DIGIT.B4 – Big Data PoC

RTD – Health papers

D03.03.Text-Mining Models

everis Spain S.L.U

Table of contents

Table of figures

1 INTRODUCTION

1.1 Context of the project

This proof of concept aims to prove how big data techniques can be applied in the research domain and to demonstrate the policy benefits that big data can bring.

Specifically, this proof of concept demonstrates the use of text mining techniques on large amounts of unstructured research papers as a means to identify trending topics in the health research field. This analysis can be used as an additional input prior to launching calls for grants.

1.2 Objective

The purpose of this document is to describe the processes carried out during the modelling phase of the CRISP-DM methodology. In this phase, segmentation algorithms are used in order to create homogeneous segments and heterogeneous segments among them.

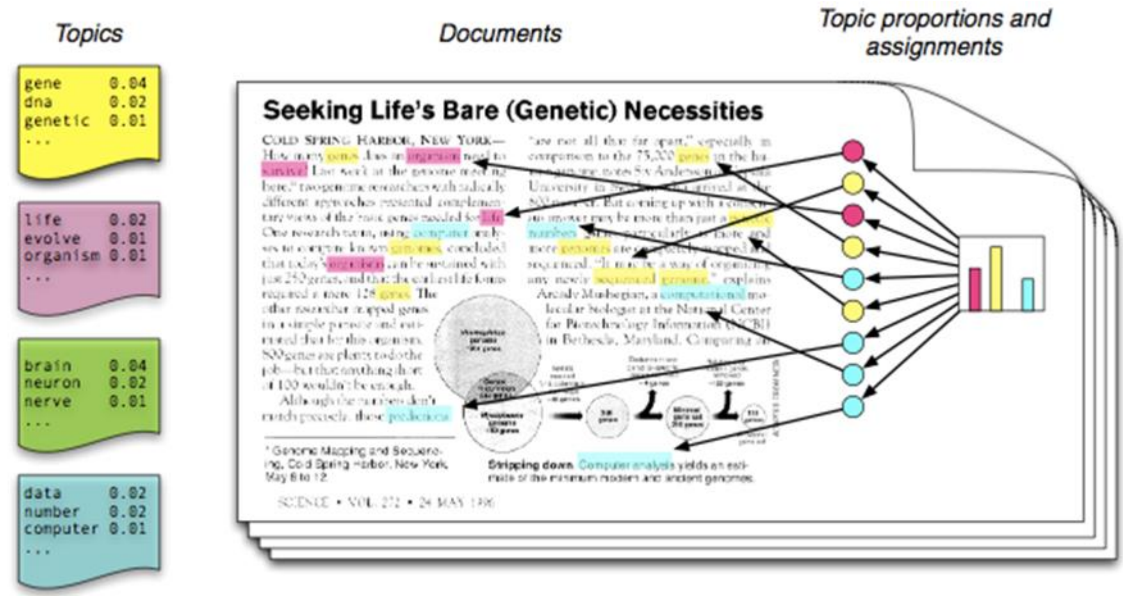
2 MODELLING

2.1 Introduction

The publications within the scope of the analysis come from two different data sources: PubMed and CORDIS.

Due to the amount of papers to be analysed and the specific language used in the text, text-mining algorithms must be applied to perform the analysis.

There are several topic-discovery algorithms based on different points of view. Algorithms called "correlated topic models" will be used for this PoC. The first step of the modelling consists of identifying the best-fitting algorithm for the purpose of the analysis.



Association algorithms can be used for simpler problems, where categories are not as well defined. Is such cases, there are common keywords in several segments and there are segments which do not have a representative sample to make the topic-discovery algorithms converge.

2.2 Topic-discovery algorithms

The original developers of LDA and CTM wrote:

> *“In technical terms, a topic model is a generative probabilistic model that uses a small number of distributions over a vocabulary to describe a document collection. When fit from data, these distributions often correspond to intuitive notions of topicality”.*

The input for both models is a TDM or a DTM (Term/Document matrix or Document/Term matrix) so it assumed that the words in a document are exchangeable and their order is not important for the document’s summary (a *bag-of-words algorithm*).

A resume of both iterative algorithms could be the execution of the following steps:

1. An initial **number of topics** that it is expected to be is fixed. In mathematical classification models, it is usually necessary to previously set the topics to detect. The fixed number could be pondered or got from a previous analysis. This number would be called N.

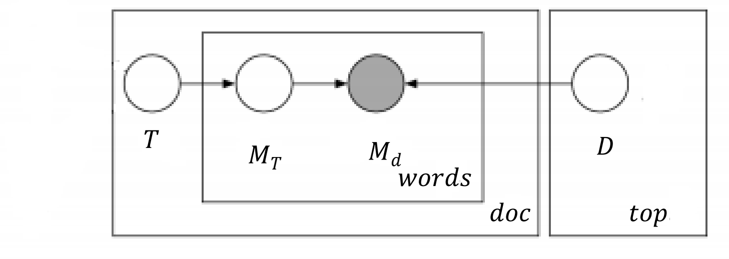
2. **Every word** is **assigned** to a temporary topic according to some function. This function will be the Dirichlet distribution (a multivariate generalisation of *Beta distribution*) in the LDA algorithm and the logistic normal distribution in the CTM one. This assignment proceeds as follows:

   a. For each topic and each document both distributions are used: $Dir(\vec{\alpha})$ and $Dir(\vec{\beta})$ with LDA, $\mathcal{N}_1(\mu,\Sigma)$ and $\mathcal{N}_2(\mu,\Sigma)$ with CTM. They would be called T and D.

   b. All the topics generated previously are distributed shaping geometrical figures with a centroid. These 'centroids' are moved according to a multinomial distribution whose parameters are T and D. They would be called M_T and M_D

3. Each topic is properly **labelled** because initially the algorithm gets numbered topics.

4. The previous steps are repeated for multiple iterations. In each iteration, the centroids move towards a convergence point. The iterations stop when the centroids do not move anymore compared to the previous iteration (the algorithm "converges"). The following graph shows the previous steps by summing up the **assignment algorithm** with the previous notation:



In order to calculate the topics probability distribution in words and documents, **some information about several parameters** must be estimated.

Due to the computational complexity of their maximum likelihood estimation, a **Gibbs sampling** has been used (a Markov chain Monte Carlo algorithm) for obtaining them within the LDA and a **variant of the expectation-maximisation** algorithm in the CTM.

2.2.1 Differences between LDA and CTM

In addition to the different distributions used in the calculation previously described, there are other important differences between them:

- LDA always converges with the Gibbs sampling and CTM does not.
- CTM sometimes achieves a better fit than LDA when topics are highly correlated.
- CTM calculates the correlation between the discover topics and LDA does not.

The CTM algorithm has been selected because when the abstracts have high correlation this model brings better results than LDA.

2.2.2 Oriented CTM model

According to one of the requirements of the PoC, the algorithm needs to provide as an output the categories of the HRCS classification model - whose names are *Blood, Cancer, Cardiovascular, Congenital, Ear, Eye, Infection, Inflammatory, Injuries, Mental Health, Metabolic, Musculoskeletal, Neurological, Oral and Gastro, Renal and Urological, Reproduction, Respiratory, Skin, Stroke* and finally *Other*.

Since the number of topics is quite large, this saturates the capacity of the algorithm, which will not be able to identify all the categories in just one execution. For this reason, an iterative CTM process is used.

The entire process is described with the following steps:

1. The content of the papers or abstracts, previously cleaned, is transformed by using a dictionary. The dictionary contains the most relevant terms, which will be used to fit the model oriented to HRCS categories. The dictionary is attached in D03.02.Dictionaries_v0.1.
2. A CTM model is built (with the entire sample in the first execution and a sub-set of documents in the rest of the iterations).
3. All topics and their terms are reviewed in order to extract the groups that are close to an HRCS category.
4. The documents assigned to accept and define categories are saved.
5. The process returns to step 2 with documents that have not been assigned to a category - until all HRCS categories are recognised or all documents have a category assigned.

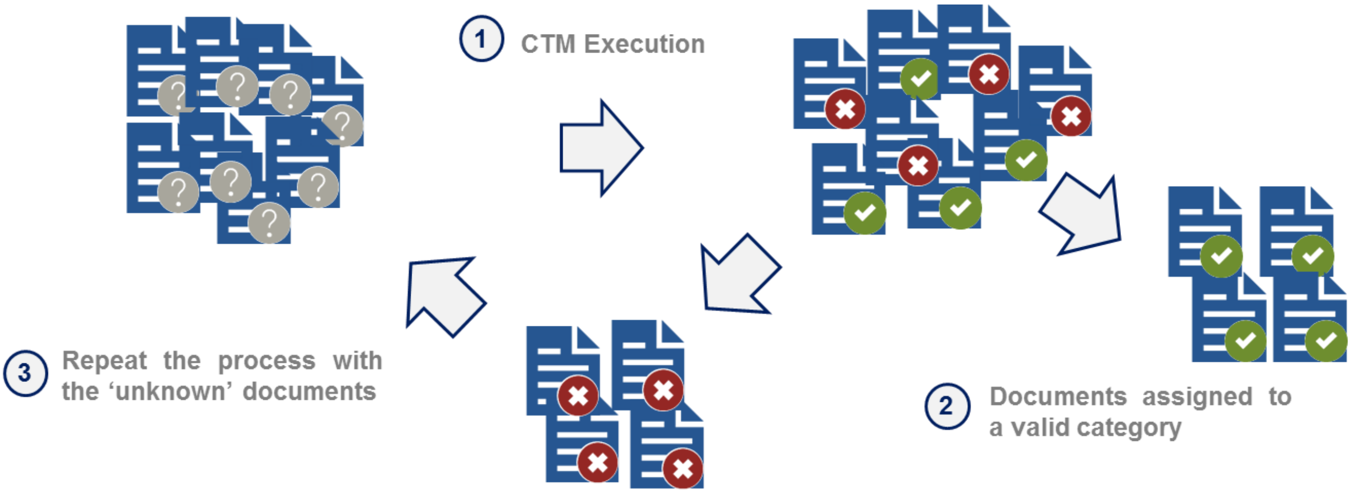

1
CTM Execution
2
Documents assigned to a valid category
3
Repeat the process with the 'unknown' documents

Figure 1 - Iterative CTM model

Each iteration allows a certain amount of papers to be classified, thus the final classification will be obtained by joining the result of all iterations.

2.3 Associative algorithms

As there is a more complex scenario (small sample for some categories, common keywords, not very specific keywords…) topic-discovery can not be used because they will not converge, so a simpler algorithm must be used.

The quality for these algorithms is determined by the work done in dictionaries. The most frequent terms are analysed and grouped into higher categories to try to increase their relevance.

After this, an array is created with as many rows as there are documents, and as many columns as there are categories. Each cell will contain a 1 if the document contains more than 3* terms associated with that category and a 0 if the document does not. This array defines which documents belong to each category (a document can belong to multiple categories).

**Distribution analyses have been made to determine the number of terms associated to a category that must appear in order to assign a 1 to the document.*

3 RESULTS

The CTM model is built with approximately 20.000 papers from the original source, which were cleaned in the previous phase (delete stopwords, punctuation, meaningless terms…). The model is subsequently applied to all of the papers (approximately 4,5 million).

Three iterations were necessary to build the oriented CTM model, explained in section 2.4. Oriented CTM model. 19 topics related to HRCS were detected, as explained in Figure 1, and the 29 topics related to OCDE were derived from the list of the most interesting keywords sent by DG RTD.

The following figures show some of the segments obtained through the CTM model in the first iteration. Each blue circle (or cluster) represents the abstracts that are classified under a topic and the terms with the highest frequency inside them.

For example, Topic 2 is the ‘*Cardiovascular’* category because many terms are related to the cardiovascular topic and Topic 4 is the ‘*Reproduction’* category for the same reason.

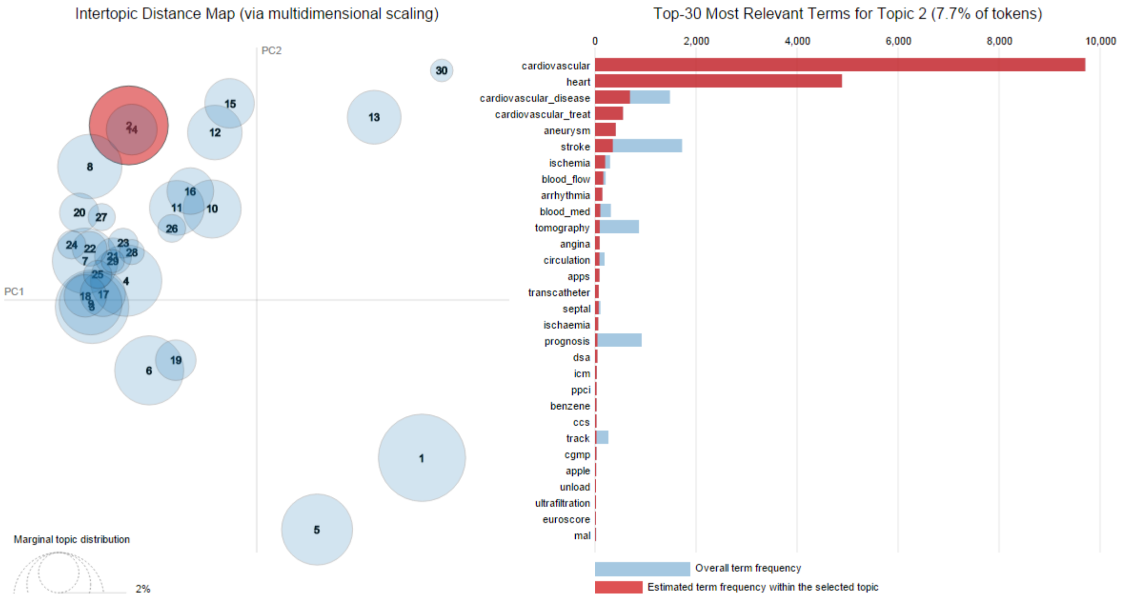


Figure 2 - CTM example: Relevant terms in Topic 2 ‘*Cardiovascular*’

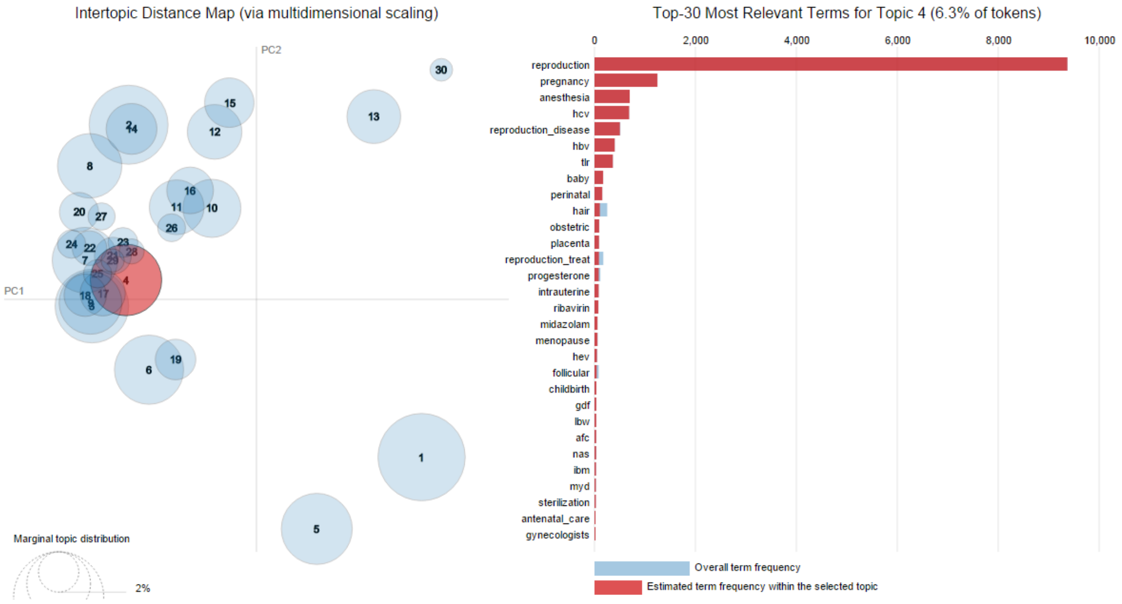


Figure 3 - CTM example: Relevant terms in Topic 4 '*Reproduction*'

3.1 Topic summary HRCS categories

Besides the 19 predefined HRCS categories, there is always a cluster with a mix of terms which are included in other clusters too. This is why this group is named '*Other*'.

The following table shows the abstracts count within each topic:

Abstracts	Topic Name
277.042	Blood
795.875	Cancer
494.361	Cardiovascular
284.318	Congenital
59.034	Ear
174.664	Eye
367.885	Infection
253.396	Inflammatory
297.008	Injuries
312.730	Mental Health
325.472	Metabolic
311.330	Musculoskeletal
314.733	Neurological
357.896	Oral and Gastro
122.696	Renal and Urological
356.336	Reproduction
181.404	Respiratory
122.838	Skin
60.635	Stroke
654.472	Other

Figure 4 – Abstracts count HRCS categories

The next subsections focus on analysing what we could find within each topic.

3.1.1 Topic '*Blood*'

The CTM model has classified 277.042 papers into '*Blood*'.

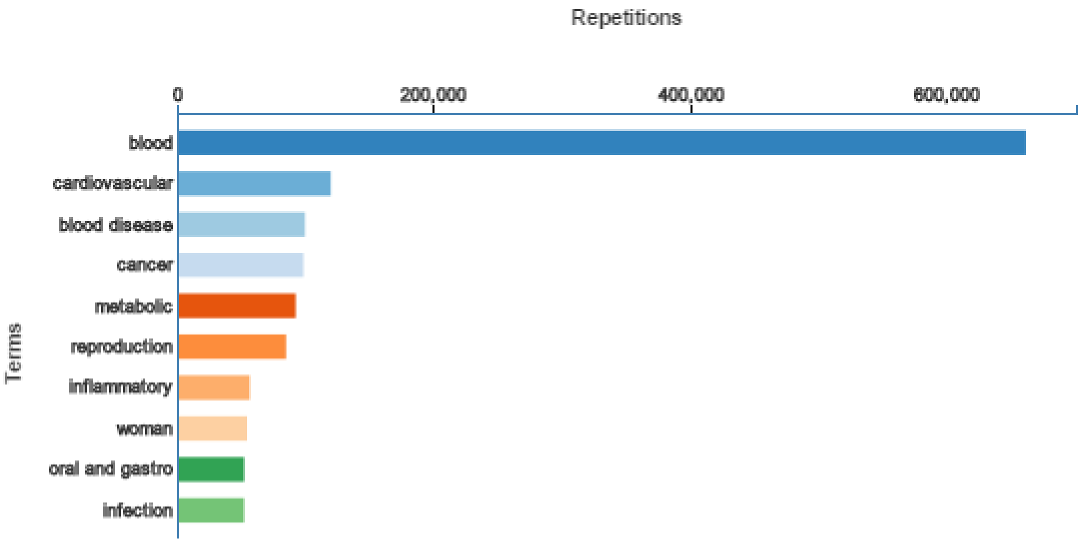


Figure 5 - Ten of the most frequent terms in '*Blood*'

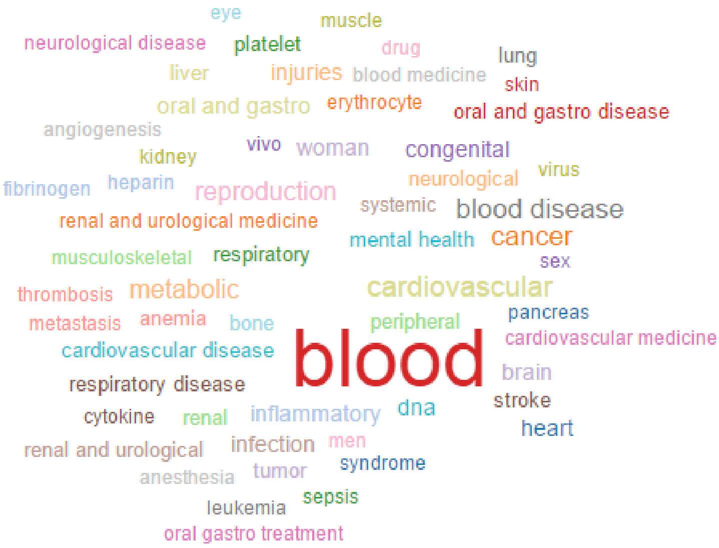


Figure 6 - '*Blood*' wordcloud

3.1.2 Topic ‘*Cancer*’

The CTM model has classified 795.875 papers into ‘*Cancer*’.

