



# MORPHOSAURUS in ImageCLEFmed 2006: The effect of subwords on biomedical IR

---

Philipp Daumke, Jan Paetzold, Kornél Markó

Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz...



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed> Search Print

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed  
Overview  
Help | FAQ  
Tutorial  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals  
Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
LinkOut  
My NCBI  
(Cubby)

Related  
Resources  
Order  
Documents  
NLM Catalog  
NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**Interaction between hypertension, apoE, and cerebral white matter lesions.**

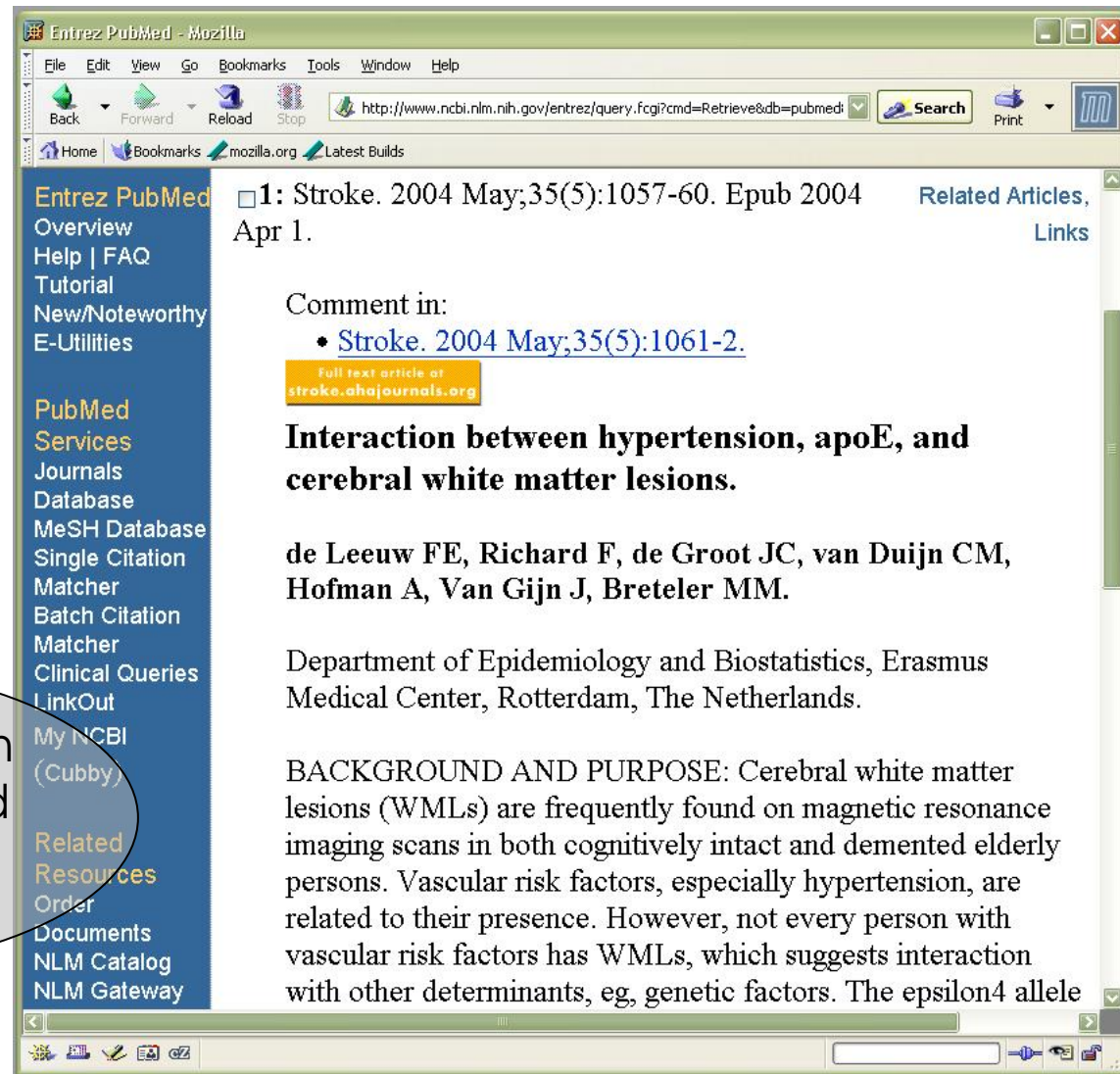
de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz...

