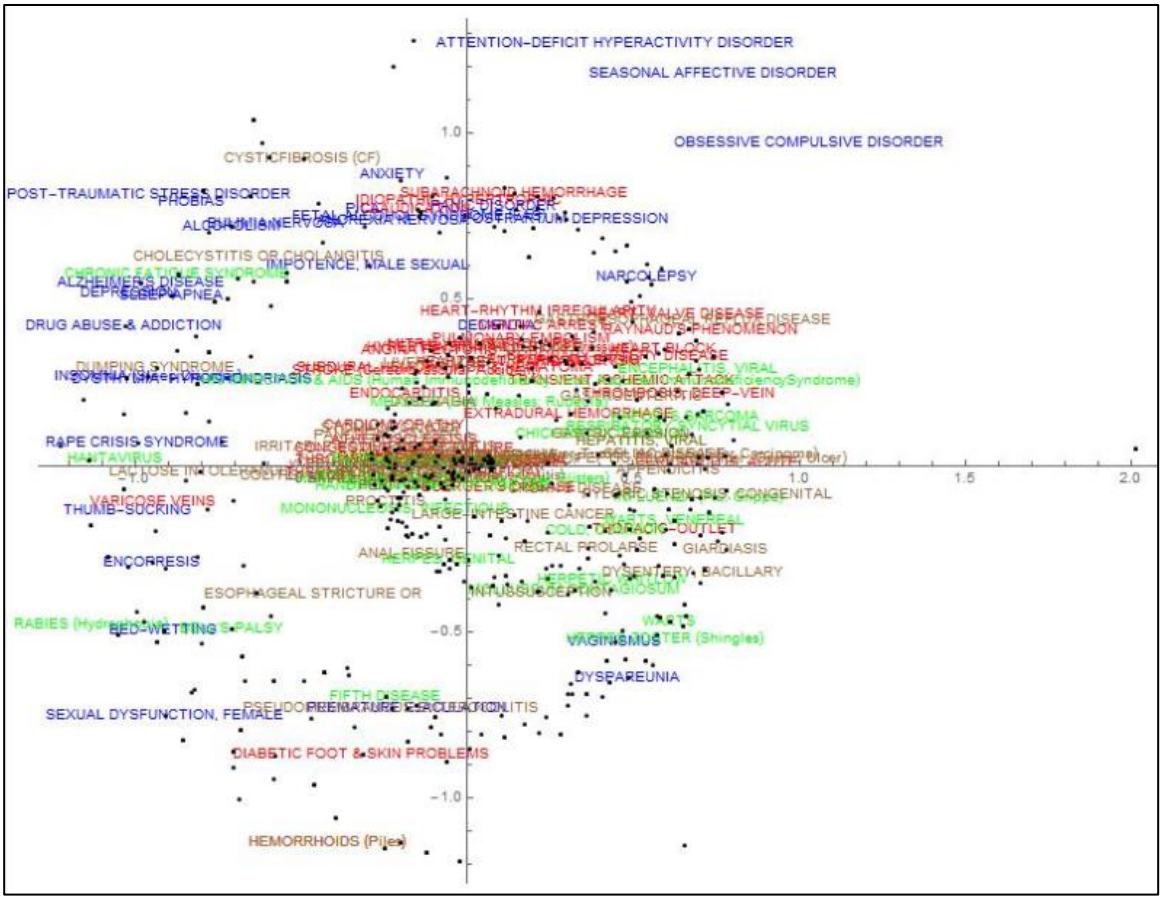

Figure 21. 2D multidimensionally scaled graph of LSA semantic space at full 514 dimensions in base rate control condition.