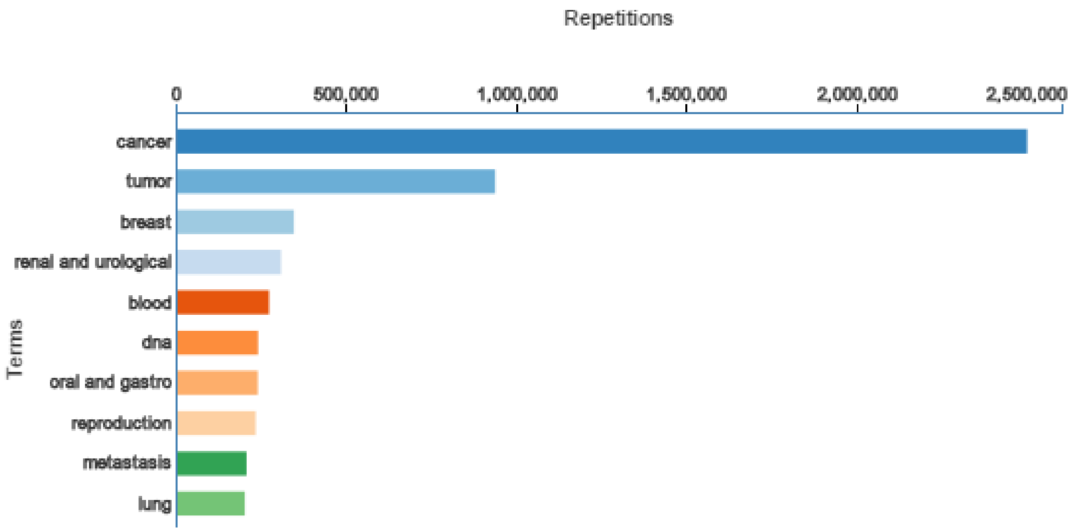


Figure 7 - Ten of the most frequent terms in *'Cancer'*

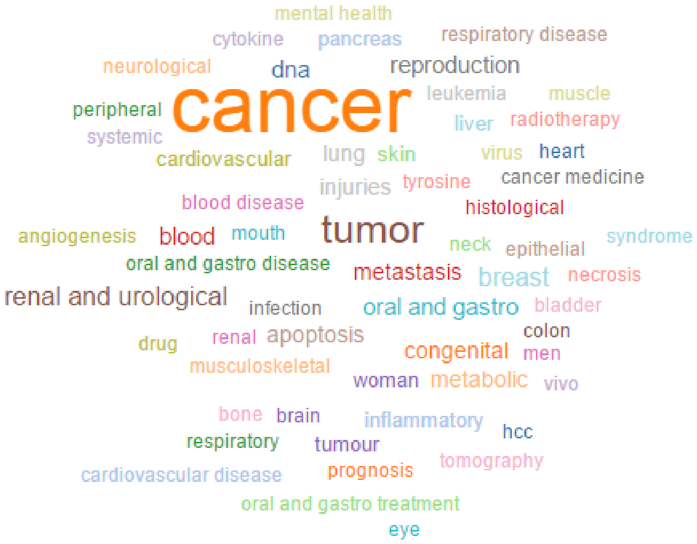


Figure 8 - *'Cancer'* wordcloud

3.1.3 Topic ‘*Cardiovascular*’

The CTM model has classified 494.361papers into ‘*Cardiovascular*’.

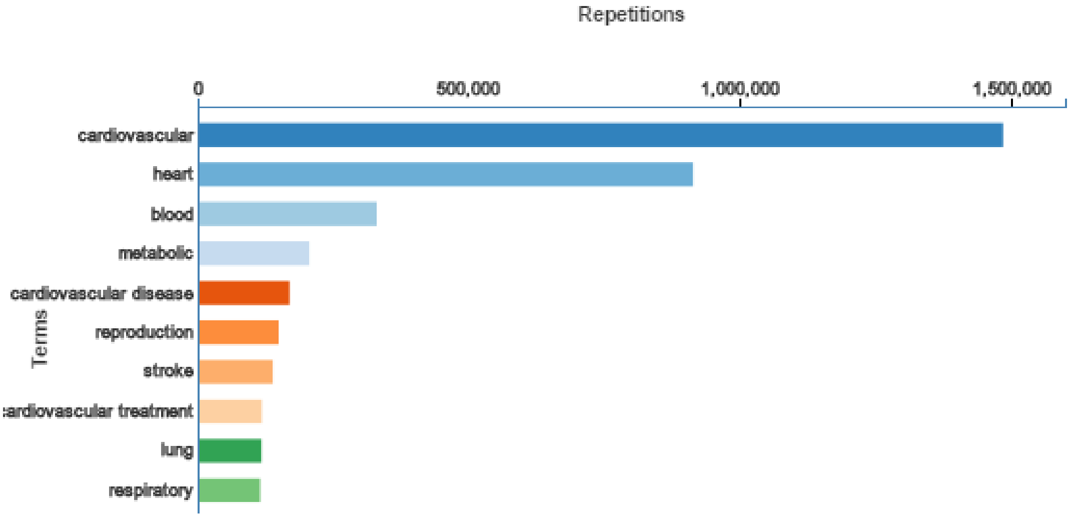


Figure 9 - Ten of the most frequent terms in '*Cardiovascular*'

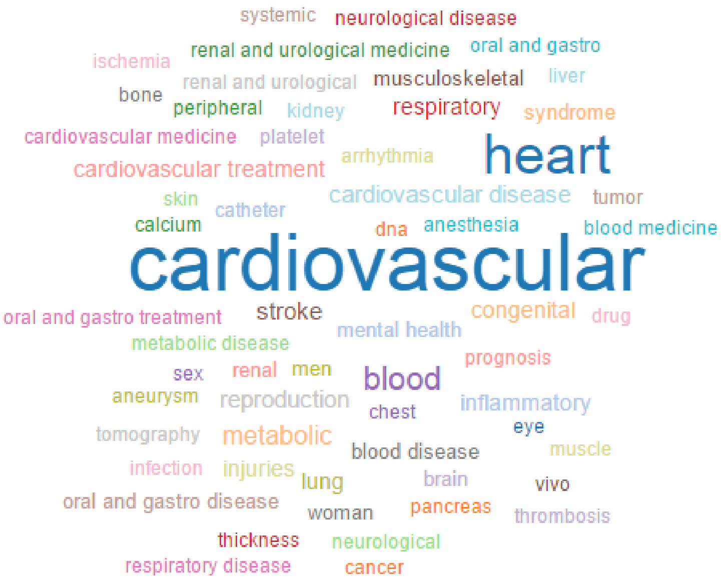


Figure 10 - '*Cardiovascular*' wordcloud

3.1.4 Topic '*Congenital*'

The CTM model has classified 284.318 papers into '*Congenital*'.

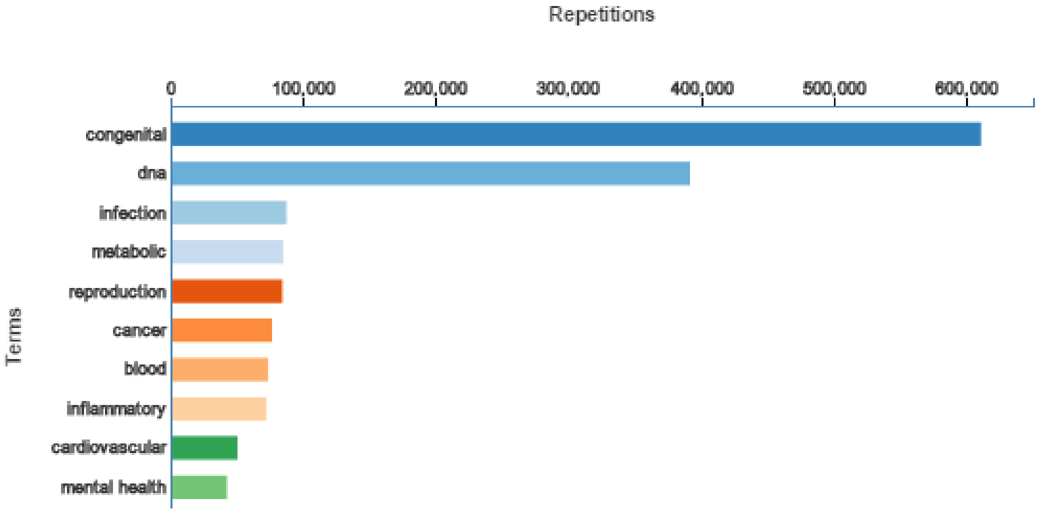


Figure 11 - Ten of the most frequent terms in '*Congenital*'

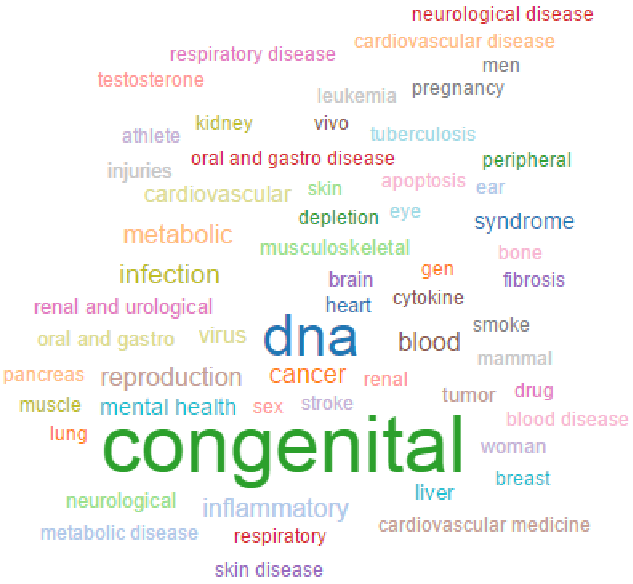


Figure 12 - '*Congenital*' wordcloud

3.1.5 Topic ‘Ear’

The CTM model has classified 59.034 papers into ‘*Ear’*.

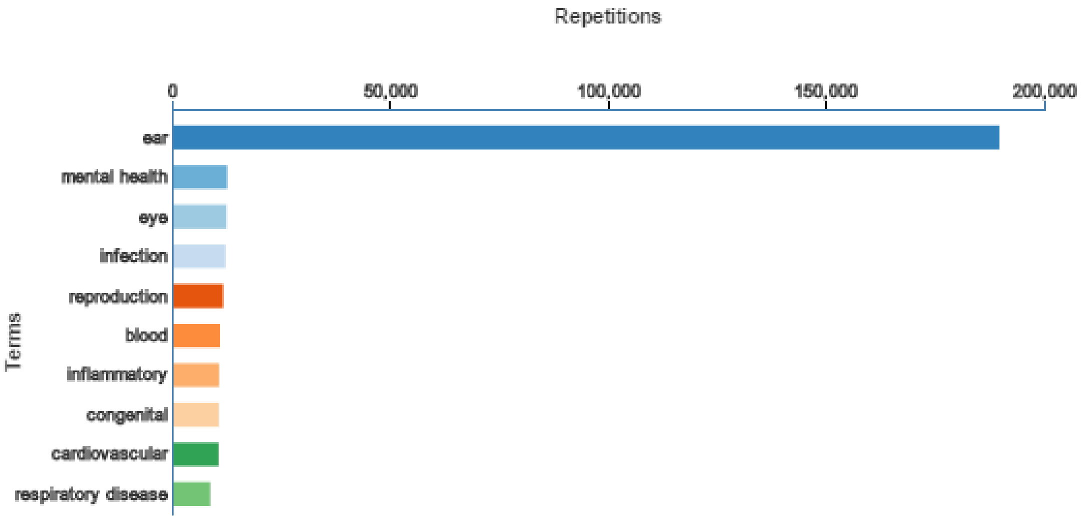


Figure 13 - Ten of the most frequent terms in *'Ear*'

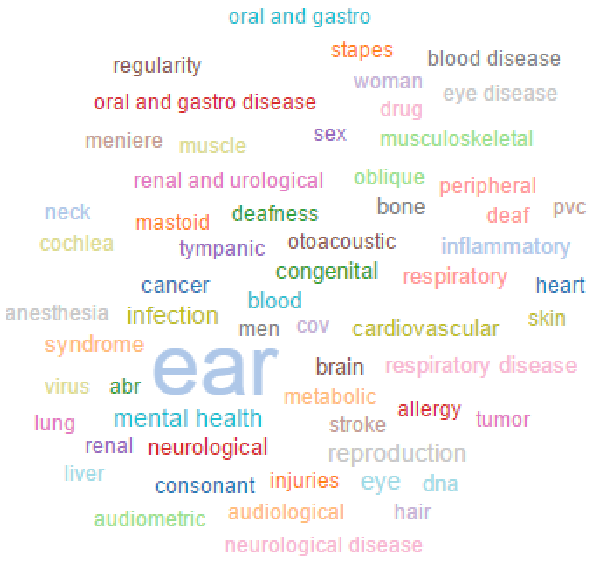


Figure 14 – ‘*Ear*' wordcloud

3.1.6 Topic '*Eye*'

The CTM model has classified 174.664 papers into '*Eye*'.

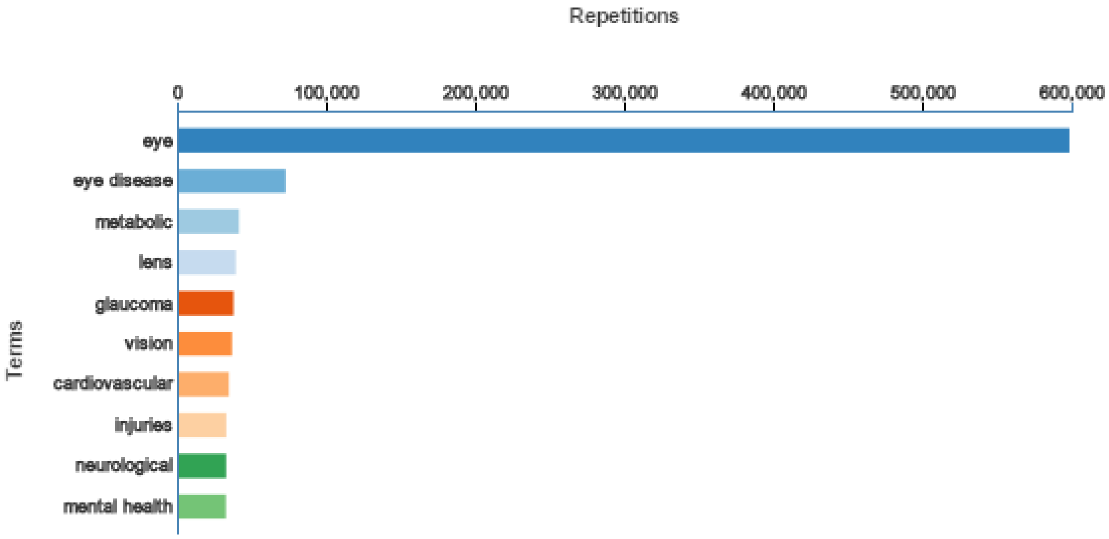


Figure 15 - Ten of the most frequent terms in '*Eye*'

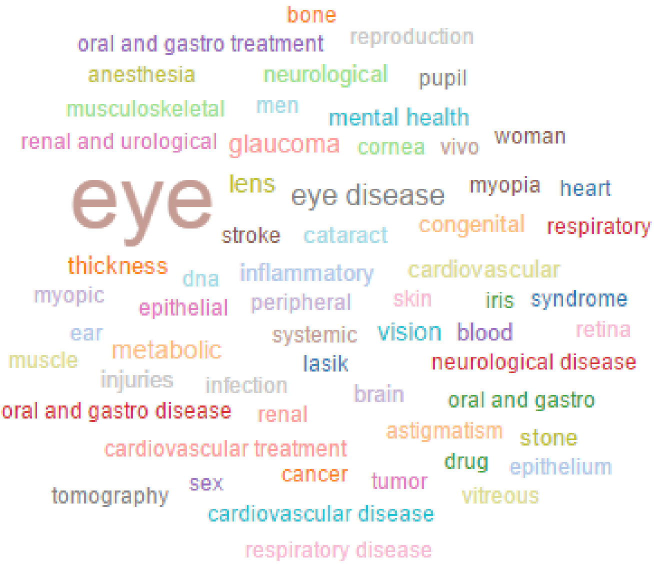


Figure 16 - '*Eye*' wordcloud

3.1.7 Topic *'Infection'*

The CTM model has classified 367.885 papers into '*Infection*'.

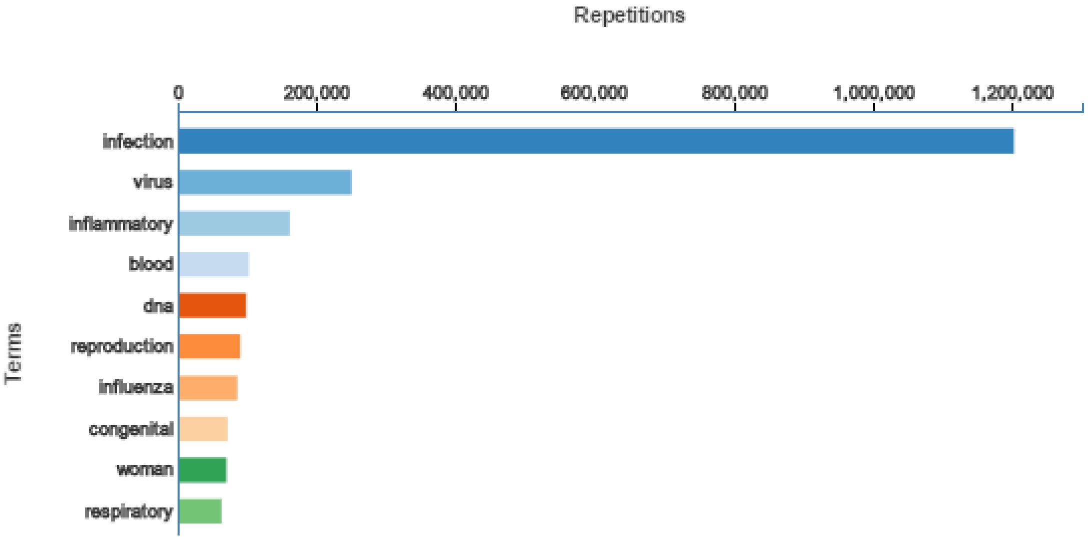


Figure 17 - Ten of the most frequent terms in *'Infection'*

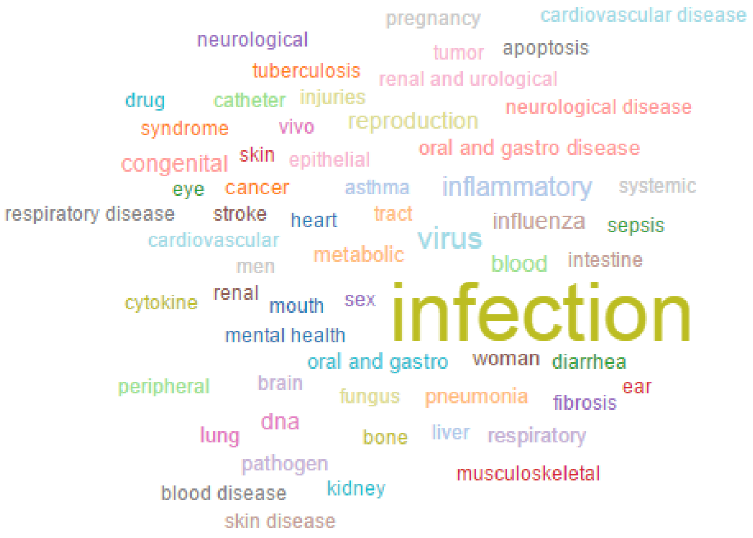


Figure 18 - *'Infection'* wordcloud

3.1.8 Topic ‘*Inflammatory*’

The CTM model has classified 253.396 papers into ‘*Inflammatory*’.

