ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 27 (2005) 520 – 532

Mining the posterior cingulate: Segregation between memory and pain components

Finn Årup Nielsen,[a,b,*] Daniela Balslev,[a,c] and Lars Kai Hansen[b]

[a]*Department of Neurology, The Neuroscience Centre, Rigshospitalet, Building 9201, Neurobiology Research Unit, Copenhagen University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark*
[b]*Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark*
[c]*Danish Research Centre for Magnetic Resonance, Hvidovre, Denmark*

Received 12 July 2004; revised 2 December 2004; accepted 21 April 2005
Available online 8 June 2005

We present a general method for automatic meta-analyses in neuroscience and apply it on text data from published functional imaging studies to extract main functions associated with a brain area—the posterior cingulate cortex (PCC). Abstracts from PubMed are downloaded, words extracted and converted to a bag-of-words matrix representation. The combined data are analyzed with hierarchical non-negative matrix factorization. We find that the prominent themes in the PCC corpus are episodic memory retrieval and pain. We further characterize the distribution in PCC of the Talairach coordinates available in some of the articles. This shows a tendency to functional segregation between memory and pain components where memory activations are predominantly in the caudal part and pain in the rostral part of PCC.

Keywords: Meta-analysis; PubMed; Positron emission tomography; Magnetic resonance imaging; Cingulate gyrus; Brain mapping; Memory; Pain; Alzheimer's disease; Automatic data processing; Bibliometrics

Introduction

Functional neuroimaging methods such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) map function to brain anatomy by manipulating behavior and measuring a change in regional activity. The resulting maps are used to infer the principles of brain function. The interpretability and value of these functional maps however depend critically on the specificity of the behavioral paradigm and on the sensitivity of the analytical tools for detecting the interesting signal. In any individual study, it is very difficult, if not impossible, to avoid the particular subject's in part irreproducible behavior, and it is not possible a priori to describe the spatiotemporal characteristics of the signal necessary for accurate modeling and detection. Investigating the consensus among results from a large number of brain imaging studies can overcome the problem of false positive and false negative results. Furthermore, neuroimaging studies may discover unexpected changes in activity in areas that so far have not been associated with a given behavior, thus generate new hypotheses. Unexpected activations may be genuine, showing a novel and interesting association between brain and behavior, or spurious, for instance, a false positive activation from some uninteresting behavior that the paradigm did not control for. If such changes are repeatedly encountered over a large number of different paradigms that addressed the same behavioral function, then it is probable that they are of significance and deserve further investigation.

The potential for data mining across neuroimaging studies has already been recognized and exploited to investigate the consistency of activation patterns associated with a given brain function (Cabeza and Nyberg, 2000) or to generate new hypotheses about the function of a given brain area (Bush et al., 2000; Maddock, 1999; Maguire, 2001). In these papers, the authors manually generate tables and figures that summarize results across a large number of studies in order to inspect for consensus. The expanding number of published studies, however, makes it increasingly time-consuming for the individual researcher to generate exhaustive result summaries. Here, we propose an automatic method which can extract from the neuroimaging literature the consensus about the functions of a given brain area. This method is thought to assist researchers by providing a quick summary of a large number of scientific publications. To illustrate the approach, we apply this method to extract functions that are consistently associated with the posterior cingulate cortex (PCC). The studies we include are not restricted to examine any particular behavioral function: we simultaneously examine multiple functions. To this end, we have used article abstracts indexed by PubMed under the keywords that relate to both PCC and functional imaging methods.

* Corresponding author. Department of Neurology, The Neuroscience Centre, Rigshospitalet, Building 9201, Neurobiology Research Unit, Copenhagen University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. Fax: +45 3545 6713.
E-mail address: fn@imm.dtu.dk (F.Å. Nielsen).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.04.034

Despite the increasing attention that the PCC has attracted over the years (see Fig. 1), there is no "textbook" consensus about the functions undertaken by this area. Different reviews associate this area with a variety of brain functions, e.g., evaluative functions (for spatial orientation and memory) (Vogt et al., 1992), successful episodic memory retrieval (Cabeza and Nyberg, 2000), emotion (Maddock, 1999), navigation (Maguire, 2001), visuospatial attention (Mesulam et al., 2001; Small et al., 2003), pain (Bromm, 2001), and "resting state" (Binder et al., 1999; Mazoyer et al., 2001; Raichle et al., 2001; Shulman et al., 1997).

Our method is two-stage: in the first stage, we will make unsupervised text mining in the form of clustering based on the words in abstracts of PubMed that mention our specific target area (the PCC), and we compare these results with the results reported by previous manual reviews to assess the validity of this automatic analysis. In the second stage, we use the Talairach coordinates (Talairach and Tournoux, 1988) within the clustered articles and describe the spatial distribution in terms of the cluster labels from the first stage. This is in order to test whether the functional clusters are anatomically segregated within the posterior cingulate cortex.

That is, in terms of PCC, we ask whether our machine-based methodology is able to capture themes in alignment with major reviews and whether there is functional segregation within the PCC.