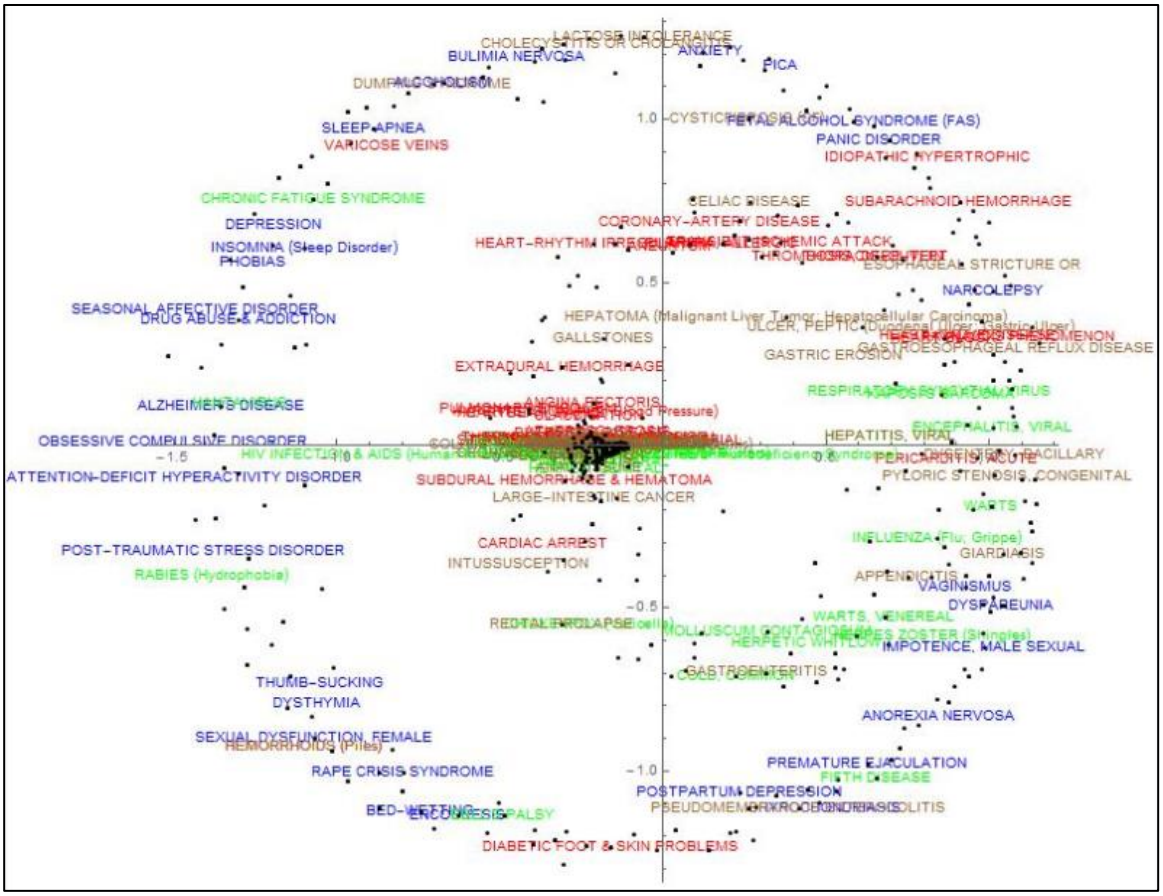

Figure 22. 2D multidimensionally scaled graph of LSA semantic space at 350 dimensions in base rate control condition.

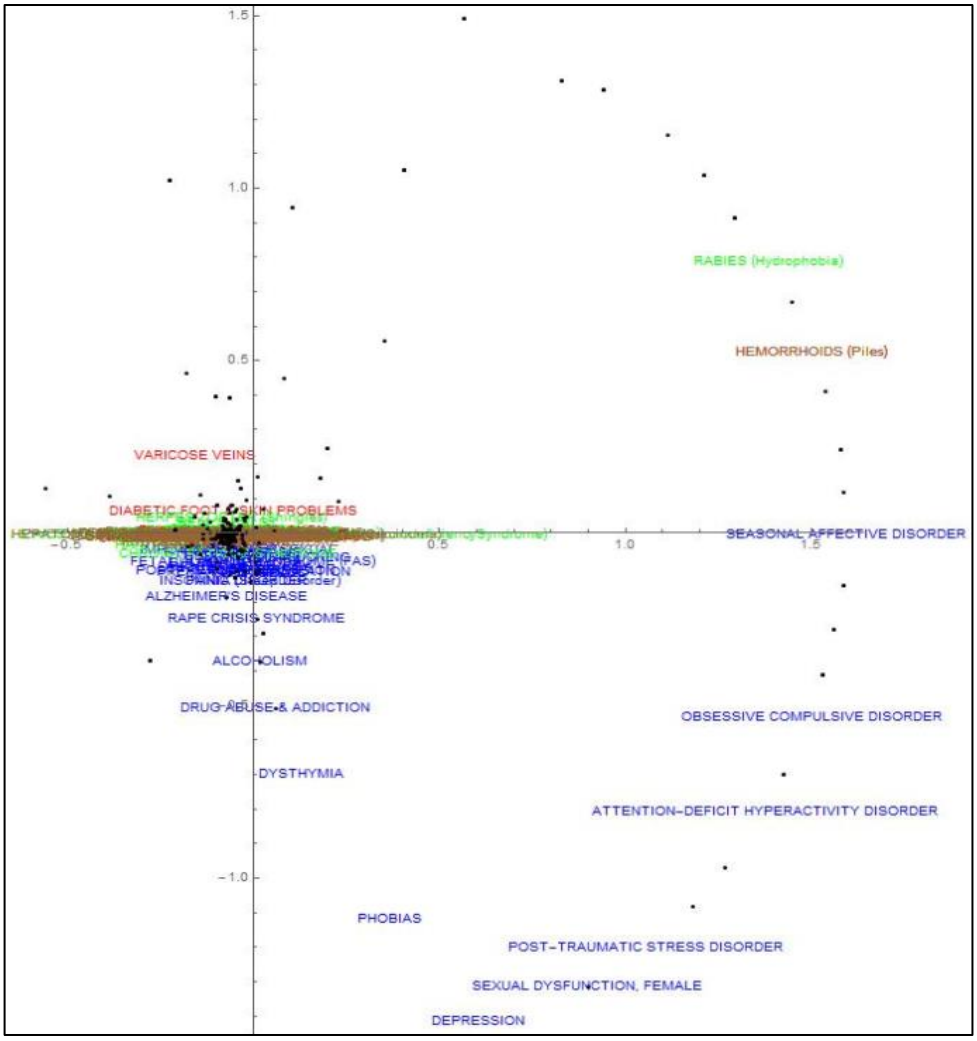

Figure 23. 2D multidimensionally scaled graph of LSA semantic space at 25 dimensions in base rate control condition.