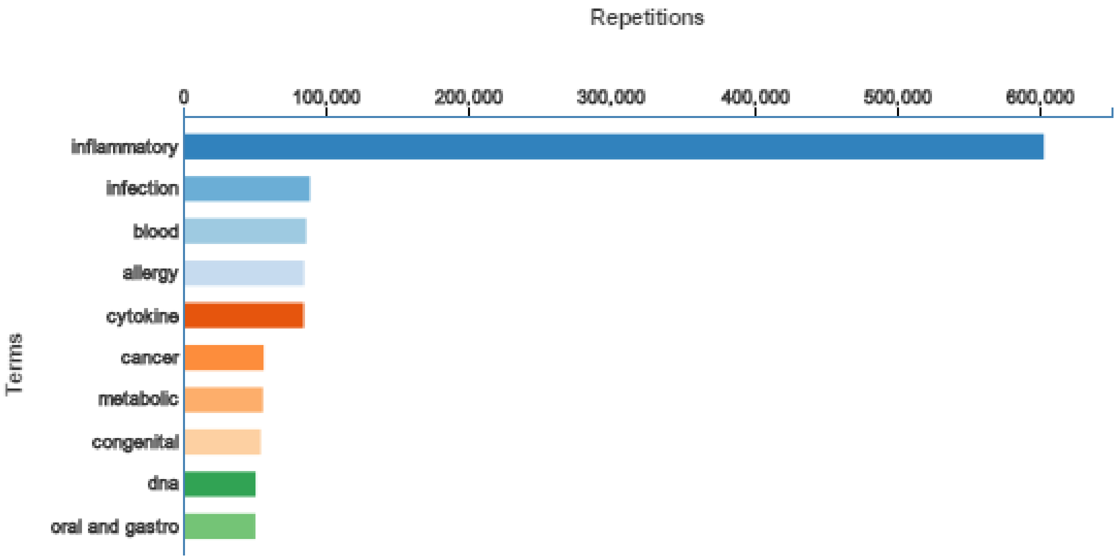


Figure 19 - Ten of the most frequent terms in *'Inflammatory'*

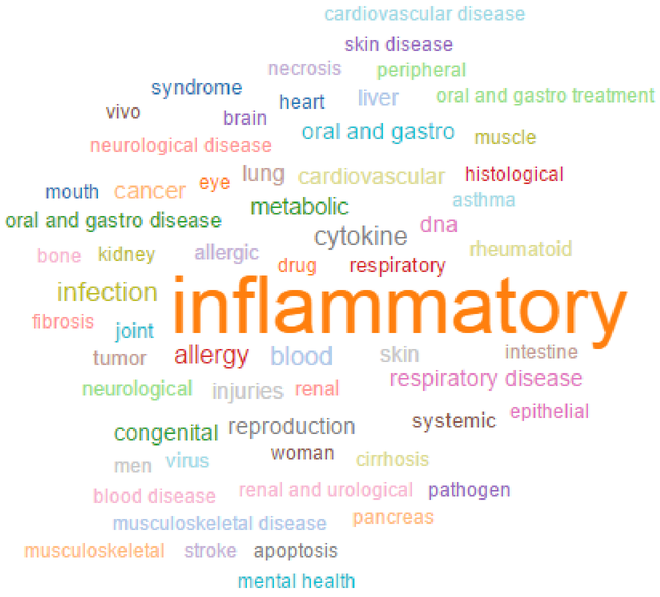


Figure 20 - *'Inflammatory'* wordcloud

3.1.9 Topic *‘Injuries’*

The CTM model has classified 297.008 papers into ‘*Injuries*’.

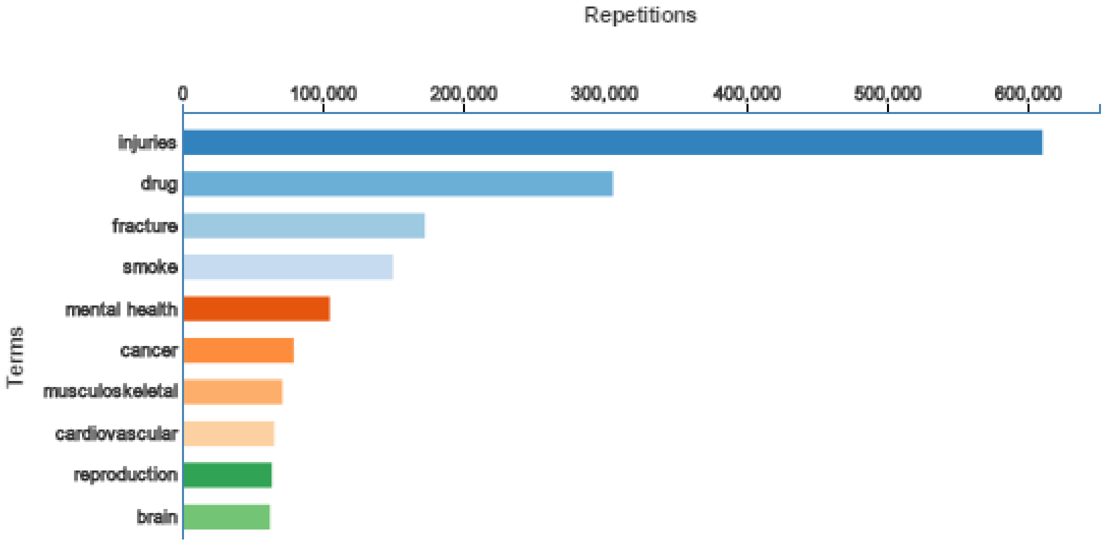


Figure 21 - Ten of the most frequent terms in *'Injuries'*

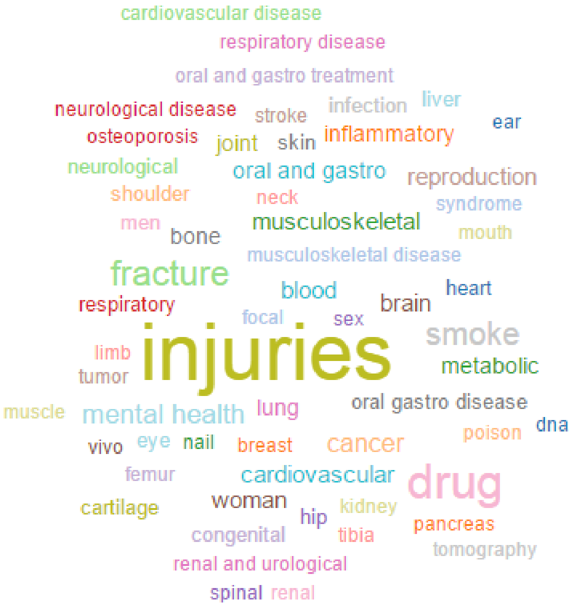


Figure 22 - *'Injuries'* wordcloud

3.1.10 Topic ‘*Mental Health*’

The CTM model has classified 312.730 papers into ‘*Mental Health*’.

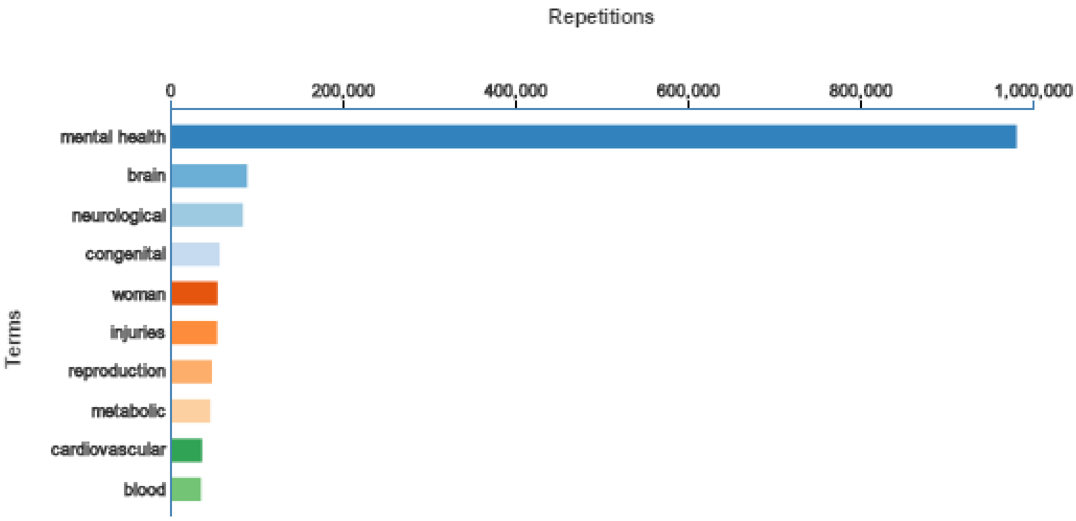


Figure 23 - Ten of the most frequent terms in *‘Mental Health’*

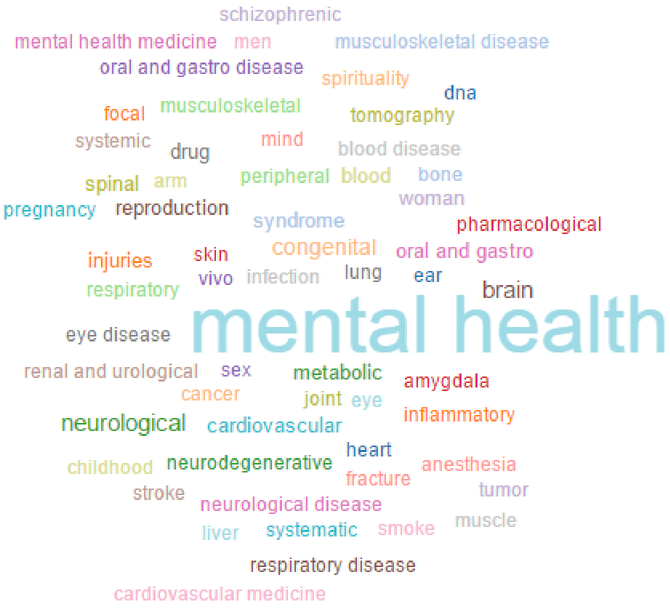


Figure 24 - *‘Mental Health’* wordcloud

3.1.11 Topic ‘*Metabolic’*

The CTM model has classified 325.472 papers into ‘*Metabolic’*.

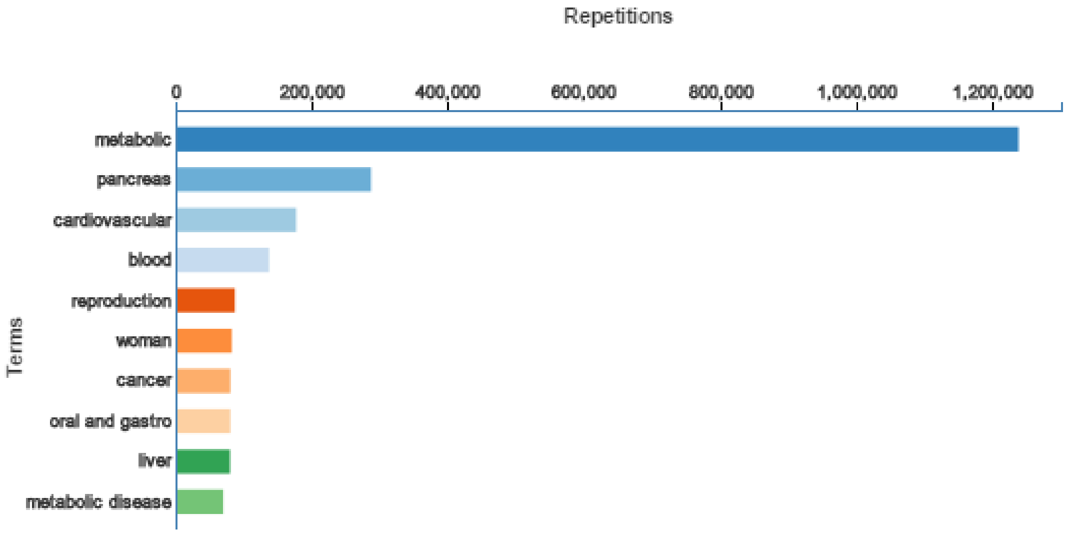


Figure 25- Ten of the most frequent terms in 'Metabolic'

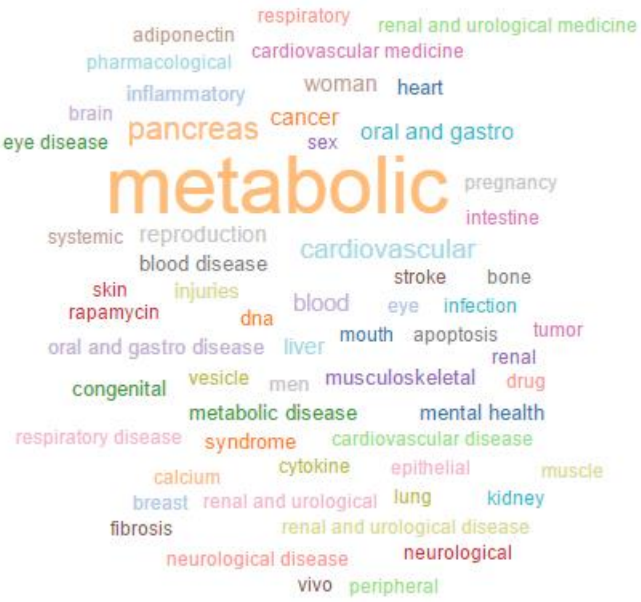


Figure 26 – ‘Metabolic’ wordcloud

3.1.12 Topic ‘*Musculoskeletal’*

The CTM model has classified 311.330 papers into ‘*Musculoskeletal’*.

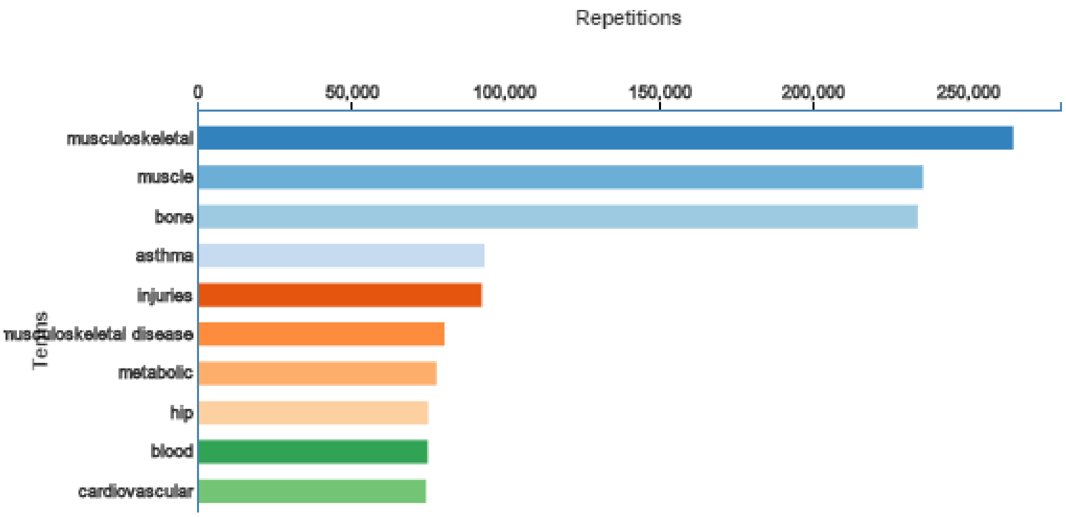


Figure 27- Ten of the most frequent terms in *‘Musculoskeletal’*

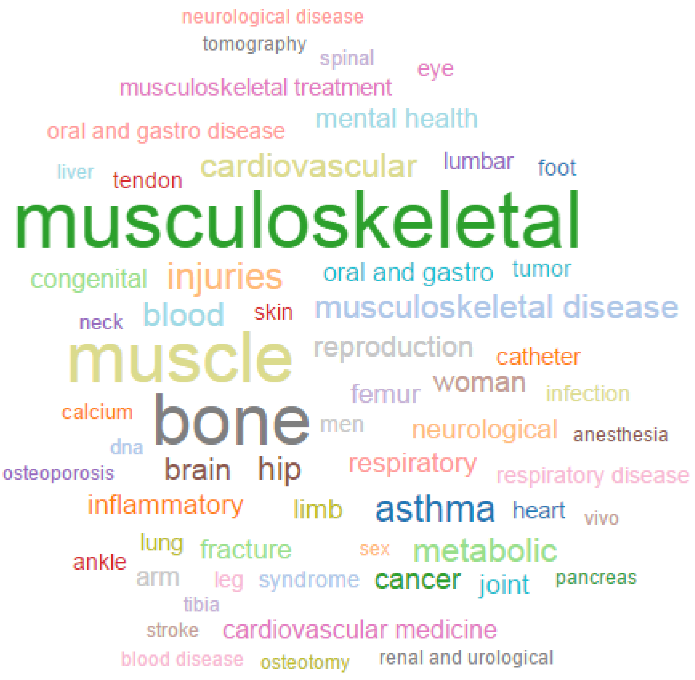


Figure 28- *‘Musculoskeletal’* wordcloud

3.1.13 Topic ‘*Neurological’*

The CTM model has classified 314.733 papers into ‘*Neurological*’.