“Correlation of high  
blood pressure and  
lesion of the white  
substance”



The screenshot shows a Mozilla browser window titled "Entrez PubMed - Mozilla". The address bar contains the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed>. The search results page displays a list of search results, with the first result selected:

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**Interaction between hypertension, apoE, and cerebral white matter lesions.**

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

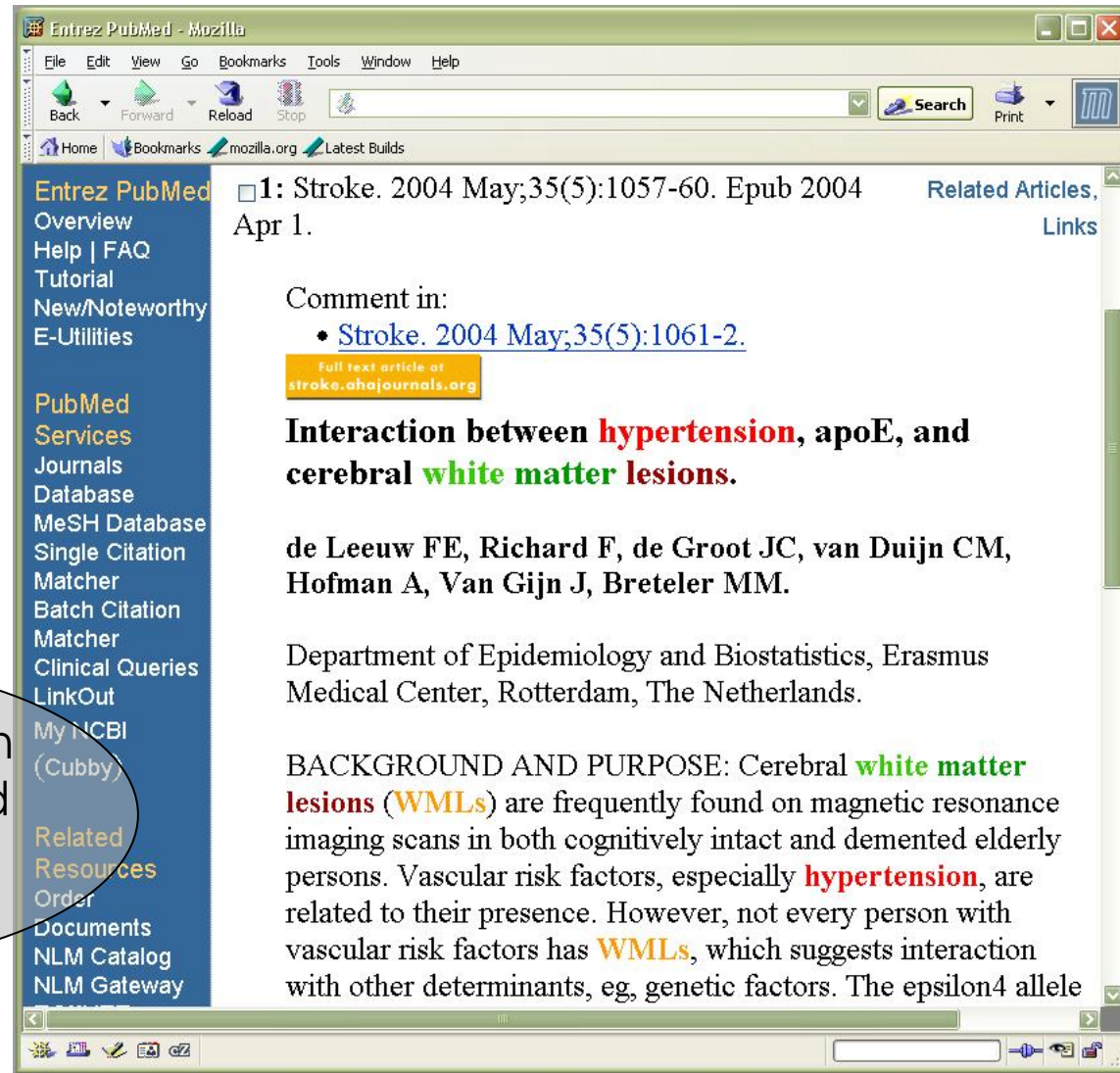
Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele



Korrelation von Hypertonie und Läsion der Weißen Substanz...