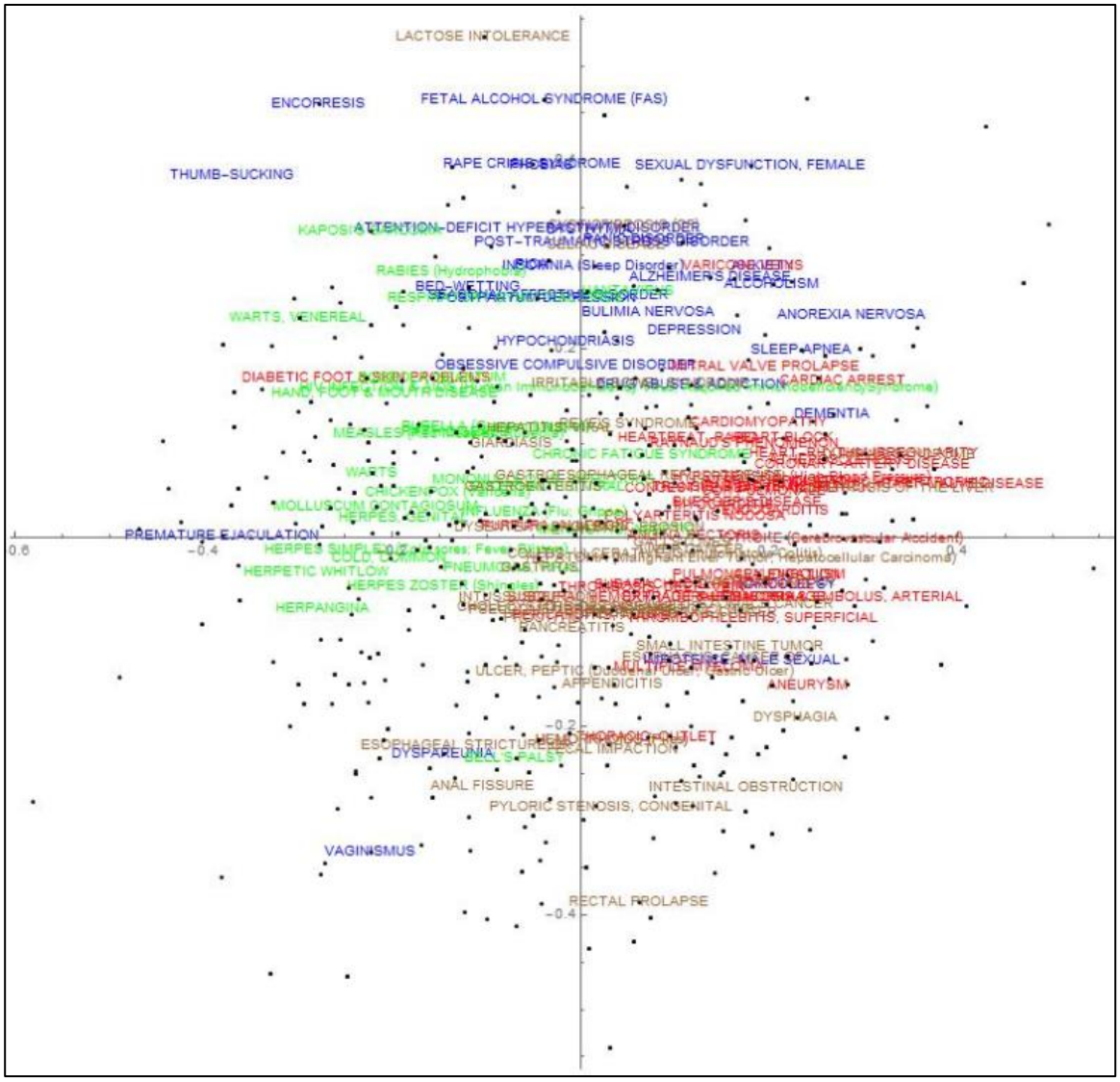

Figure 24. 2D multidimensionally scaled graph of Ideal Observer HiMean semantic space in base rate control condition.