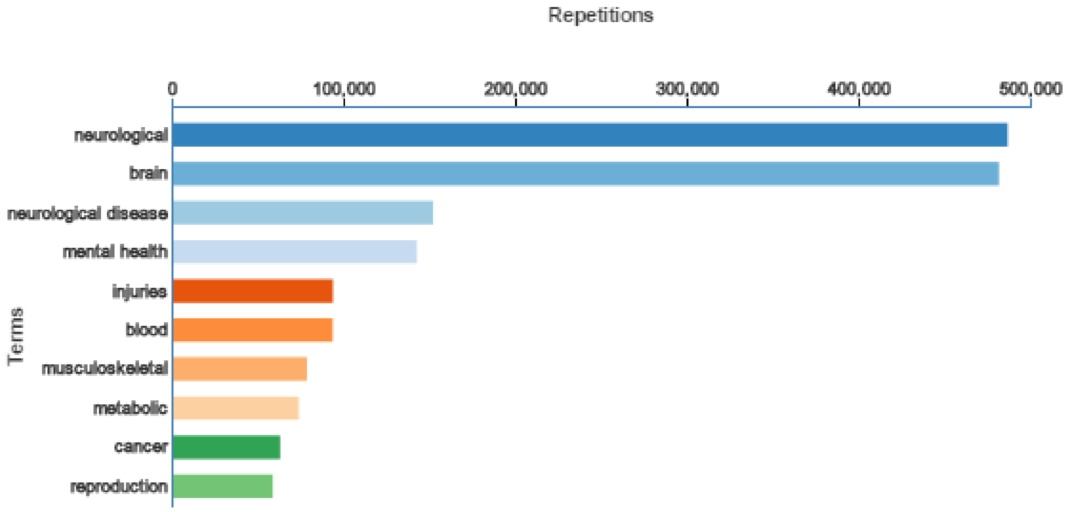


Figure 29- Ten of the most frequent terms in *‘Neurological’*

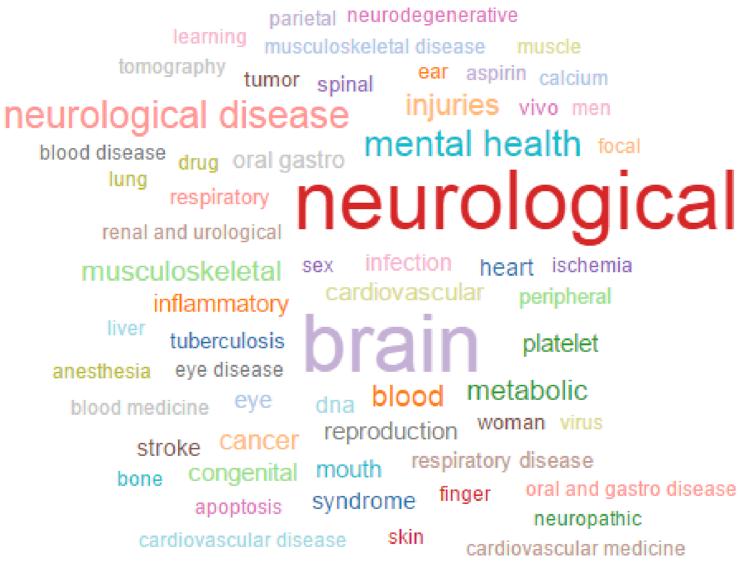


Figure 30 - *‘Neurological’* wordcloud

3.1.14 Topic ‘*Oral and Gastro’*

The CTM model has classified 357.896 papers into ‘*Oral and Gastro’*.

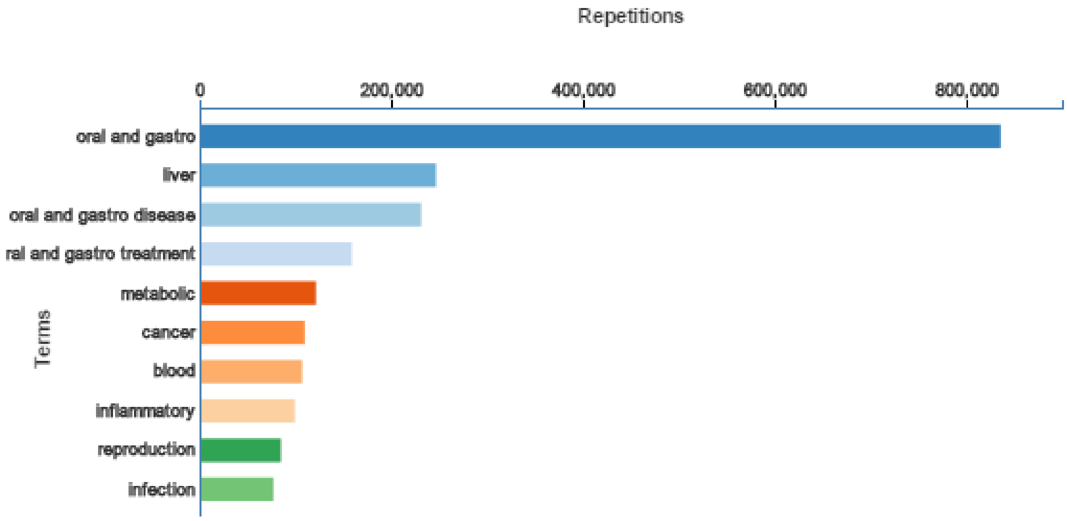


Figure 31 - Ten of the most frequent terms in *‘Oral and Gastro’*

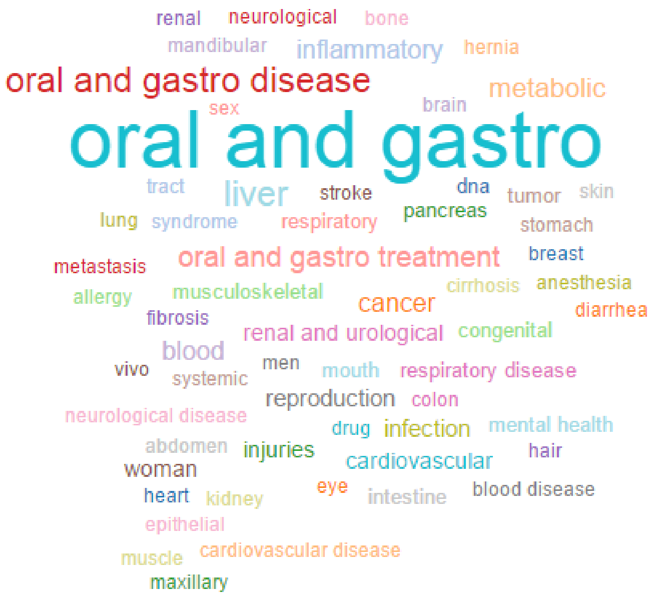


Figure 32 - *‘Oral and Gastro’* wordcloud

3.1.15 Topic *'Renal and Urological'*

The CTM model has classified 122.696 papers into '*Renal and Urological*'.

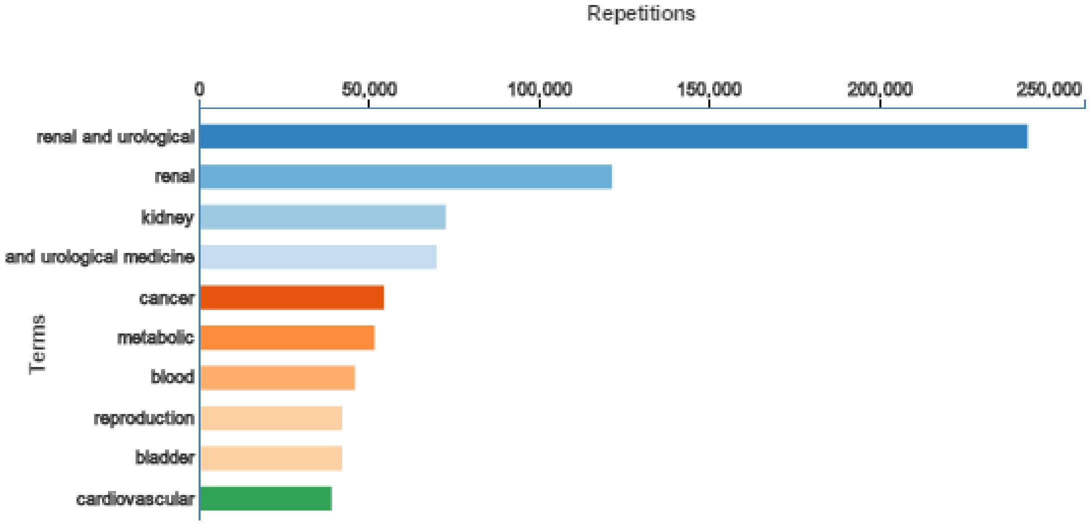


Figure 33- Ten of the most frequent terms in *'Renal and Urological'*

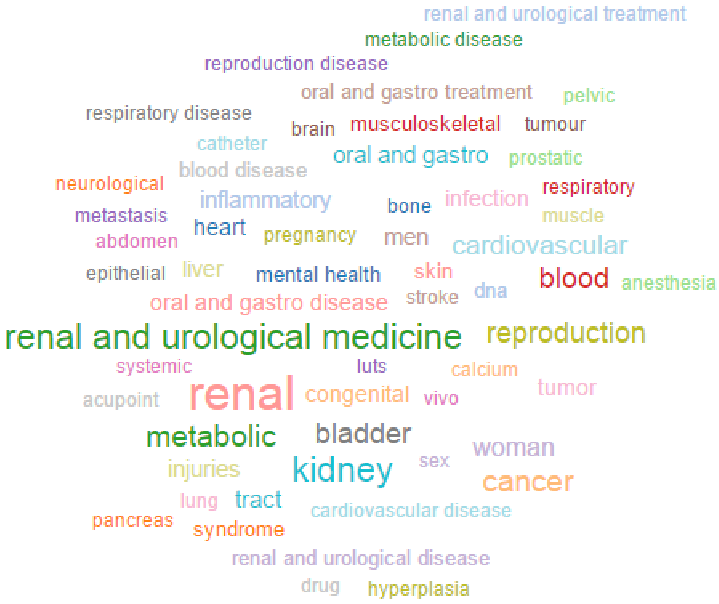


Figure 34 - *'Renal and Urological'* wordcloud

3.1.16 Topic '*Reproduction*'

The CTM model has classified 356.336 papers into '*Reproduction*'.

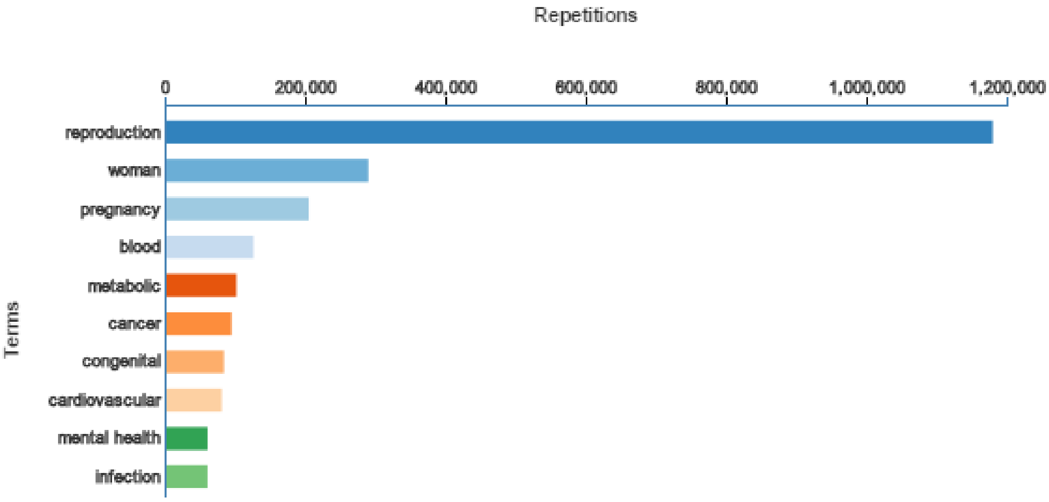


Figure 35 - Ten of the most frequent terms in *'Reproduction'*

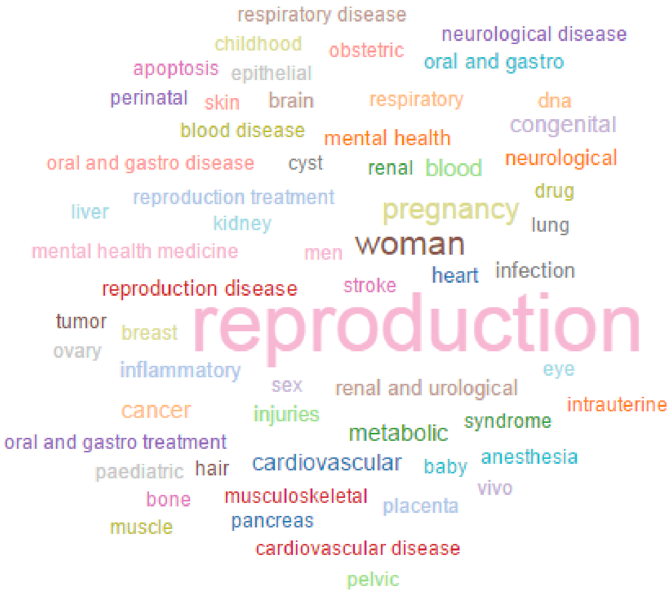


Figure 36 - *'Reproduction'* wordcloud

3.1.17 Topic ‘*Respiratory’*

The CTM model has classified 181.404 papers into ‘*Respiratory*’.

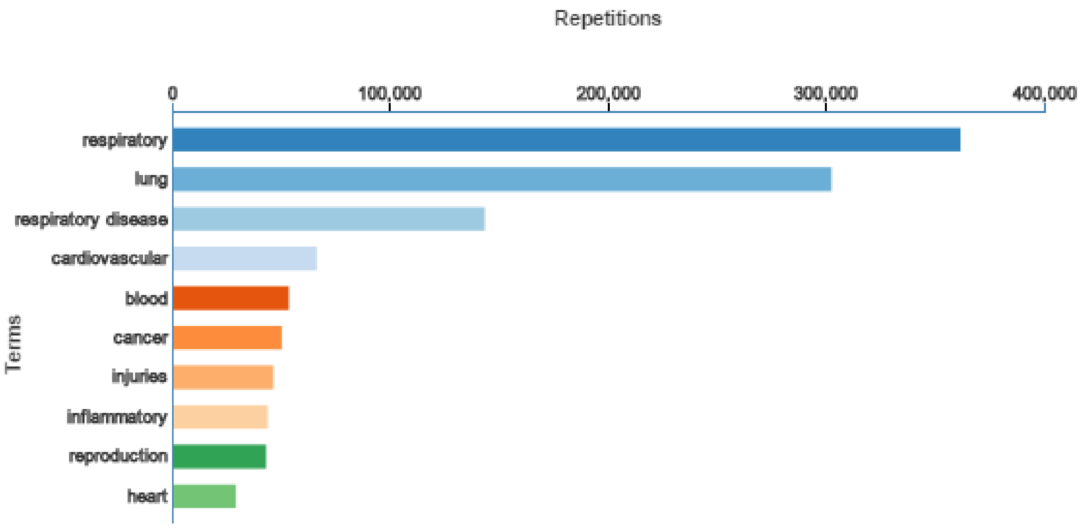


Figure 37 - Ten of the most frequent terms in *‘Respiratory’*

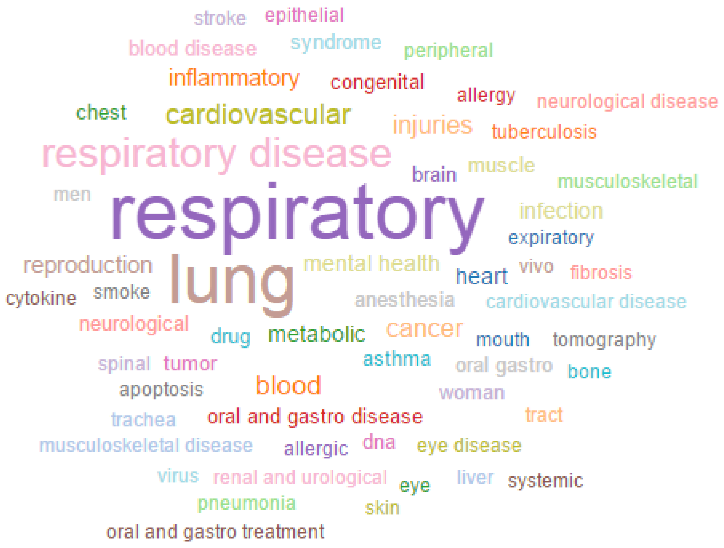


Figure 38 - *‘Respiratory’* wordcloud

3.1.18 Topic ‘*Skin’*

The CTM model has classified 122.838 papers into ‘*Skin*’.

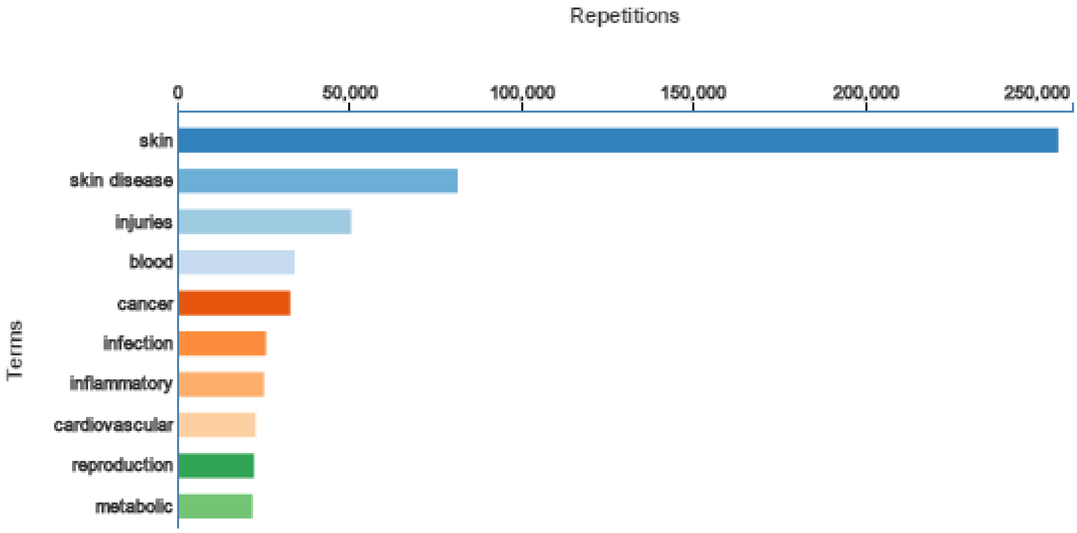


Figure 39 - Ten of the most frequent terms in *‘Skin’*

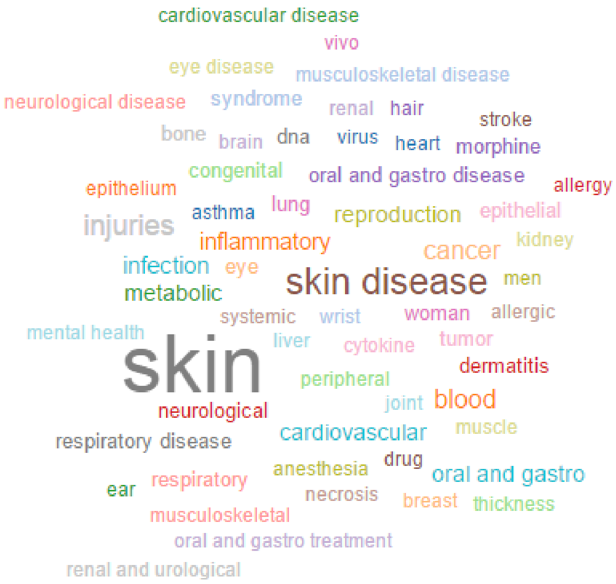


Figure 40 - *‘Skin’* wordcloud

3.1.19 Topic ‘*Stroke’*

The CTM model has classified 60.635 papers into ‘*Stroke*’.