“Correlation of high blood pressure and lesion of the **white** substance”



The screenshot shows a Mozilla browser window with the address bar set to 'Entrez PubMed'. The page content includes a search result for a stroke article from May 2004. A sidebar on the left lists various PubMed services and resources. The main text area contains the article title, authors, and a brief background paragraph.

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop Search Print

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed  
Overview  
Help | FAQ  
Tutorial  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals  
Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
LinkOut  
My NCBI  
(Cubby)

Related  
Resources  
Order  
Documents  
NLM Catalog  
NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](#)

**Interaction between **hypertension**, apoE, and cerebral **white matter lesions**.**

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral **white matter lesions (WMLs)** are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially **hypertension**, are related to their presence. However, not every person with vascular risk factors has **WMLs**, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

Korrelation von Hypertonie und Läsion der Weißen Substanz...

"Correlation of high blood pressure and lesion of the **white** substance"

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed  
Overview  
Help | FAQ  
Tutorial  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals  
Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
LinkOut  
My NCBI  
(Cubby)

Related  
Resources  
Order  
Documents  
NLM Catalog  
NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 15. [Related Articles, Links](#)

Comment in:  
• [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text available at [stroke.ahajournals.org](#)

**Interaction between hypertension, apoE, and cerebral white matter lesions.**

de Leeuw FE, Richard IH, de Groot JC, van Duijn CM, Hofman A, Van Gijn JB, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral **white matter lesions (WMLs)** are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially **hypertension**, are associated to their presence. However, not every person with vascular risk factors has **WMLs**, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

- **Inflection:** leukocytes, appendix – appendices
- **Derivation:** leukocyte → leukocytic
- **Composition:** *para /sympath /ectomy,*  
*Magen /schleim /haut /entzünd /ung*
- **Acronyms, Abbreviations:** AIDS(en,de) ↔ SIDA(fr), SARS, OECD
- **Proper Names:** Aspirin ↔ ASS, ...
- **Orthographic Variants:**
  - Kolonkarzinom, Colocarzinom, Colocarzinom
  - Oesophagus, Esophagus
- (Multi Word) **Synonyms:**
  - Hypertension ↔ High Blood Pressure
- **Homonyms/Polysemes:** Cold → Common Cold, Low Temperature, Chronic Obstructive Lung Disease

**Biomedical language** is characterized by:

- Diverse user groups → **diverse documents**
  - **Telegraphic** style in Electronic Patient Records („58 yo patient c/o pain, o/e NAD“) vs.  
**flowery** language of pathological reports („size of a ripe peach“)
  - **Laymen** terminology in health portals („inflammation of the gall bladder“) vs.  
**Professional** terminology in scientific papers („cholecystitis“)
- **Neoclassical compounds** (mix of latin and greek roots)  
(„esophagogastroduodenoscopy“)
- **Medical Neologisms / Collocations / Phraseologisms**  
(„cardiac muscle bridges“)

**Subwords** as smallest units of meaning, whose meaning cannot be derived from smaller meaningful parts:

- Stems: *stomach, gastr, diaphys, vitamin c*
- Prefixes: *anti-, bi-, hyper-*
- Suffixes: *-ary, -ion, -itis*
- Infixes: *-o-, -s-*

**Equivalence classes** contain synonymous subwords and their translations in a thesaurus:

#derma = { **derm**, **cutis**, **skin**, **haut**, **kutis**, **pele**, **cutis**, **piel**, ... }  
#inflamm = { **inflam**, **-itic**, **-itis**, **entzuend**, **-itis**, **-itisch**, **inflam**, **flog**,  
**inflam**, **flog**, **-iolitis**, ... }