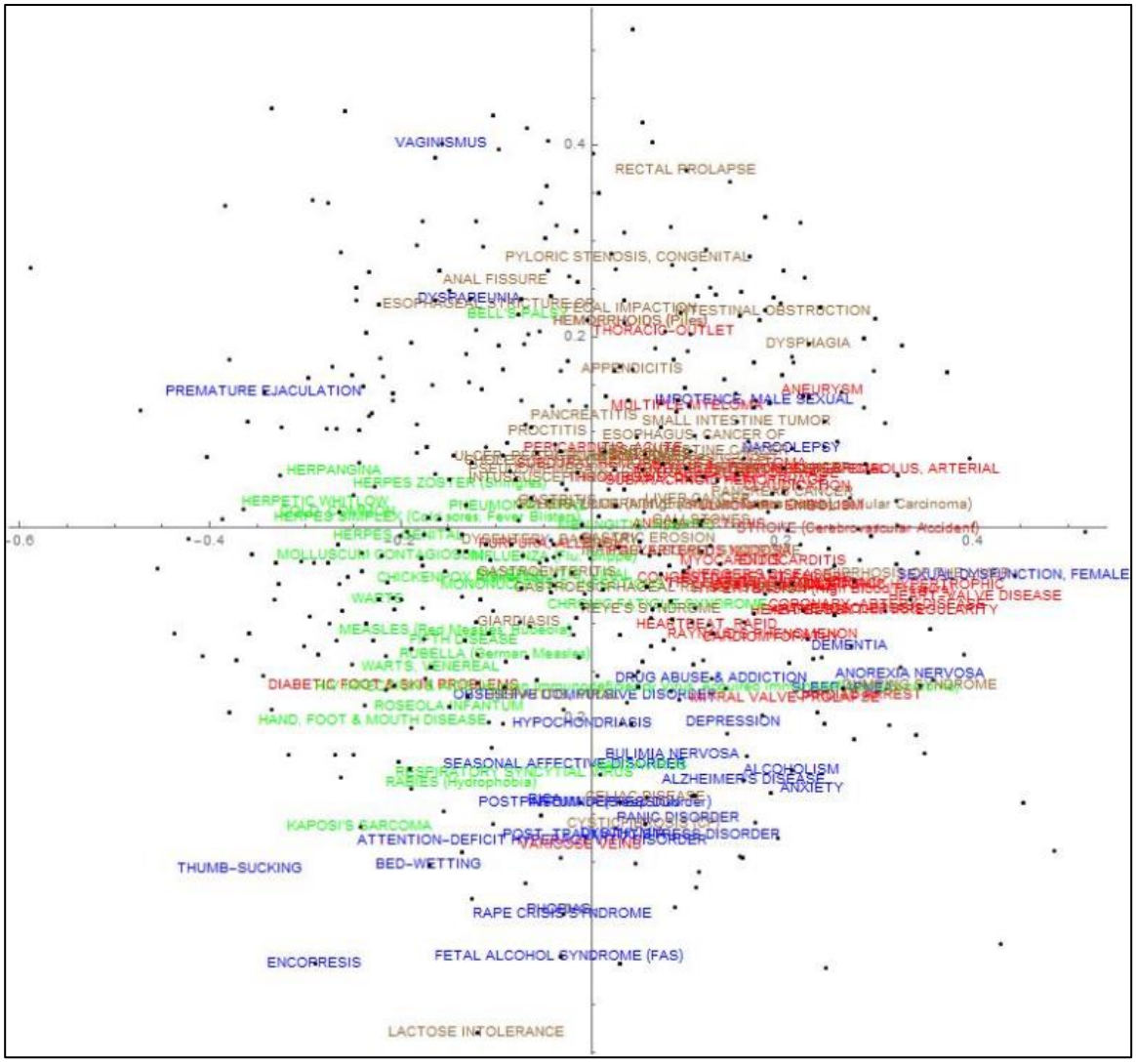

Figure 25. 2D multidimensionally scaled graph of Standard HiMean semantic space in base rate control condition.

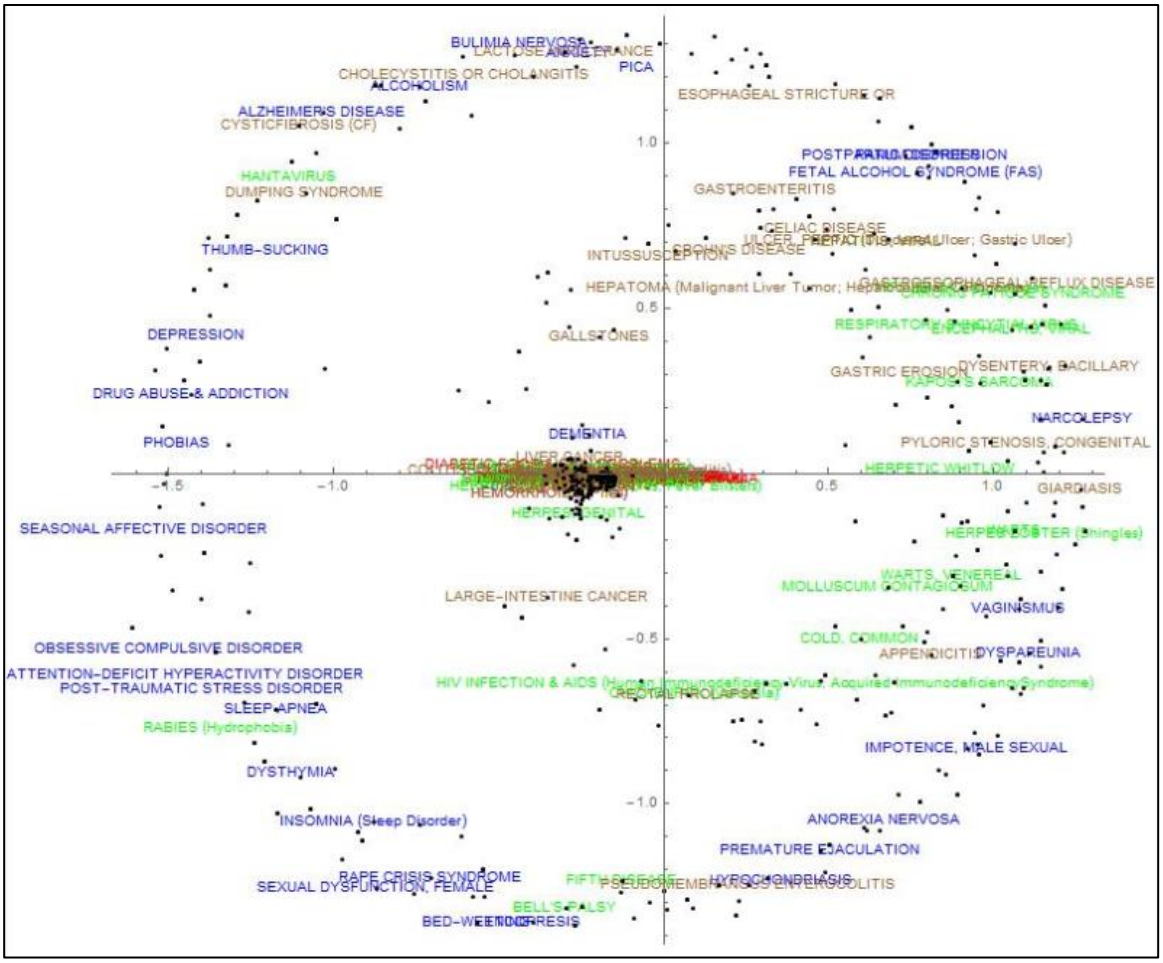

Figure 26. 2D multidimensionally scaled graph of LSA semantic space in cardiovascular base rate 10 condition.