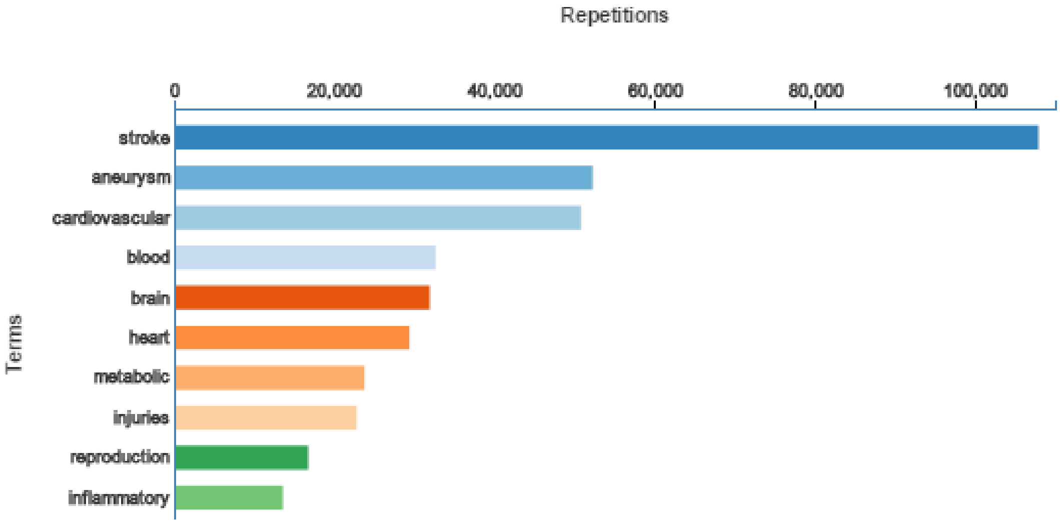


Figure 41 - Ten of the most frequent terms in ***‘Stroke’***

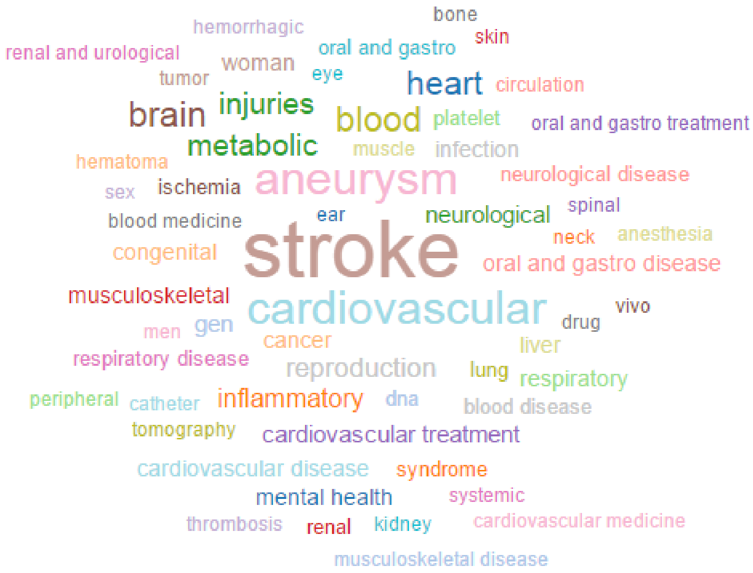


Figure 42 - ***‘Stroke’*** **wordcloud**

3.1.20 Topic '*Other'*

The CTM model has classified 654.472 papers into '*Other*'.

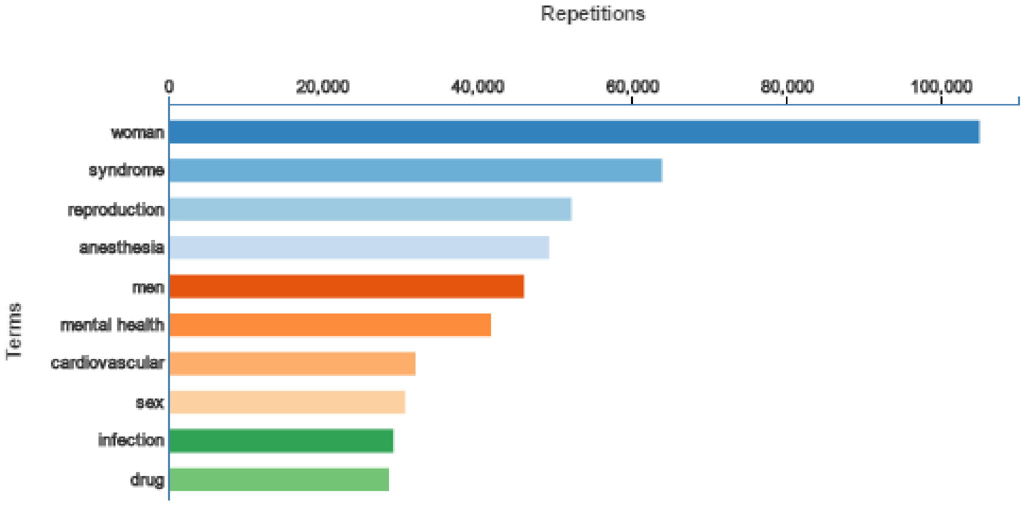


Figure 43 - Ten of the most frequent terms in *'Other'*

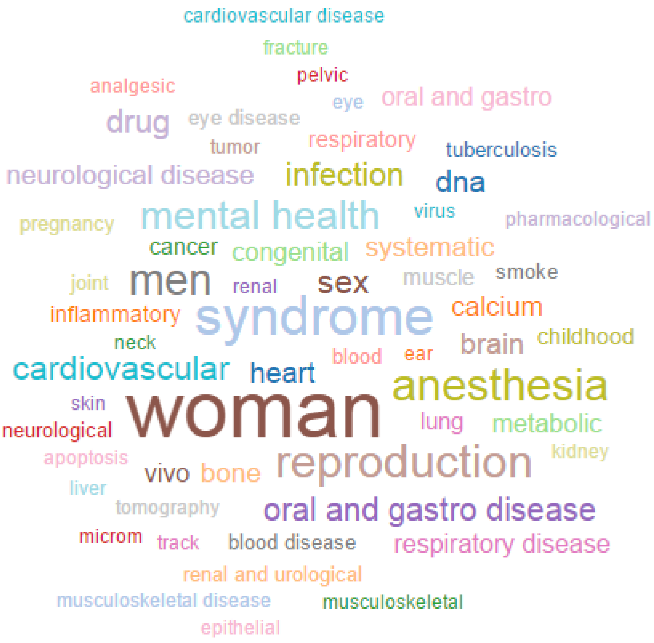


Figure 44 - *'Other'* wordcloud

3.2 Topic summary OCDE categories

Besides the 28 predefined OCDE categories, there is an additional cluster with a mixture of terms which are also included in other clusters. This group is named '*Other*'.

The following table shows the abstracts count within each category:

Abstracts	Topic Name
83.841	Andrology
669.363	Cardiovascular
464.116	Clinical Neurology
6.884	Complementary Medicine
412.291	Endocrinology
746.563	Environmental Heath
664.104	Epidemiology
244.968	Genetics Heredity
275.816	Human Genetics
271.452	Immunology
188.715	Medical Devices
698.270	Neuroscience
200.591	Nuclear Medicine
78.418	Nutrition and Dietetics
164.697	Obstetrics
612.521	Oncology
183.676	Ophthalmology
294547	Paediatrics
144.246	Personalised Medicine
120.338	Pharmacology
6.926	Physiology
316.641	Psychiatry
419.709	Public Health
112.896	Regenerative Medicine
315.496	Rheumatology
236.828	Surgery
917.091	System Biology
121.354	Toxicology
728.403	Other

Figure 45 - Abstracts count OCDE categories

3.2.1 Topic '*Andrology*'

The CTM model has classified 83.841 papers into '*Andrology*'.

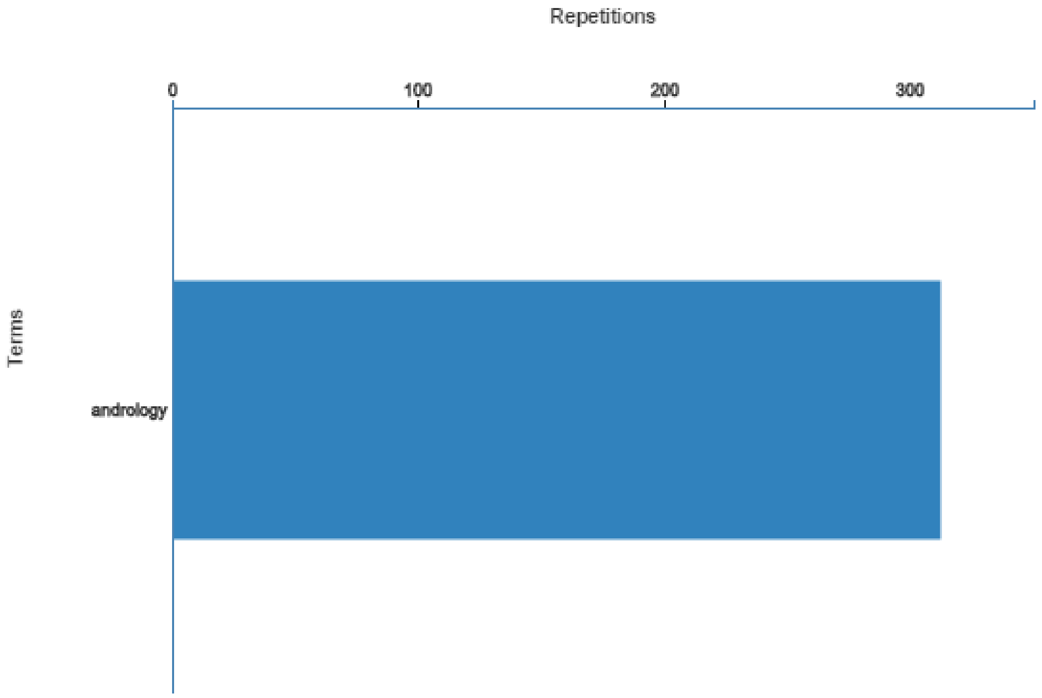


Figure 46 - Ten of the most frequent terms in *'Andrology'*

Figure 47 - *'Andrology'* wordcloud

3.2.2 Topic '*Cardiovascular'*

The CTM model has classified 669.363 papers into '*Cardiovascular*'.

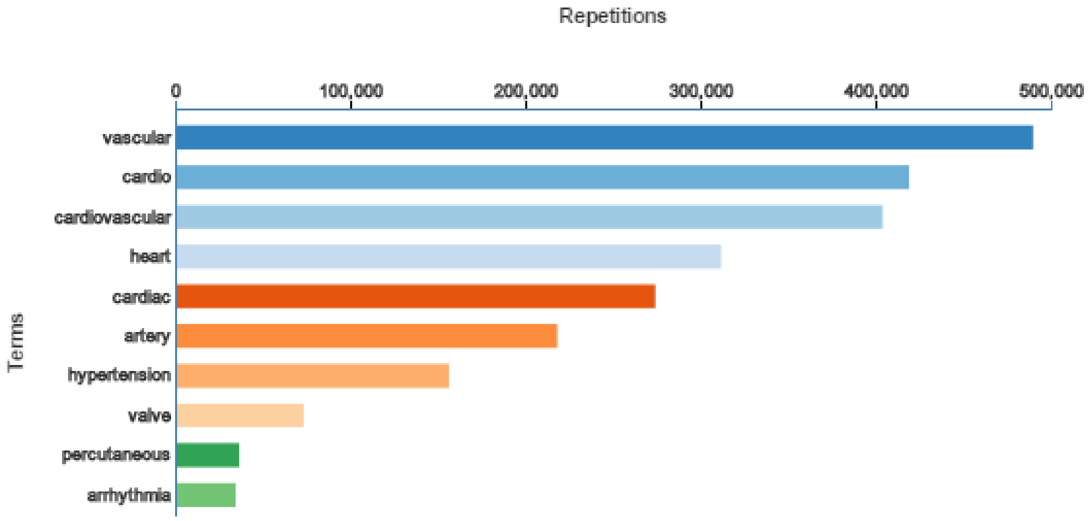


Figure 48 - Ten of the most frequent terms in *'Cardiovascular'*

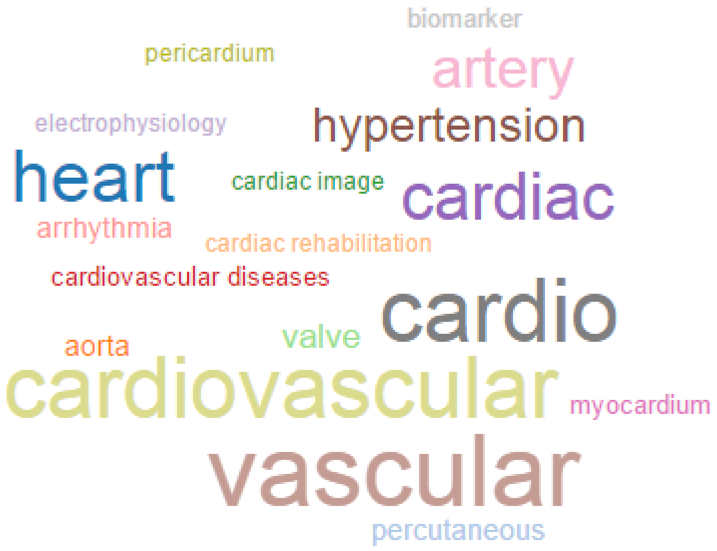


Figure 49 - '*Cardiovascular'* wordcloud

3.2.3 Topic '*Clinical Neurology'*

The CTM model has classified 464.116 papers into '*Clinical Neurology*'.

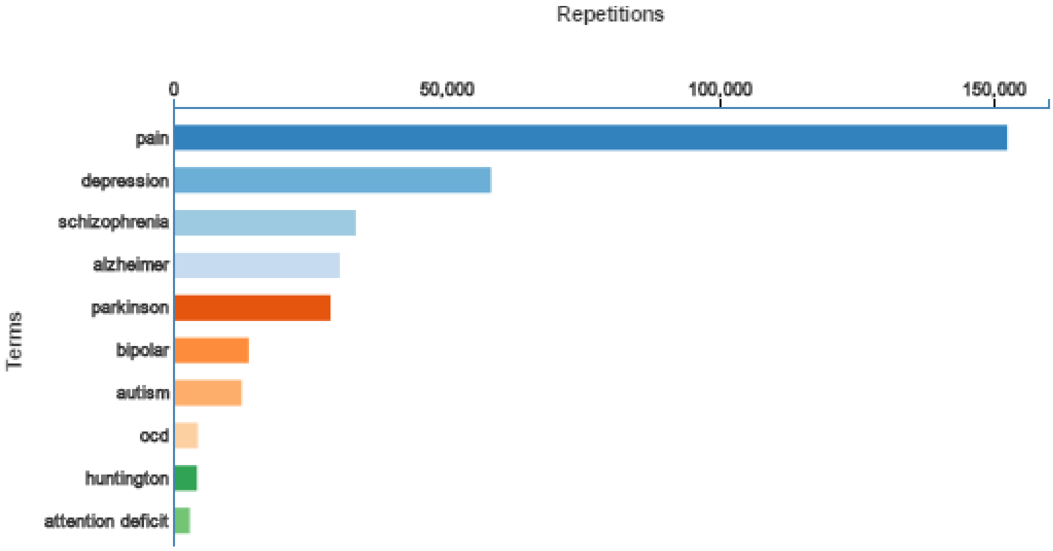


Figure 50 - Ten of the most frequent terms in *'Clinical Neurology'*

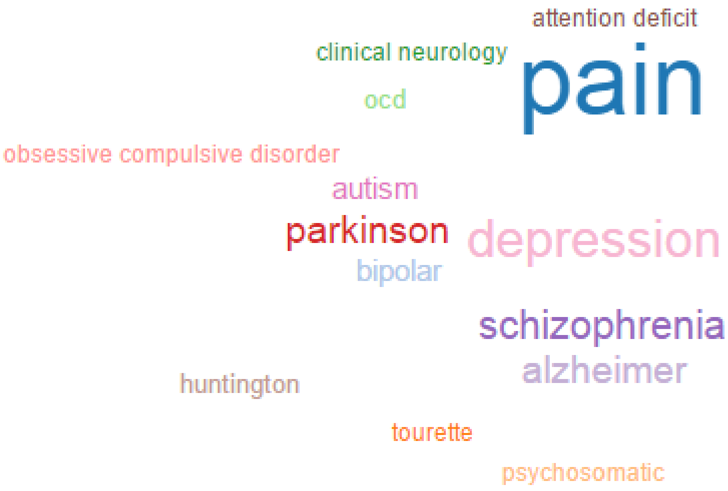


Figure 51 - '*Clinical Neurology'* wordcloud

3.2.4 Topic '*Complementary Medicine'*

The CTM model has classified 6.884 papers into '*Complementary Medicine'*.

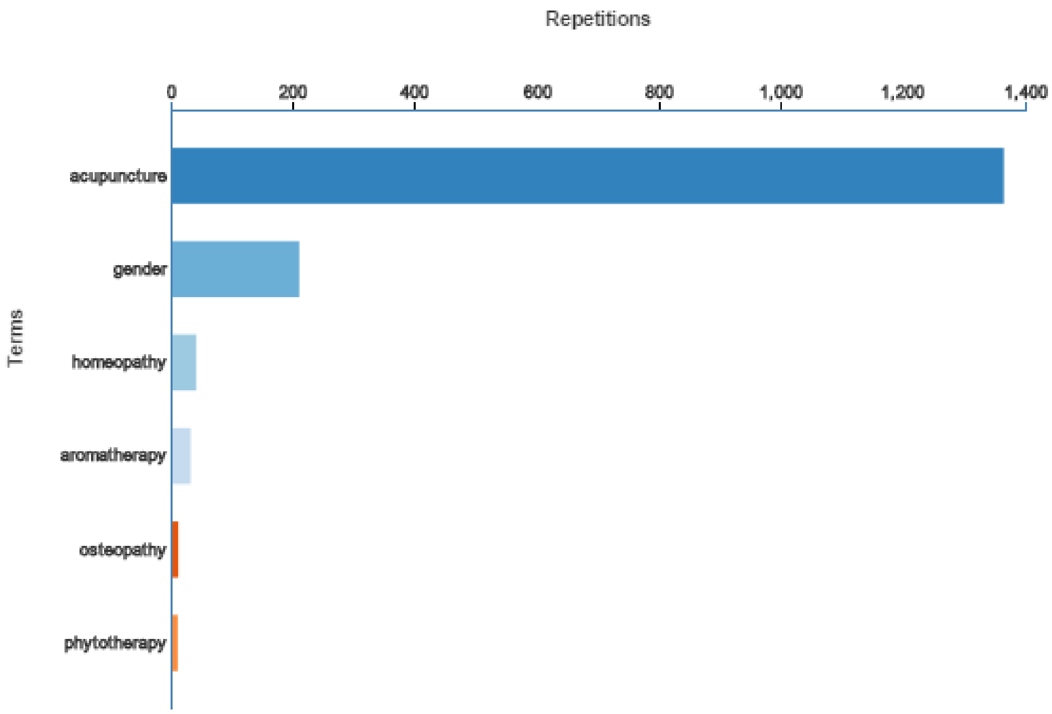


Figure 52 - Ten of the most frequent terms in *'Complementary Medicine'*

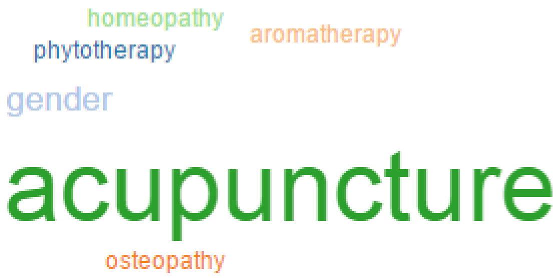


Figure 53 -*'Complementary Medicine'* wordcloud

3.2.5 Topic *'Endocrinology'*

The CTM model has classified 412.291 papers into '*Endocrinology*'.