**Two lexical relations** between equivalence classes:

- expandsTo: #hypertens → #high #blood #pressure
- senseOf: #head → {#chief;#caput}



### *Subword Lexicon:*



gastr  
stomach  
Magen

ventric  
chamber

hepat,hepar  
liver  
leber

-itis, inflamm,  
entzünd

neph-  
ren  
kidney  
niere

#GASTR

#CHAMBER

#HEPAR

#INFLAMM

#NEPHR

### *Subword Thesaurus:*

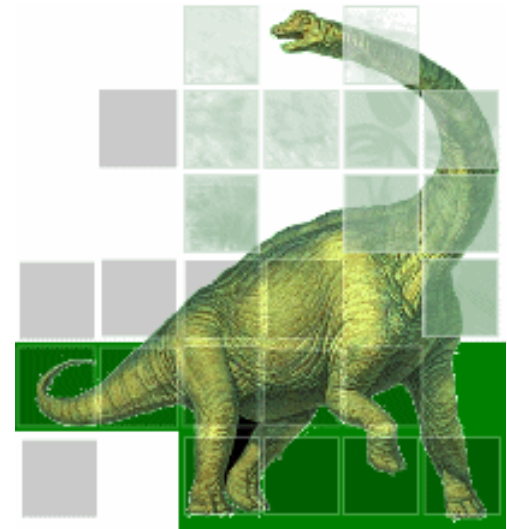
Grouping of synonymous subwords

### Lexicon statistics

- six languages
- 100,000 subwords  
25,000 (en,de), 15,000 (pt)...
- 20,000 eqClasses
- 500 exandsTo relations
- 1500 senseOf relations

- **Subword-Lexicon:**
  - Organizes subwords in several languages  
(English, German, Portuguese, Spanish, French, Swedish)
- **Subword-Thesaurus:**
  - Contains equivalence classes, groups synonymous subwords (within and between languages)
- **Subword-Segmenter:**
  - Extraction of Subwords and Assignment of *Equivalence Classes*

- **Subword-Lexicon:**
  - Organizes subwords in several languages  
(English, German, Portuguese, Spanish, French, Swedish)
- **Subword-Thesaurus:**
  - Contains equivalence classes, groups synonymous subwords (within and between languages)
- **Subword-Segmenter:**
  - Extraction of Subwords and Assignment of *Equivalence Classes*



**MORPHOSAURUS**  
([www.morphosaurus.net](http://www.morphosaurus.net))

# MORPHOSAURUS

## Example

---

High TSH values suggest the diagnosis of primary hypothyroidism ...

Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

**Original**



High TSH values suggest the diagnosis of primary hypothyroidism ...

Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

**Original**

**Orthographic  
Normalisation**



*Orthographic  
Rules*

high tsh values suggest the diagnosis of primary hypothyroidism ...

erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

High TSH values suggest the diagnosis of primary hypothyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

**Original**

**Orthographic  
Normalisation**



**Orthographic  
Rules**

high tsh values suggest the diagnosis of primary hypothyroidism ...
erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...



**Segmenter  
Subword Lexicon**

high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en schilddruese n unter funktion

# MORPHOSAURUS

# Example

High TSH values suggest the diagnosis of primary hypothyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

## Original

## Interlingua

#up	tsh	#value	#suggest
#diagnost		#primar	#hypo
#thyre			
#up	tsh	#value	#permit
#diagnost		#primar	#thyre
#hypo	#function		

**Orthographic Normalisation**



*Orthographic Rules*

high tsh values suggest the diagnosis of primary hypothyroidism ...
erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...



**Segmenter Subword Lexicon**

**Semantic Normalisation**



*Subword Thesaurus*

high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en schilddrues en unter funktion

# MORPHOSAURUS

# Example

High TSH values suggest the diagnosis of primary hypothyroidism ...

Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

## Original

## Interlingua

#up	tsh	#value	#suggest
#diagnost		#primar	#hypo
#thyre			
#up	tsh	#value	#permit
#diagnost		#primar	#thyre
#hypo	#function		

**Orthographic Normalisation**

*Orthographic Rules*

high tsh values suggest the diagnosis of primary hypothyroidism ...

erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

**Segmenter Subword-Lexicon**

**Semantic Normalisation**

*Subword-Thesaurus*

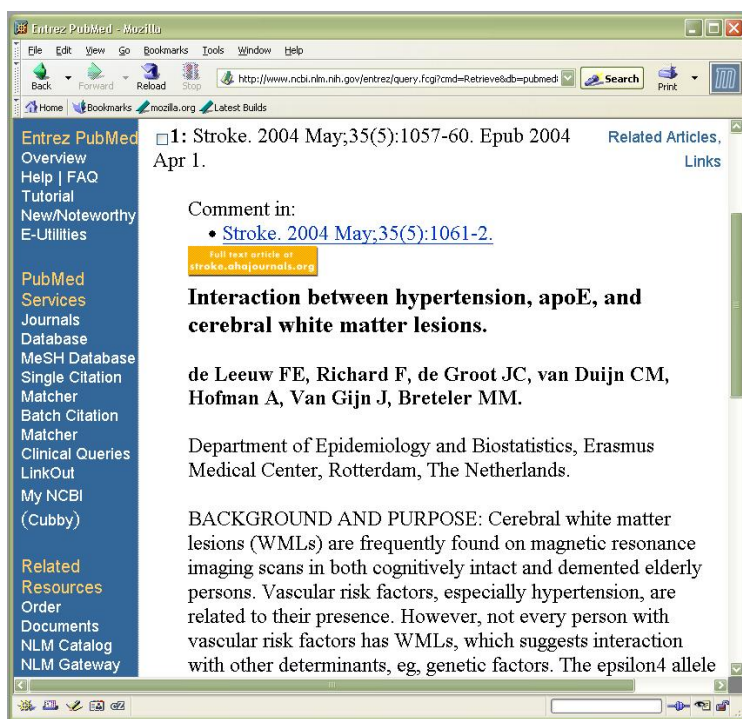
high tsh value s suggest the diagnos is of primar y hypo thyroid ism

er hoeh te tsh wert e erlaub en die diagnos e einer primaer en schilddruese n unter funktion



# MORPHOSAURUS

## MORPHOSAURUS Search



The screenshot shows a Mozilla browser window with the address bar containing the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed>. The page content includes a search result for a stroke article, a comment section, and a full-text link to a journal article.

Entrez PubMed  
Overview  
Help | FAQ  
Tutorial  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals  
Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
LinkOut  
My NCBI  
(Cubby)

Related  
Resources  
Order  
Documents  
NLM Catalog  
NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

[Full text article at stroke.ahajournals.org](#)

**Interaction between hypertension, apoE, and cerebral white matter lesions.**

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

# MORPHOSAURUS

# MORPHOSAURUS Search

Entrez PubMed - Mozilla

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. Related Articles, Links

Comment in:

- Stroke. 2004 May;35(5):1061-2.

Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**Interaction between hypertension, apoE, and cerebral white matter lesions.**

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

Entrez PubMed - Mozilla

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. Related Articles, Links

Comment in:

- Stroke. 2004 May;35(5):1061-2.

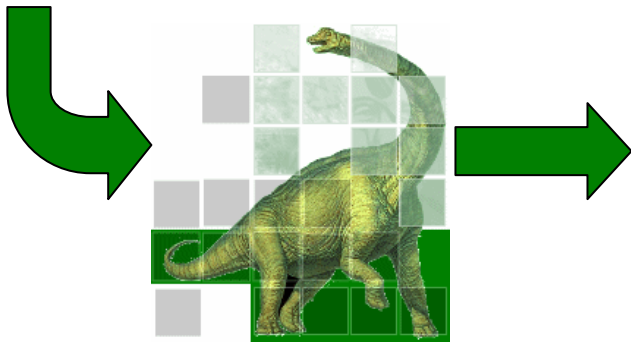
Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**#interact #hyper #tens , apoE , #cerebr #whit #matter #lesion .**

**de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .**

**#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .**

**#back #ground #purpos : #cerebr #whit #matter #lesion ( wmls ) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls , # suggest #interact #other #determin , eg ,**



Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz...