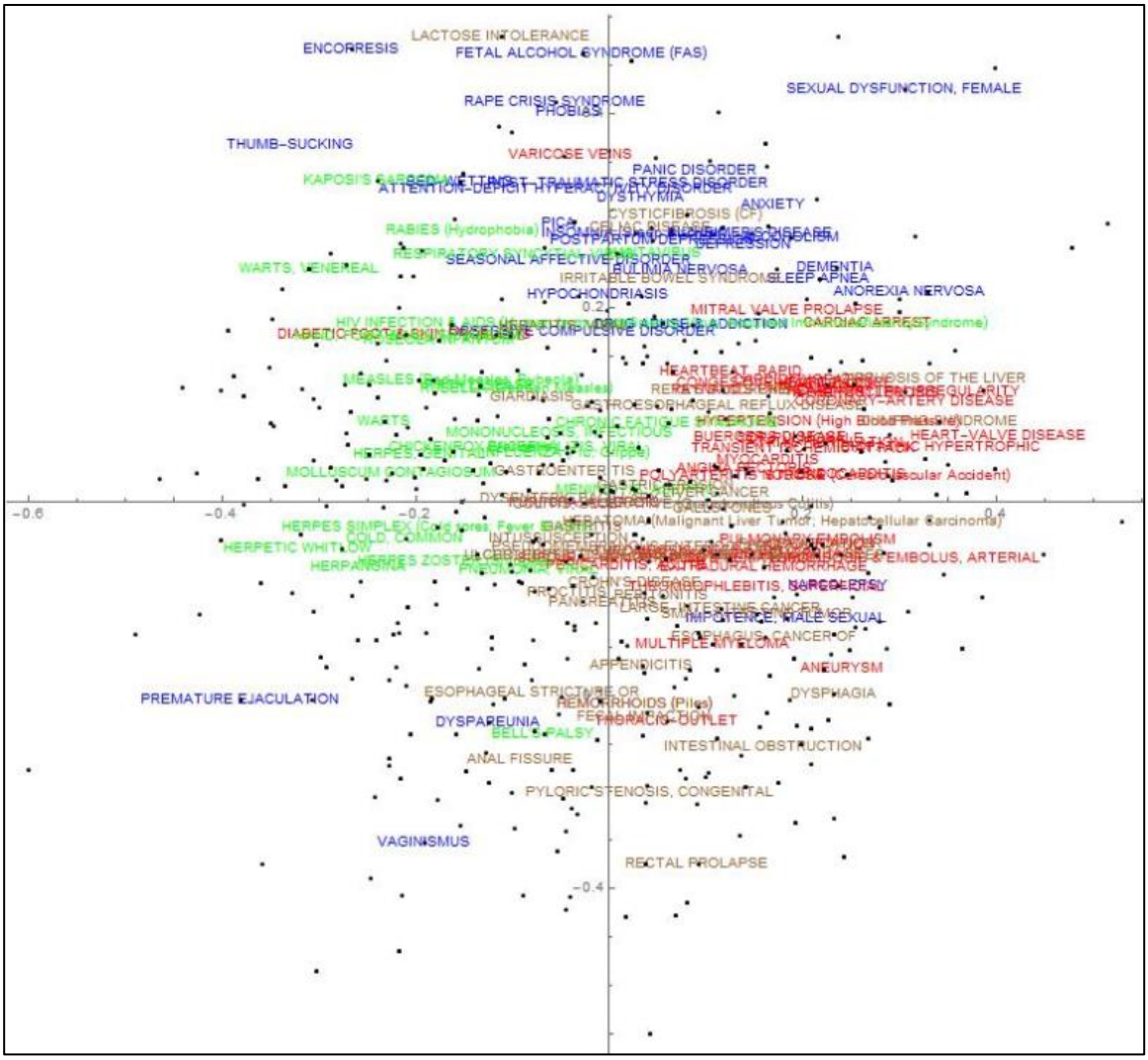

Figure 27. 2D multidimensionally scaled graph of the Ideal Observer HiMean semantic space in cardiovascular base rate 10 condition.