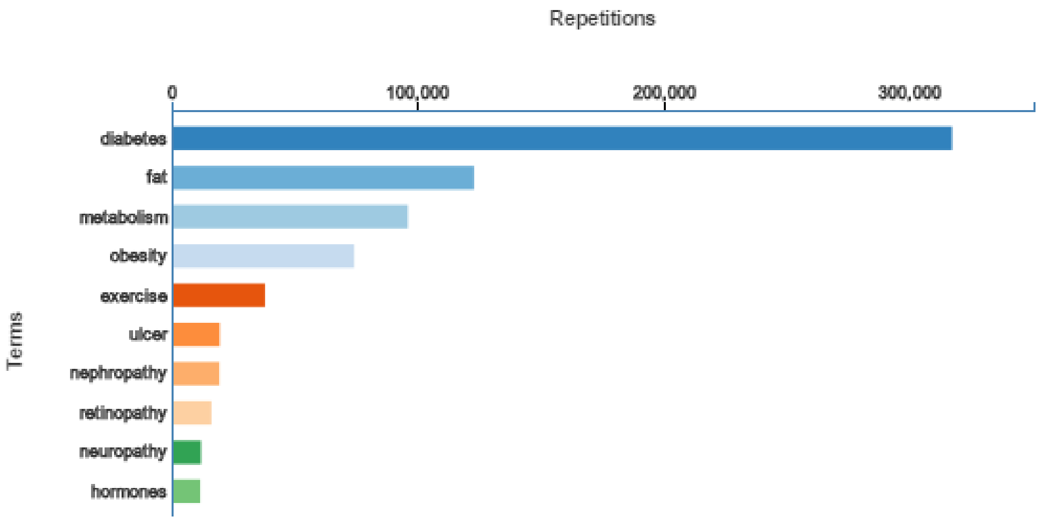


Figure 54 - Ten of the most frequent terms in *'Endocrinology'*

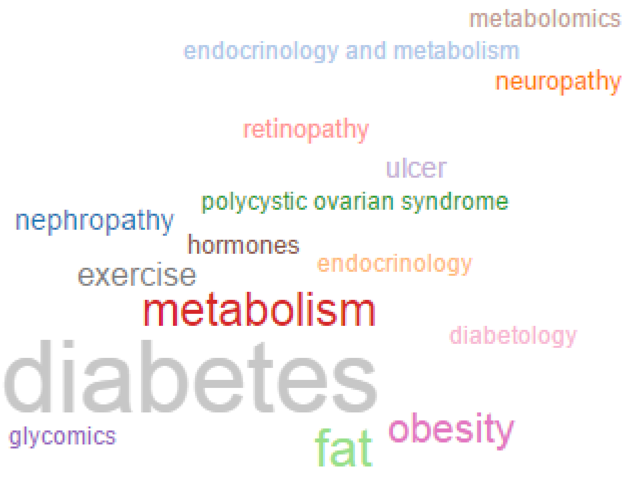


Figure 55 - *'Endocrinology'* wordcloud

3.2.6 Topic '*Environmental Health'*

The CTM model has classified 746.563 papers into '*Environmental Health*'.

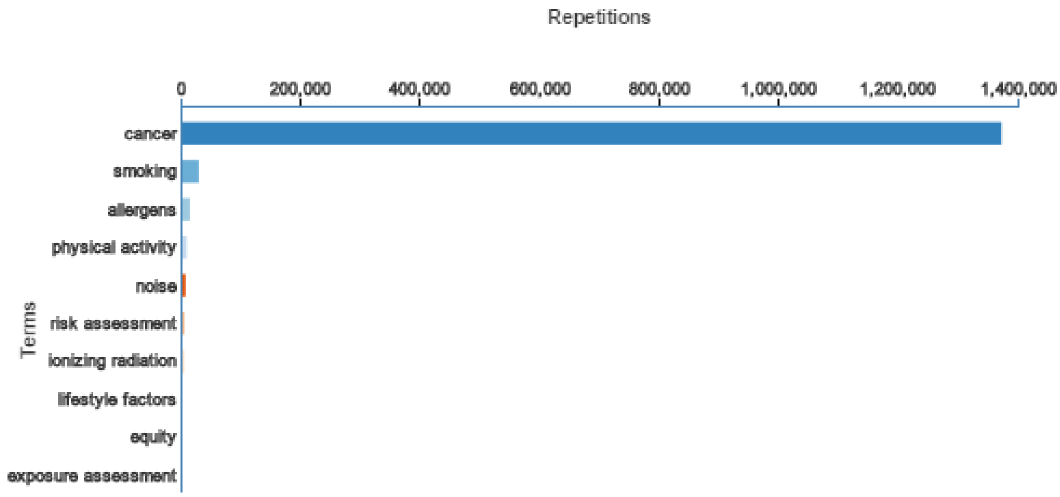


Figure 56 - Ten of the most frequent terms in *'Environmental Health'*

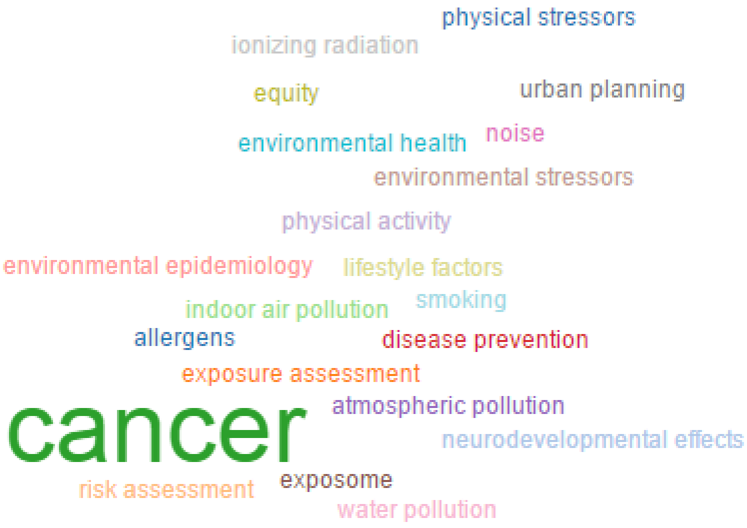


Figure 57 - *'Environmental Health'* wordcloud

3.2.7 Topic '*Epidemiology'*

The CTM model has classified 664.104 papers into '*Epidemiology*'.

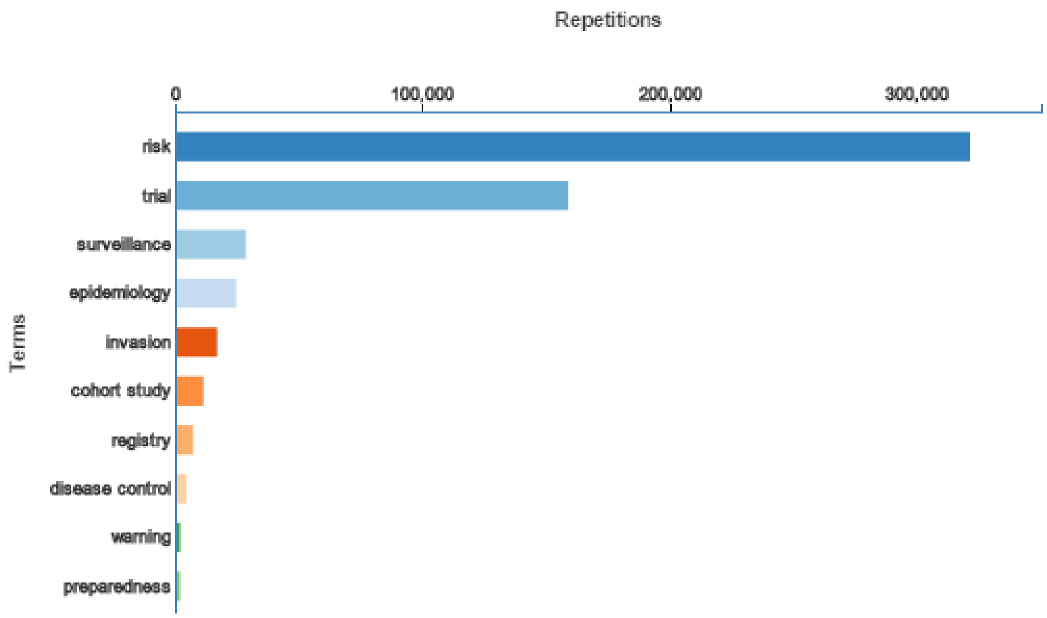


Figure 58 - Ten of the most frequent terms in *'Epidemiology'*

registry
metapopulation
warning
cohort study
fungicide
trial
epidemiology
preparedness
invasion
risk
disease control
surveillance

Figure 59 -*'Epidemiology'* wordcloud

3.2.8 Topic '*Genetics Heredity'*

The CTM model has classified 244.968 papers into '*Genetics Heredity*'.

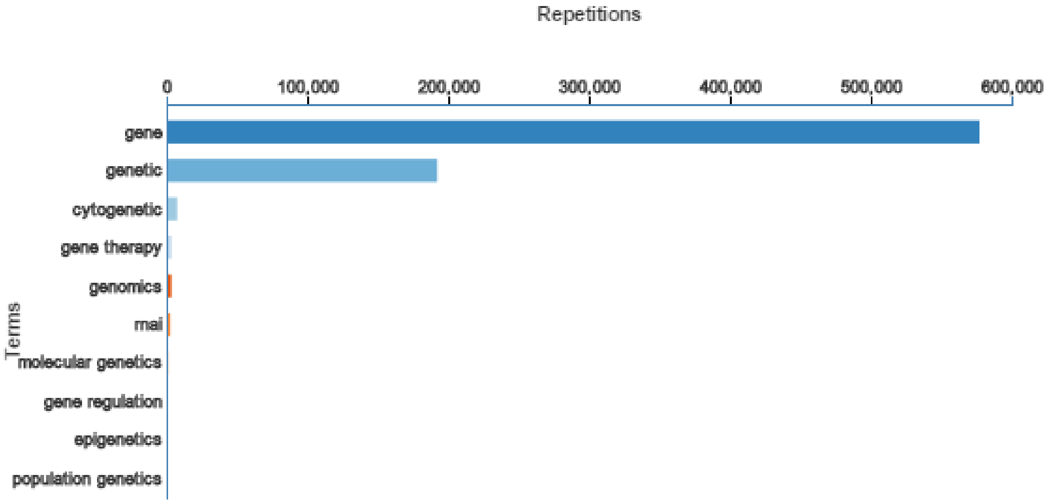


Figure 60 - Ten of the most frequent terms in *'Genetics Heredity'*

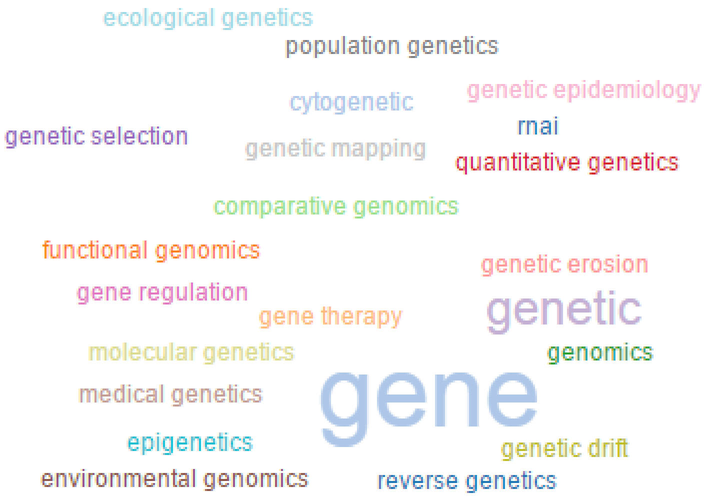


Figure 61 - *'Genetics Heredity'* wordcloud

3.2.9 Topic '*Human Genetics'*

The CTM model has classified 275.816 papers into '*Human Genetics*'.

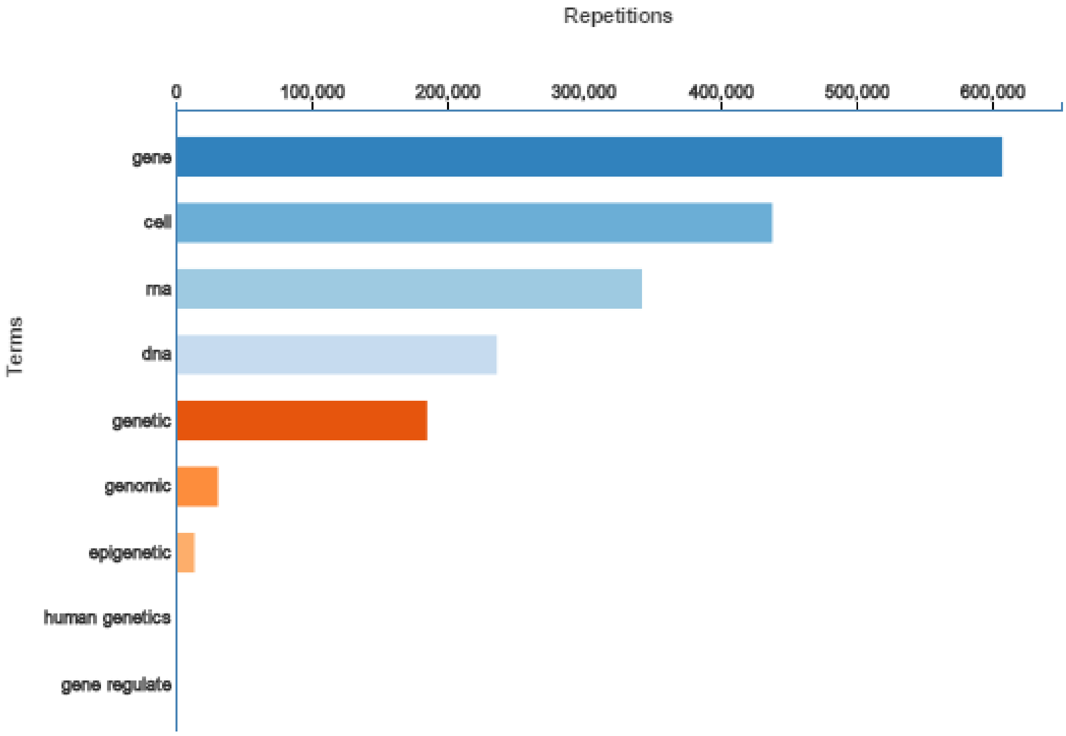


Figure 62 - Ten of the most frequent terms in *'Human Genetics'*

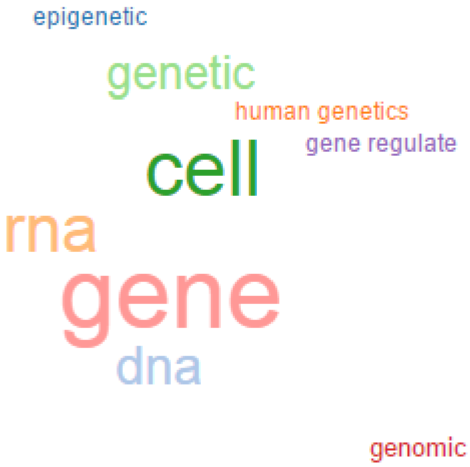


Figure 63 -*'Human Genetics'* wordcloud

3.2.10 Topic '*Immunology'*

The CTM model has classified 271.452 papers into '*Immunology*'.

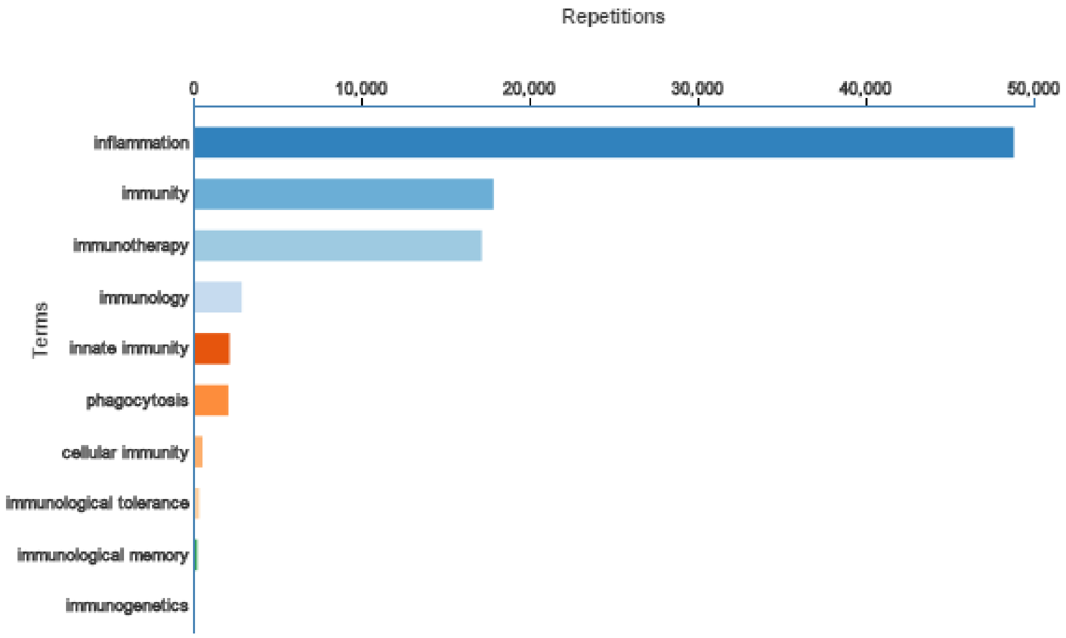


Figure 64 - Ten of the most frequent terms in *'Immunology'*

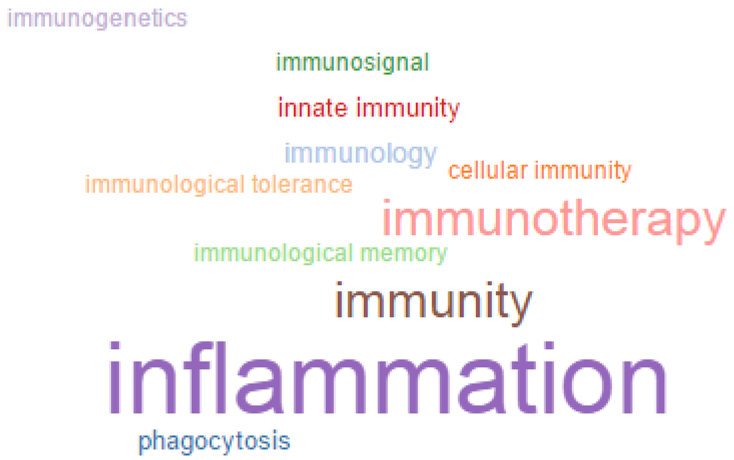


Figure 65 - *'Immunology'* wordcloud

3.2.11 Topic '*Medical Devices'*

The CTM model has classified 188.715 papers into '*Medical Devices*'.