Method

We download abstracts from the PubMed Web service by restricting the search to posterior cingulate area and functional neuroimaging with the following query: ("posterior cingulate" OR "posterior cingulum" OR "retrosplenial" OR "retrosplenium") AND ("magnetic resonance imaging" OR "positron emission tomography").

This query will return functional neuroimaging as well as other types of neuroimaging studies. With the present capabilities of PubMed, this is unavoidable. However, other restrictions are possible, e.g., only inclusion of human studies and exclusion of the "review" publication type.

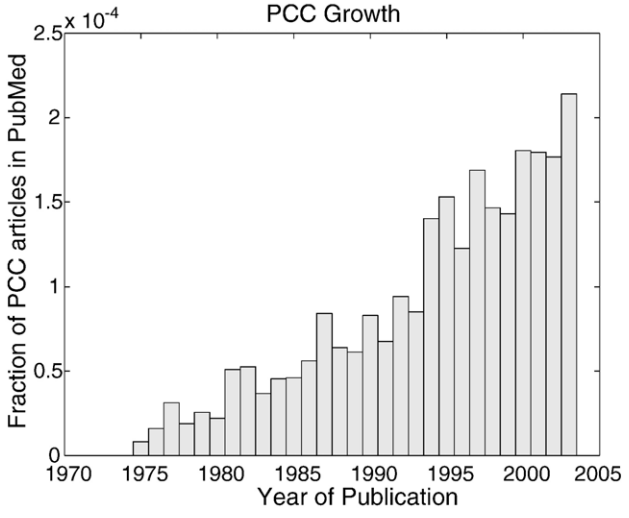

Fig. 1. Number of posterior cingulate entries in the PubMed database as a function of the year of publication. The query was "("posterior cingulate" OR "posterior cingulum" OR "retrosplenial" OR "retrosplenium")". It is normalized with the total number of entries of each year in the PubMed database.

The abstracts are converted into matrix form: $\mathbf{X}(N \times Q)$ where N corresponds to the number of abstracts and Q corresponds to the number of words. This is the so-called "vector space model" or bag-of-words representation (Salton et al., 1975). Ignoring case, we count the words in each abstract and set the element x_{nq} to the number of times the qth word occurs in the nth abstract. When the matrix contains the raw frequency, it means that abstracts which contain many occurrences of the same word will affect the subsequent analysis more than an abstract where the word occurs just once. This scheme is not necessarily optimal, but it is simple, and it is likely that, if a word is mentioned many times in an abstract, it is because the word is important. Words that only occur in a single abstract are discarded from the matrix, and certain stop words are eliminated, resulting in a matrix with fewer columns. The stop words are a compound list consisting of ordinary English words such as "the", "of" augmented with medical stop words used in MEDLINE/PubMed (http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords) and by a large manually constructed list for elimination of words that does not directly describe cognitive functions, e.g., words for brain anatomy are eliminated.

A number of methods for discovering latent classes ("clusters" or "components") in texts have been described, e.g., spherical K-means and singular value decomposition (SVD) (Dhillon and Modha, 2001), simple counting (Goldman et al., 1999), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), hierarchical clustering (Gaussier et al., 2002), non-negative matrix factorization (NMF) (Lee and Seung, 1999), Generalizable Gaussian Mixture (GGM) model (Hansen et al., 2000), and independent component analysis (ICA) (Isbell and Viola, 1999; Kolenda et al., 2000). Some of these have been applied for data mining medical literature (Dobrokhotov et al., 2003; Goldman et al., 1999), and, for example, the XplorMed Internet-based tool enables interactive exploration of noun relatedness in PubMed abstracts producing simple two-word classes (Perez-Iratxeta et al., 2001). In our case, we will use the NMF with an optimization algorithm for the Euclidean distance cost function (Lee and Seung, 2001), where a matrix $\mathbf{X}(N \times Q)$ is factorized into two non-negative matrices $\mathbf{W}(N \times K)$ and $\mathbf{H}(K \times P)$.

$$\mathbf{X} = \mathbf{WH} + \mathbf{E}, \tag{1}$$

and the cost function (reconstruction error) E is defined as

$$E = \mathrm{trace}\left(\mathbf{EE}^{\mathrm{T}}\right) = \mathrm{trace}\left(\mathbf{E}^{\mathrm{T}}\mathbf{E}\right) = \|\mathbf{E}\|_F^2, \tag{2}$$

This cost function is minimized for a fixed number of latent classes K. Each of the column vectors in $\mathbf{W}$ corresponds to a latent class with loadings for each abstract, and each of the rows in $\mathbf{H}$ contains a latent class with loadings over words. The algorithm is dependent on the initialization of $\mathbf{W}$ and $\mathbf{H}$. To avoid unfavorable local minima, we run the algorithm multiple times and choose the result with the lowest cost function value. In practice, we run the algorithm 3 times. To be reasonably sure that the global minima is found, many more runs would be needed and that will be time-consuming. NMF has the advantage that each vector (in either $\mathbf{W}$ or $\mathbf{H}$) corresponds to one latent class. In SVD and (ordinary) ICA, there might be two classes in each vector since the algorithms implicitly assume a symmetric distribution around zero, for an