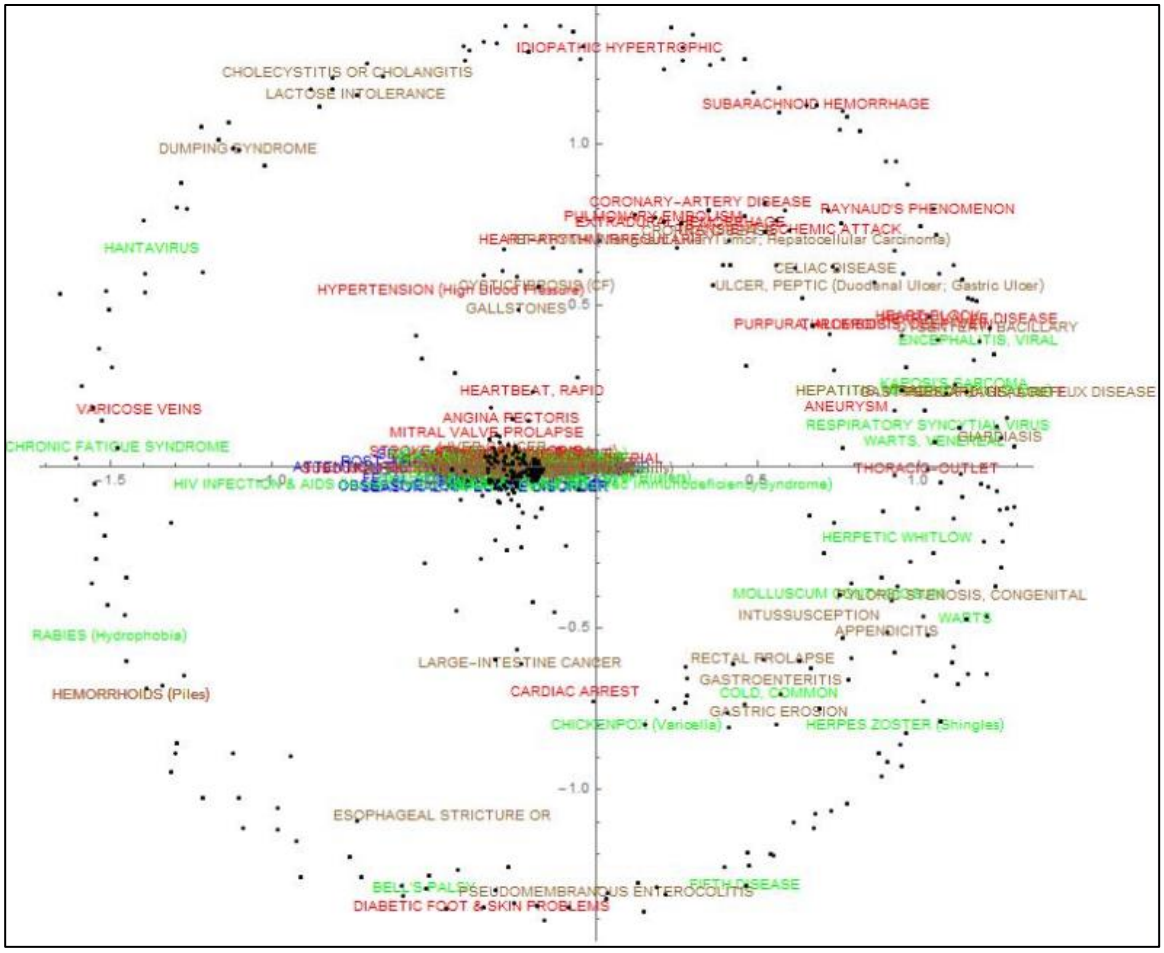

Figure 28. 2D multidimensionally scaled graph of LSA semantic space in psychological base rate 10 condition.