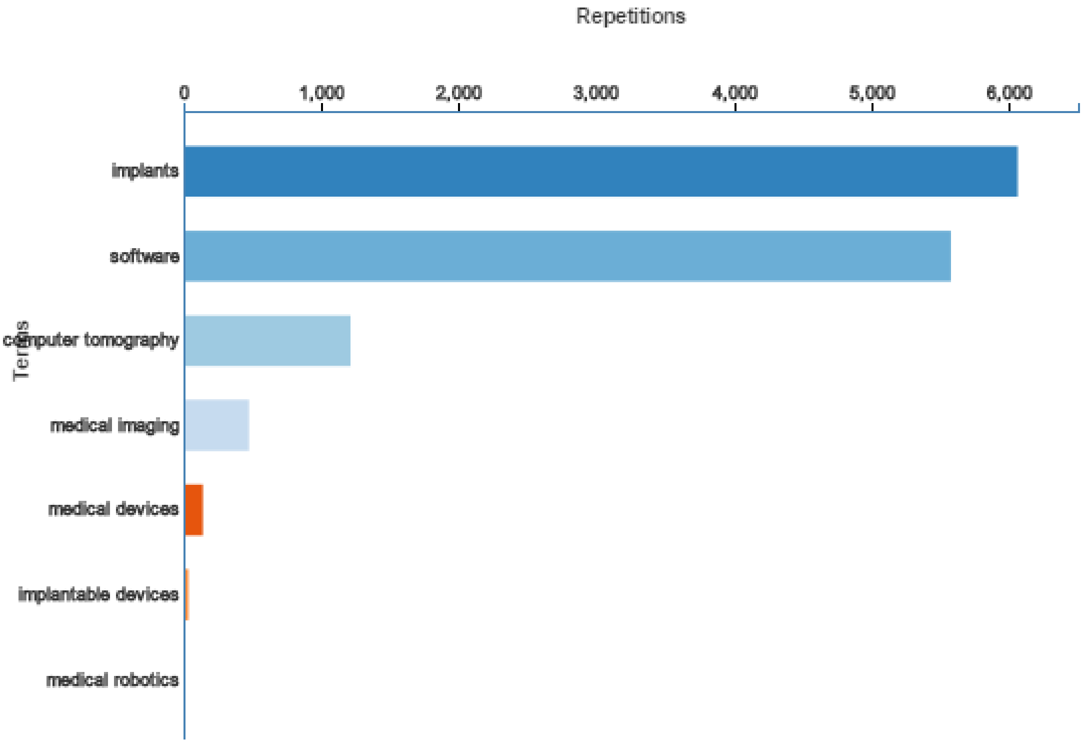


Figure 66 - Ten of the most frequent terms in *'Medical Devices'*

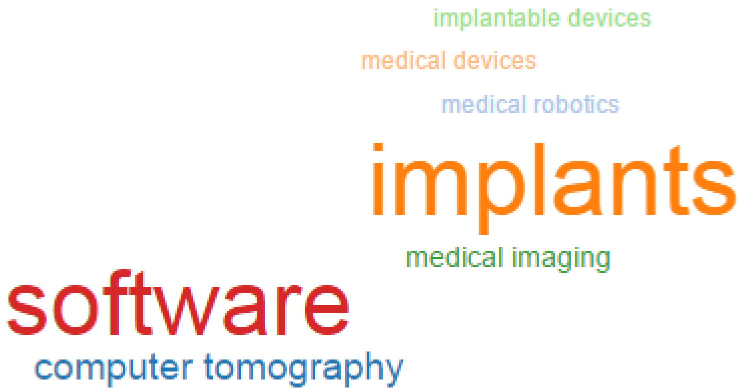


Figure 67 - *'Medical Devices'* wordcloud

3.2.12 Topic '*Neuroscience'*

The CTM model has classified 698.270 papers into '*Neuroscience*'.

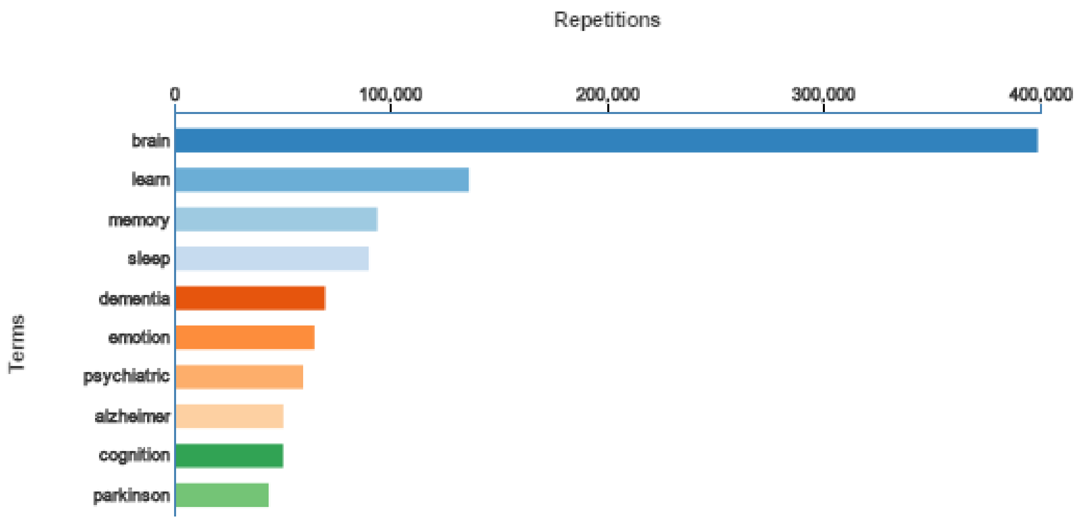


Figure 68 - Ten of the most frequent terms in *'Neuroscience'*

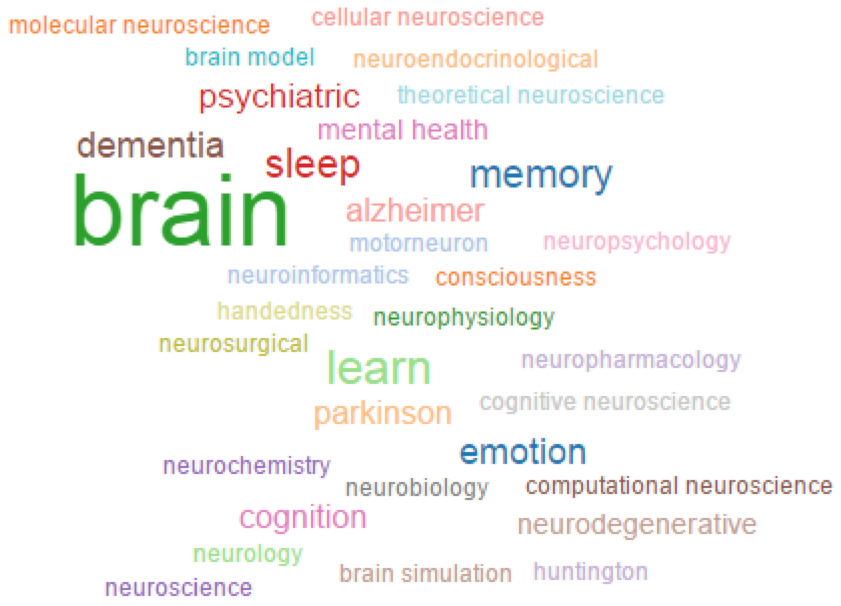


Figure 69 -*'Neuroscience'* wordcloud

3.2.13 Topic *'Nuclear Medicine'*

The CTM model has classified 200.591 papers into '*Nuclear Medicine*'.

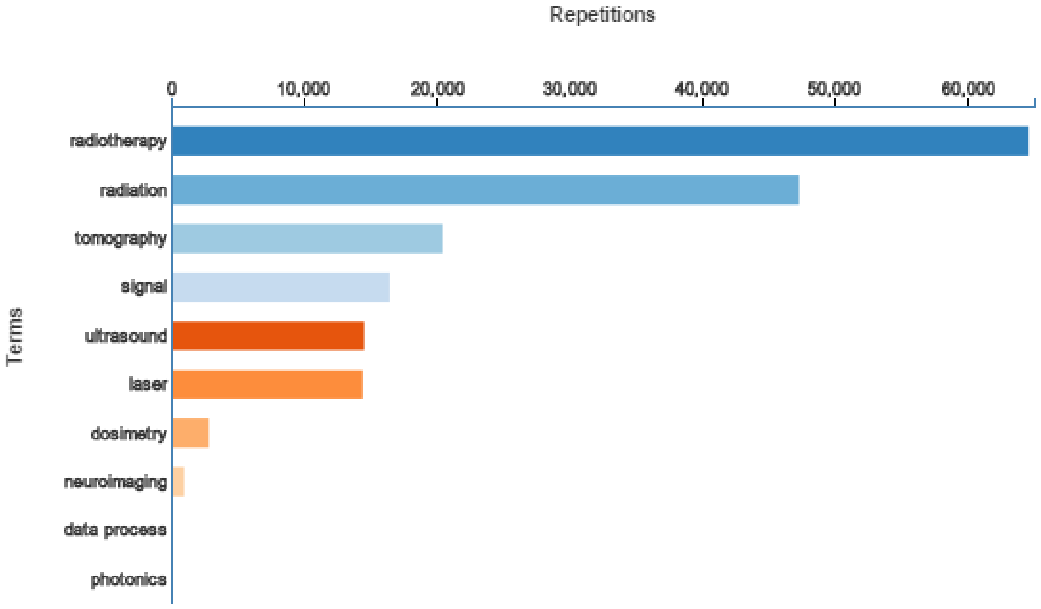


Figure 70 - Ten of the most frequent terms in *'Nuclear Medicine'*

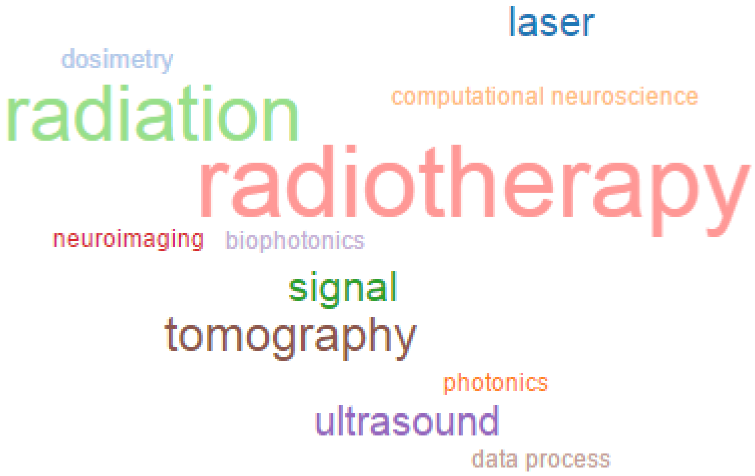


Figure 71 - '*Nuclear Medicine*' wordcloud

3.2.14 Topic '*Nutrition and Dietetics*'

The CTM model has classified 78.418 papers into '*Nutrition and Dietetics*'.

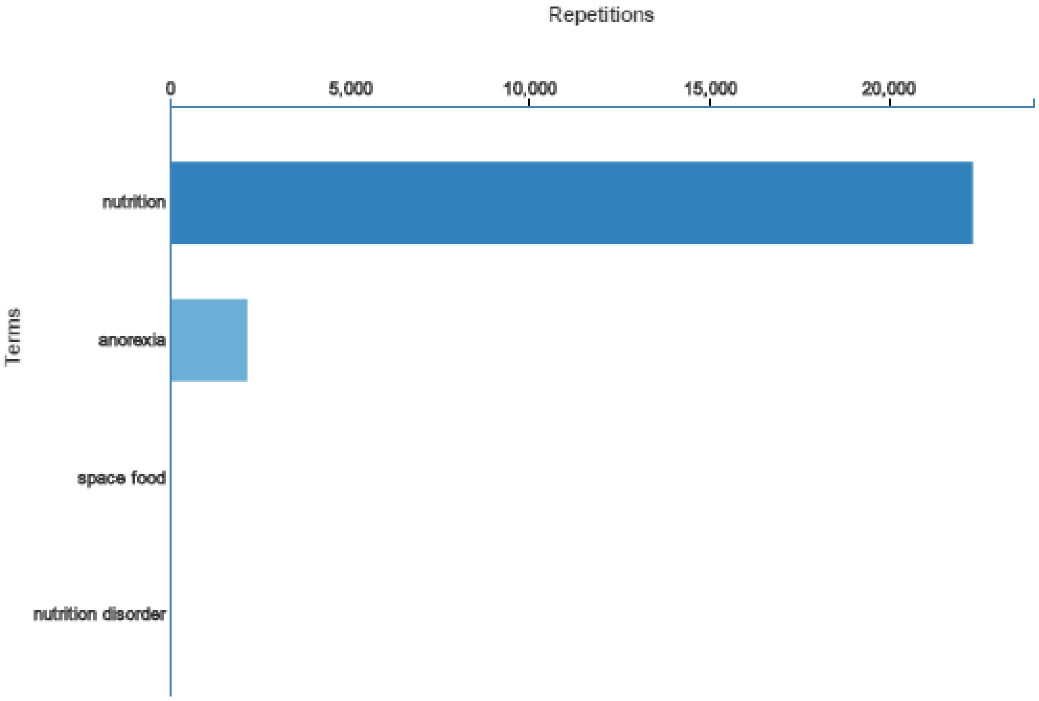


Figure 72 - Ten of the most frequent terms in *‘Nutrition and Dietetics’*

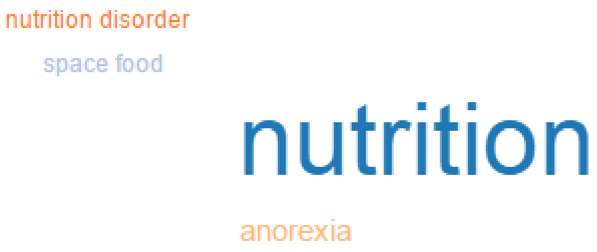


Figure 73 -*‘Nutrition and Dietetics’* wordcloud

3.2.15 **Topic ‘*Obstetrics’***

The CTM model has classified 164.697 papers into ‘*Obstetrics*’.

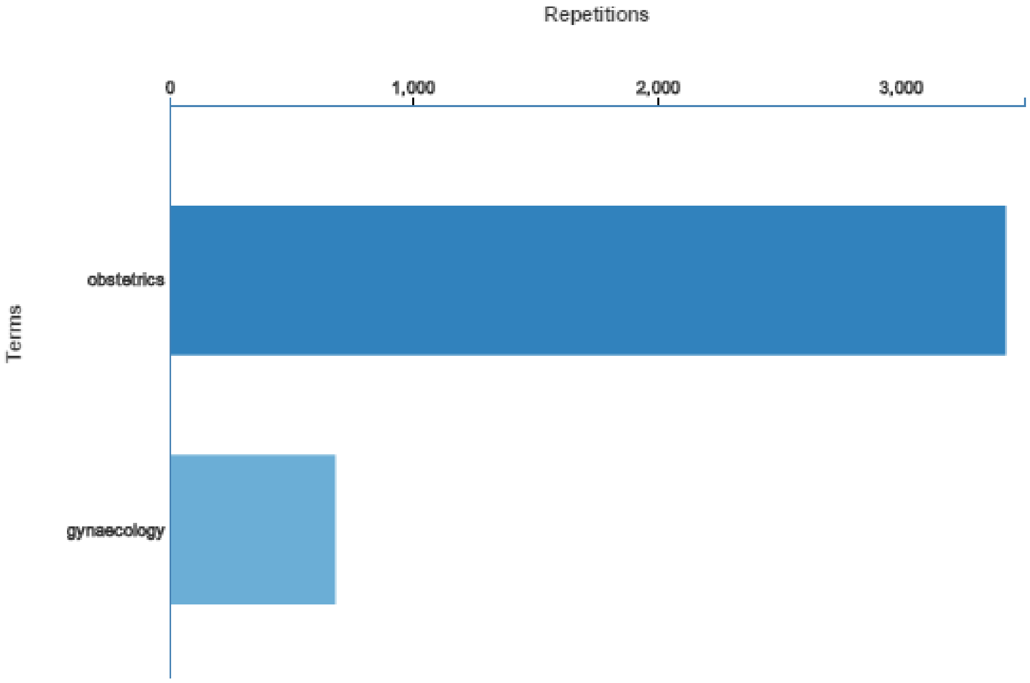


Figure 74 - Ten of the most frequent terms in *'Obstetrics'*

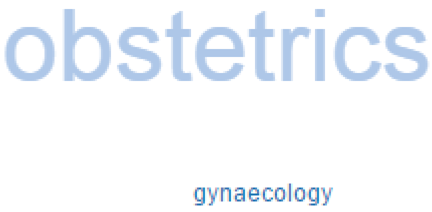


Figure 75- *'Obstetrics'* wordcloud

3.2.16 Topic '*Oncology*'

The CTM model has classified 612.521 papers into '*Oncology*'.

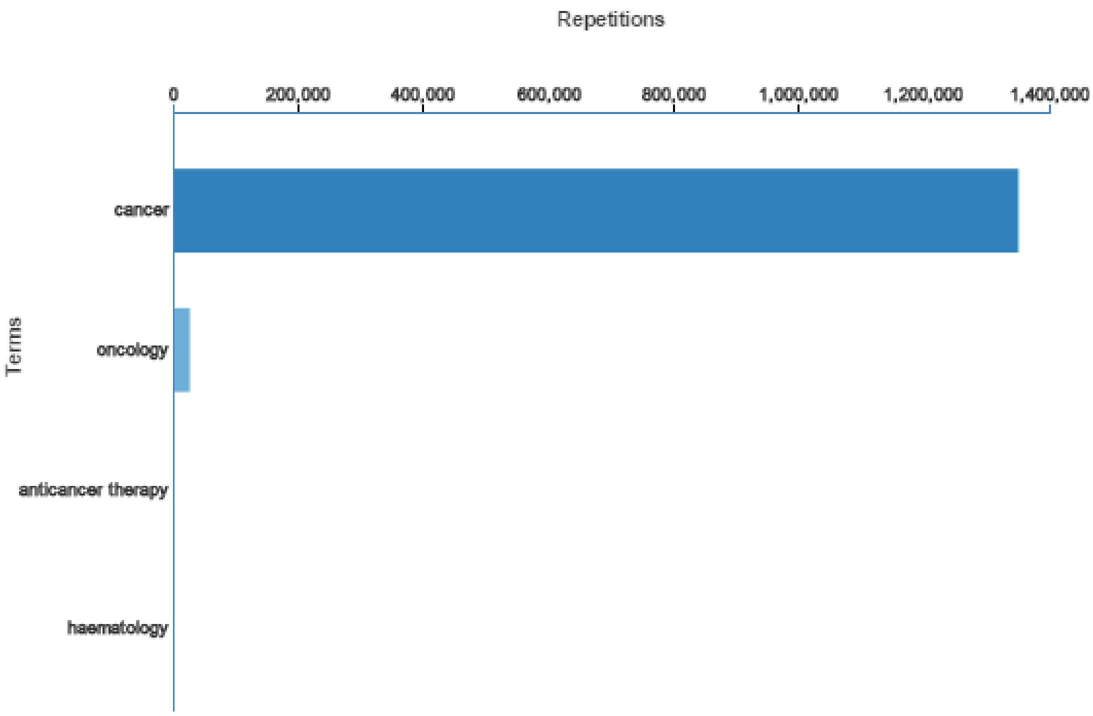


Figure 76 - Ten of the most frequent terms in *'Oncology'*

cancer
oncology
anticancer therapy
haematology

Figure 77 - '*Oncology'* wordcloud

3.2.17 **Topic '*Ophthalmology'***

The CTM model has classified 183.676 papers into '*Ophthalmology'*.