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby)

Related Resources Order Documents NLM Catalog NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](#)

**#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .**

**de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .**

**#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .**

**#back #ground #purpos : #cerebr #whit #matter #lesion ( wmls ) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls , # suggest #interact #other #determin, eg ,**



Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz...

*#correl #hyper  
#tens #lesion  
#whit #matter*

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby)

Related Resources Order Documents NLM Catalog NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. [Related Articles, Links](#)

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .**

**de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .**

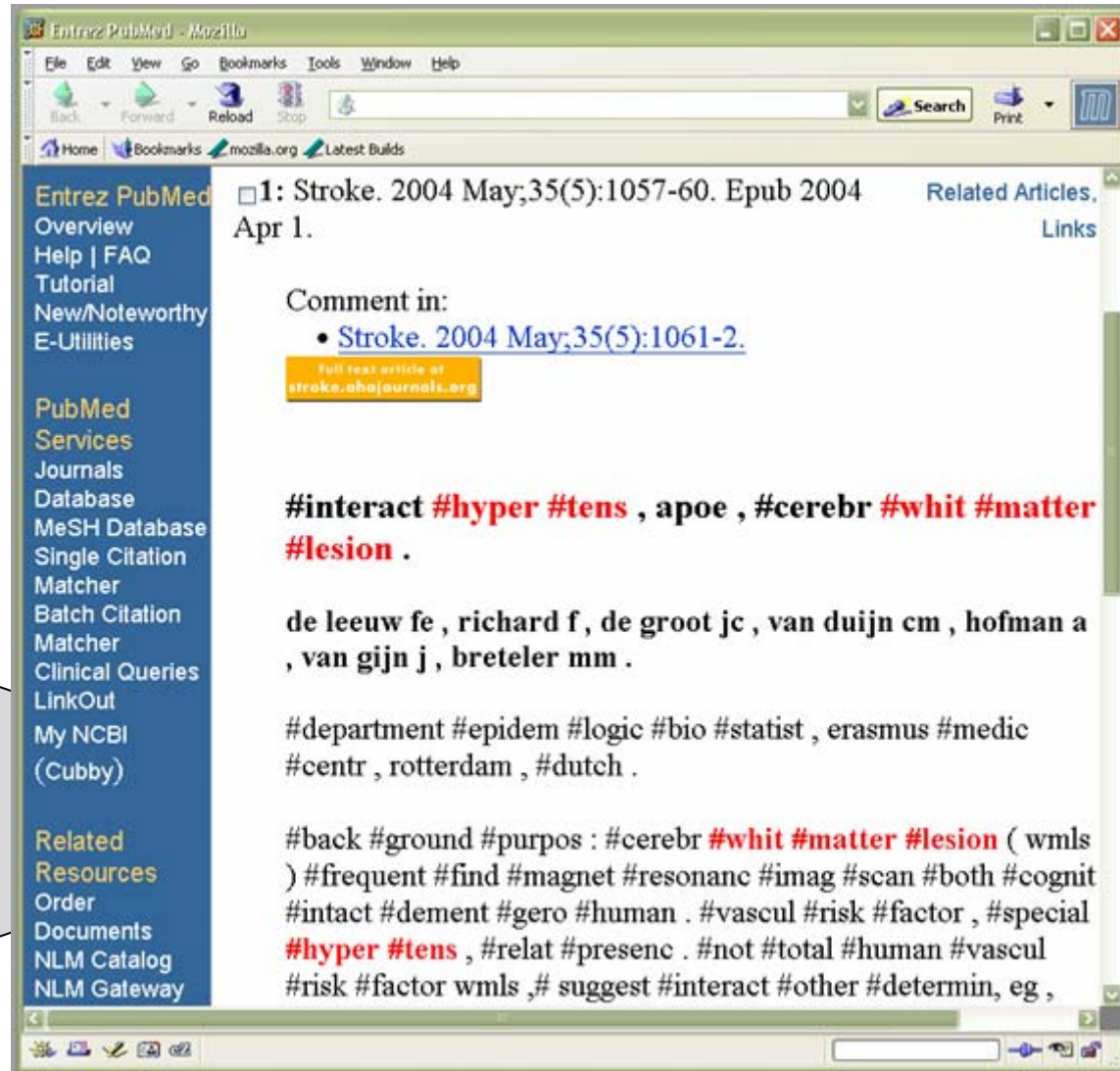
**#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .**

**#back #ground #purpos : #cerebr #whit #matter #lesion ( wmls ) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls , # suggest #interact #other #determin, eg ,**



Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz...