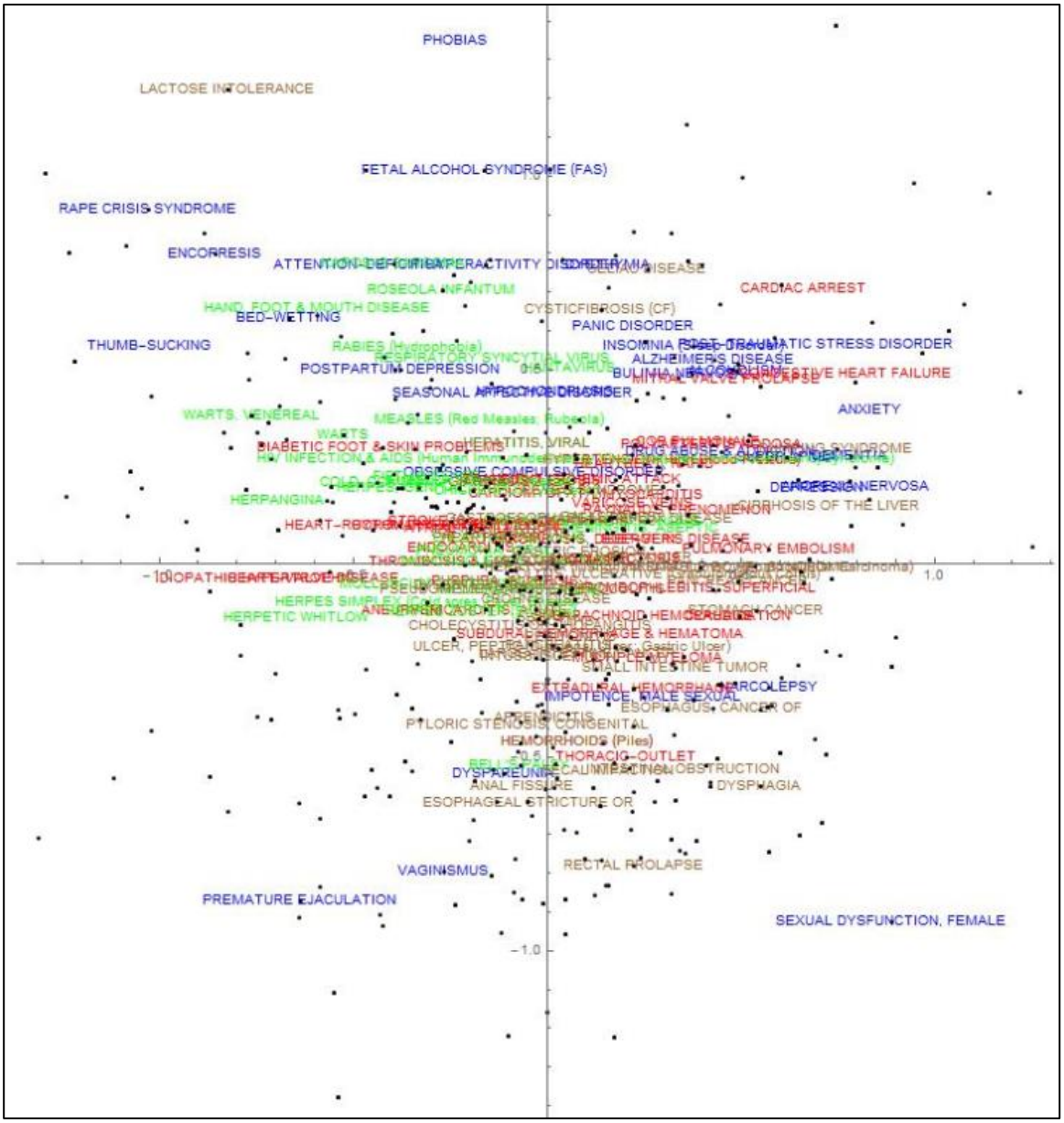

Figure 29. 2D multidimensionally scaled graph of the Ideal Observer HiMean semantic space in psychological base rate 10 condition.

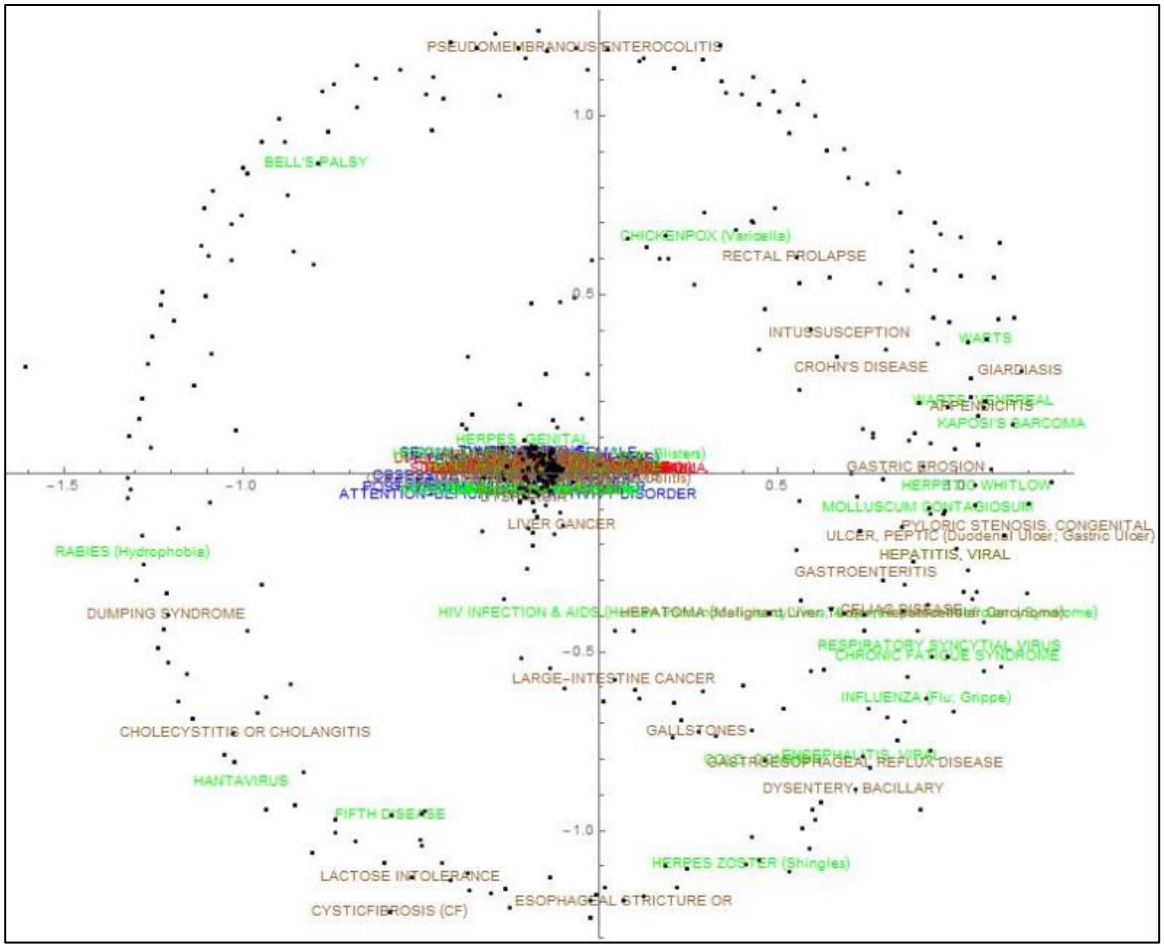

Figure 30. 2D multidimensionally scaled graph of LSA semantic space in cardiovascular and psychological base rate 10 condition.

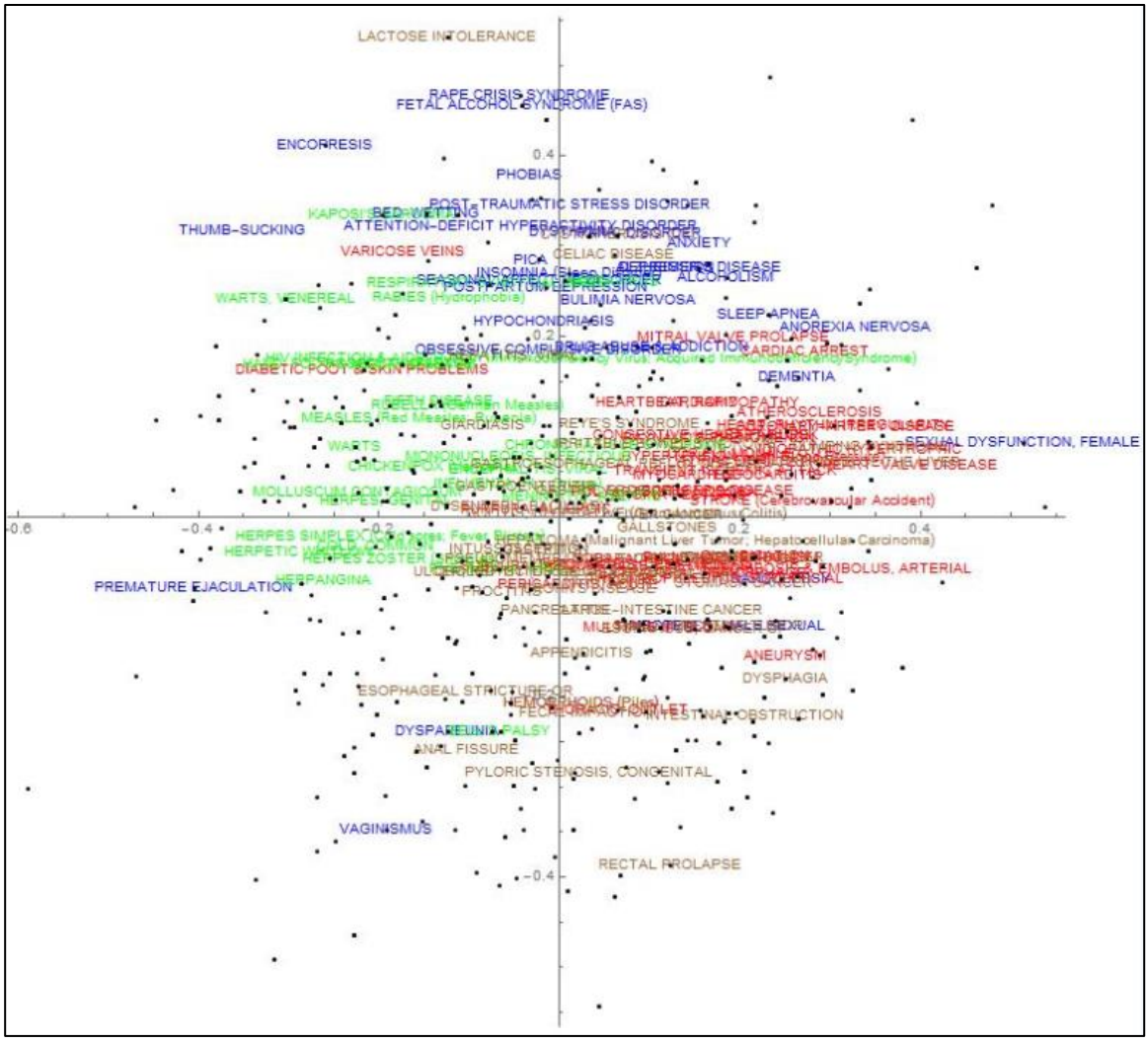

Figure 31. 2D multidimensionally scaled graph of the Ideal Observer HiMean semantic space in cardiovascular and psychological base rate 10 condition.

Appendix C: Disease Clusters

Disease categories used and number of disease per cluster.

Category Name	# Diseases in Cluster	Cluster
Psychiatry and Psychology	31	1
Immune System Diseases	29	1
Musculoskeletal Diseases	32	1
Nutritional and Metabolic Diseases	29	1
Virus Diseases	28	1
Wounds and Injuries	29	1
Bacterial Infections and Mycoses	68	2
Female Urogenital and Pregnancy Diseases	62	2
Nervous System Diseases	59	2
Pathological Conditions, Signs, and Symptoms	73	2
Skin and Connective Tissue Diseases	78	2
Cardiovascular Diseases	38	3
Digestive System Diseases	40	3
Male Urogenital Diseases	37	3
Neoplasms	50	3
Respiratory Tract Diseases	40	3
Chemically Induced Disorders	10	4
Congenital Hereditary & Neonatal Diseases & Abnormalities	16	4
Parasitic Diseases	11	4
Stomatognathic Diseases	15	4
Endocrine System Diseases	22	5
Eye Diseases	20	5
Hemic and Lymphatic Diseases	23	5
Otorhinolaryngologic Diseases	22	5