*#correl #hyper  
#tens #lesion  
#whit #matter*



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop Search Print

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed  
Overview  
Help | FAQ  
Tutorial  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals  
Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
LinkOut  
My NCBI  
(Cubby)

Related  
Resources  
Order  
Documents  
NLM Catalog  
NLM Gateway

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. Related Articles, Links

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)  
Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .**

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

#back #ground #purpos : #cerebr **#whit #matter #lesion** ( wmls ) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special **#hyper #tens** , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin, eg ,

- **ImageCLEFmed 2006 Corpus**
  - 4 Subsets (CasImage, MIR, PEIR, PathoPic)
  - 40,708 English, 7805 German, 1899 French documents
  - 30 Queries in all three languages
- Search Engine: **Apache Lucene**
- Image Retrieval Tool: **GIFT (the GNU Image-Finding Tool)**
- **Text Only Scenarios**

## **Mono-Lingual Test Runs** (English Subset)

- Original (Baseline)
- Stemming (Porter)
- Subwords
- Original + Stemming
- Original + Subwords

## **Cross-Lingual Test Runs**

- Original (En-En, De-De, Fr-Fr)
- Subwords (En – All)
- Subwords (De – All)
- Subwords (Fr – All)

- **Text + Image Scenarios:** combining above scenarios with GIFT results

**Text Only**

Scenario	Original (Baseline)	Stemming (Porter)		Subwords		Original + Stemming		Original + Subwords	
<b>map</b>	<b>0.1625</b>	<b>0.1482</b>	<b>91%</b>	<b>0.1297</b>	<b>80%</b>	0.1732	106%	<b>0.1792</b>	<b>110%</b>
top2avg	0.3778	0.3669	97%	0.3175	84%	0.4146	110%	<b>0.4525</b>	<b>120%</b>
P5	<b>0.4867</b>	<b>0.4667</b>	<b>96%</b>	<b>0.4867</b>	<b>100%</b>	0.5133	106%	<b>0.5933</b>	<b>122%</b>
P20	0.3600	0.4000	111%	0.3550	99%	0.4017	112%	<b>0.4500</b>	<b>125%</b>

**Findings:**

- neither Stemming nor Subword Search alone performed as good as Original Search regarding MAP
- Combination of Original and Subword Search increases precision values up to 25%

## Text Only

Scenario	Orig-All-All	Orig-En-En		Subwords-En-All		Subwords-De-All		Subwords-Fr-All	
map	0.1068	0.1625	152%	0.1366	128%	0.1439	135%	0.0734	69%
top2avg	0.2881	0.3778	131%	0.3970	138%	0.3751	130%	0.2028	70%
P5	0.3200	0.4867	152%	0.4200	131%	0.4000	125%	0.3103	97%
P20	0.2600	0.3600	138%	0.3467	133%	0.3383	130%	0.2293	88%

### Findings:

- Baseline Search on the whole document collection performs much worse than on the English Subset
- Subword Search in English and German performs clearly better than the baseline (original)
- German Subword Search performs (almost) as good as the English Subword Search

## Text + Image

- Image results provided by GIFT (the GNU Image-Finding Tool)
- Several merging algorithm were tested → All performed worse than *Text Only* results

- Mono- and Cross-Language Text Retrieval based on a sophisticated **interlingual layer**
- Can be combined with **any search engine**
- Mono-Lingual: **Combination of original and subword search remarkably boosts performance (up to 25%)**
- Cross-Lingual: **Subword approach clearly outperforms baseline, German as good as English**
- **Mean map values, good top-5(20)-precision values**
- A capable combination with Image Retrieval Tools is due



MORPHOSAURUS

---

[www.morphosaurus.net](http://www.morphosaurus.net)