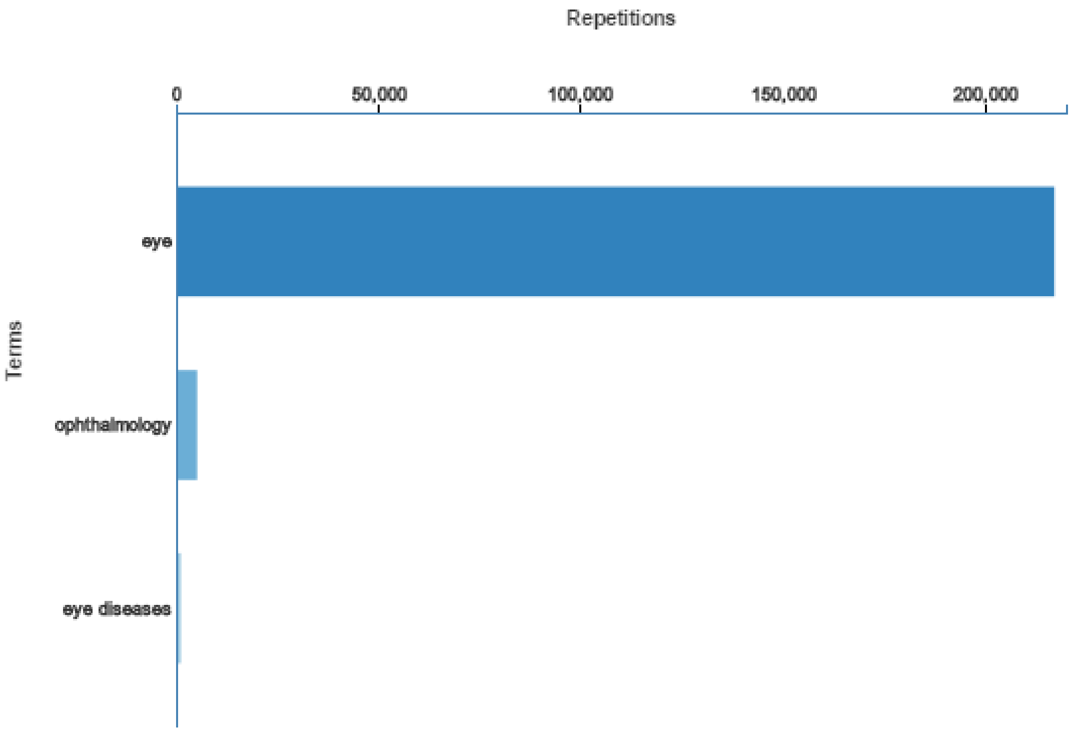


Figure 78 - Ten of the most frequent terms in *'Ophthalmology'*

eye diseases
ophthalmology
eye

Figure 79 - 'Ophthalmology' wordcloud

3.2.18 Topic '*Other*'

There are no defined terms under this category.

3.2.19 Topic ‘*Paediatrics’*

The CTM model has classified 294.547 papers into ‘*Paediatrics*’.

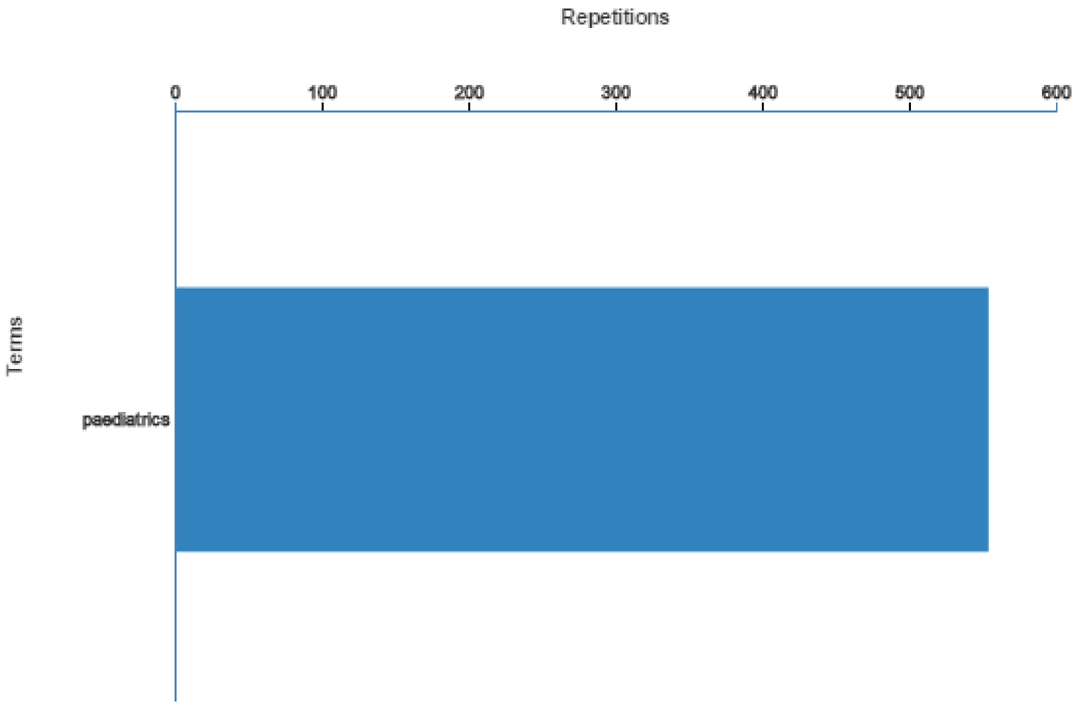


Figure 80 - Ten of the most frequent terms in *‘Paediatrics’*

Figure 81 - ‘*Paediatrics’* wordcloud

3.2.20 Topic ‘*Personalised Medicine’*

The CTM model has classified 144.246 papers into ‘*Personalised Medicine*’.

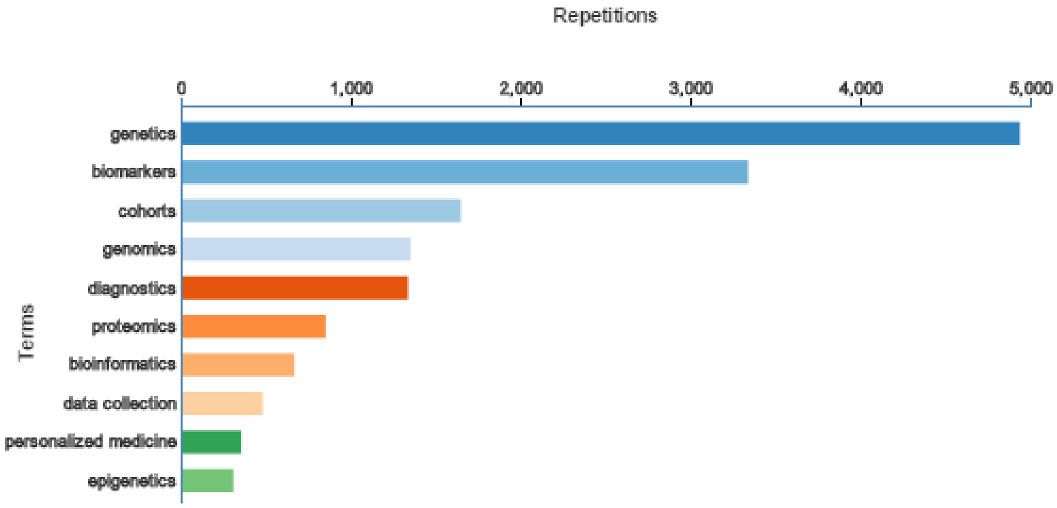


Figure 82 - Ten of the most frequent terms in *'Personalised Medicine'*

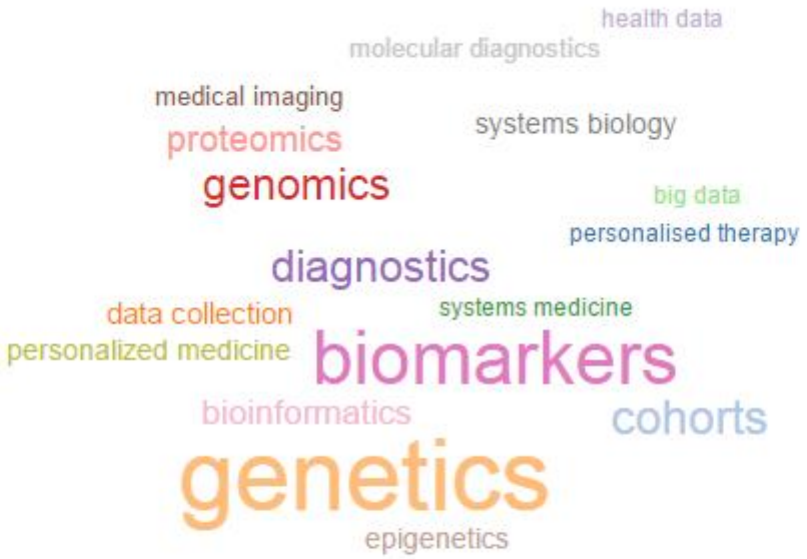


Figure 83 - '*Personalised Medicine'* wordcloud

3.2.21 Topic '*Pharmacology'*

The CTM model has classified 120.338 papers into '*Pharmacology*'.

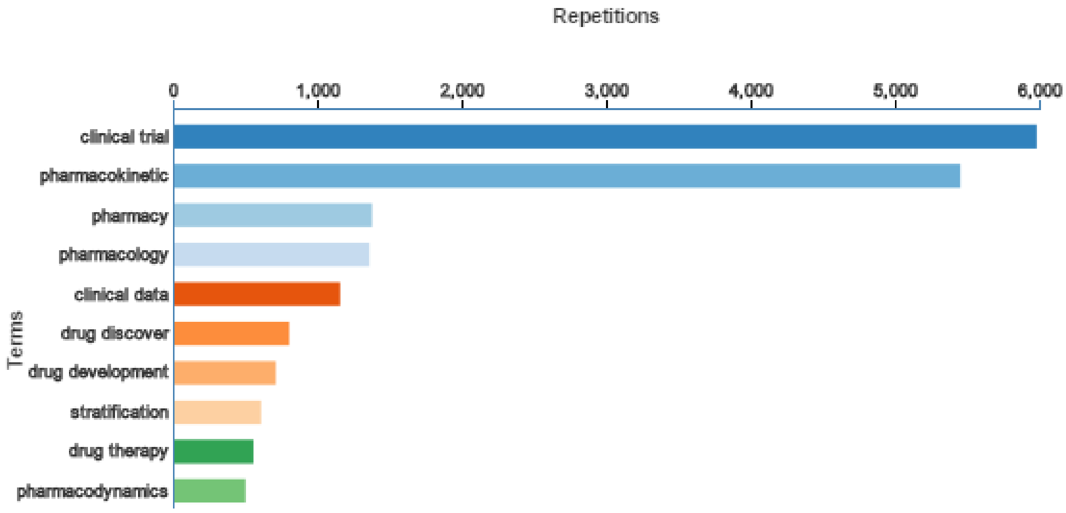


Figure 84- Ten of the most frequent terms in *'Pharmacology'*

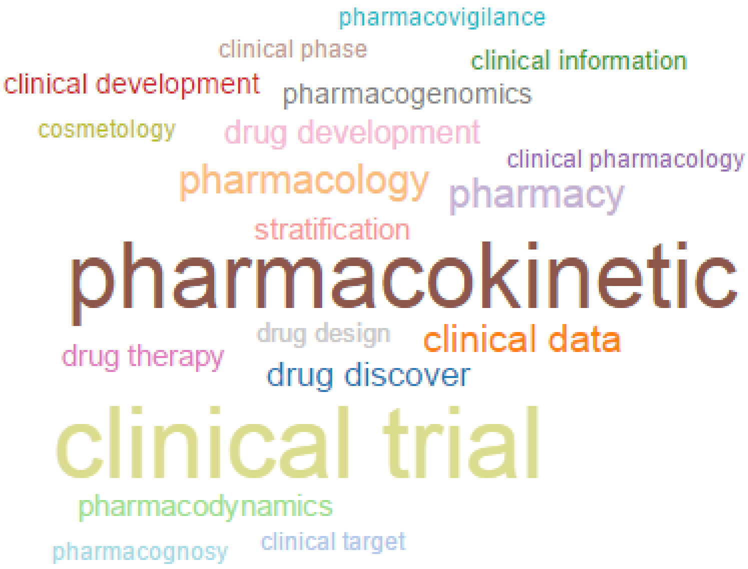


Figure 85 - '*Pharmacology*' wordcloud

3.2.22 Topic '*Physiology'*

The CTM model has classified 6.926 papers into '*Physiology*'.

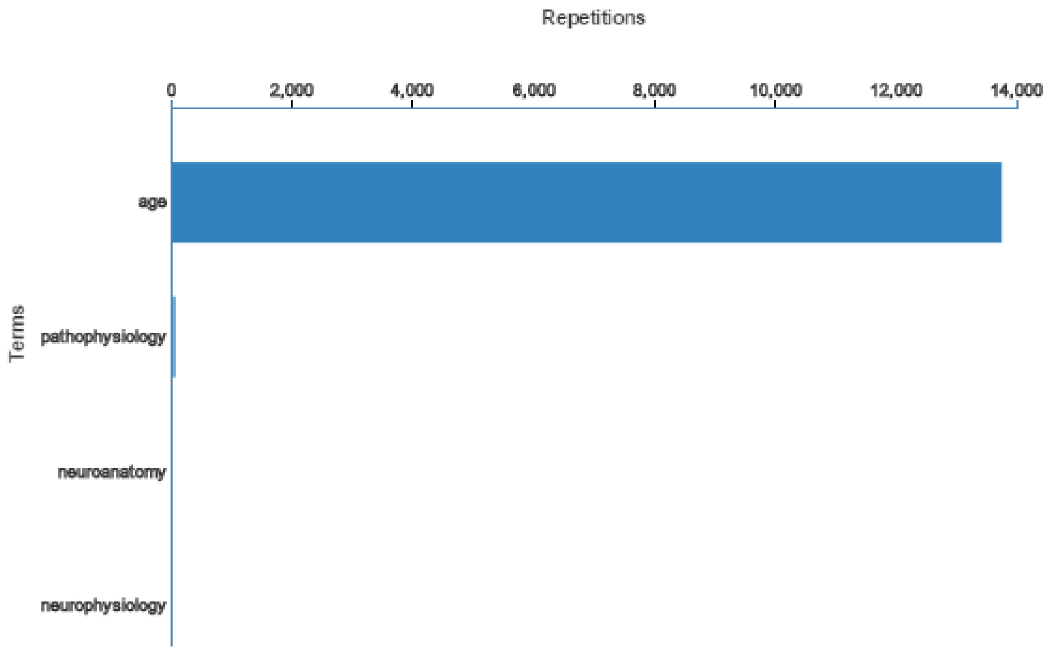


Figure 86- Ten of the most frequent terms in *'Physiology'*

age
pathophysiology
neurophysiology
neuroanatomy

Figure 87 - '*Physiology'* wordcloud

3.2.23 **Topic '*Psychiatry'***

The CTM model has classified 316.641 papers into '*Psychiatry*'.

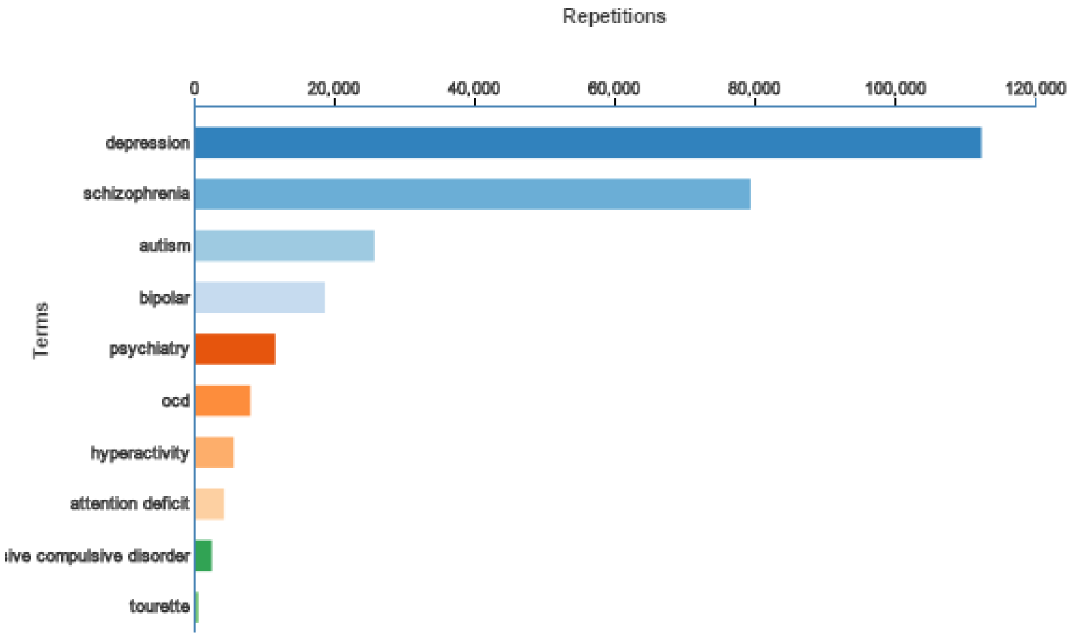


Figure 88 - Ten of the most frequent terms in *'Psychiatry'*

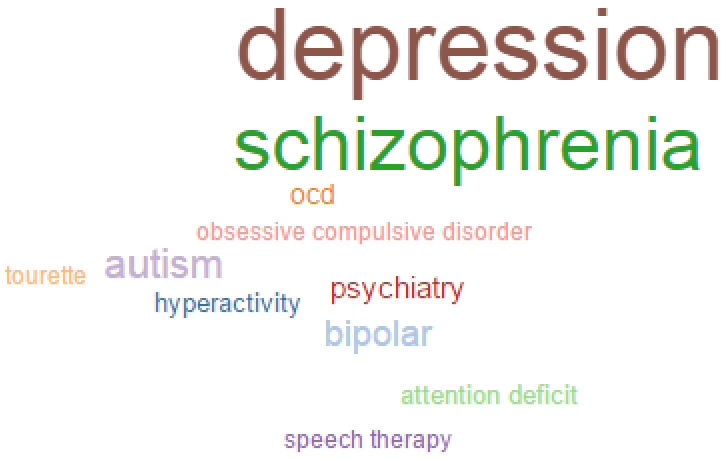


Figure 89 - '*Psychiatry*' wordcloud

3.2.24 **Topic '*Public Health'***

The CTM model has classified 419.709 papers into '*Public Health*'.

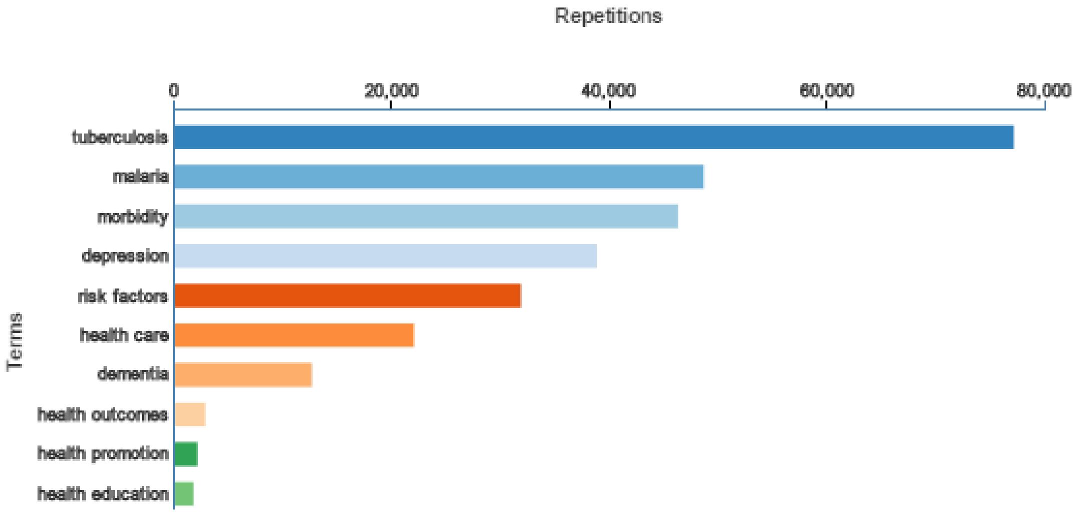


Figure 90 - Ten of the most frequent terms in *‘Public Health’*

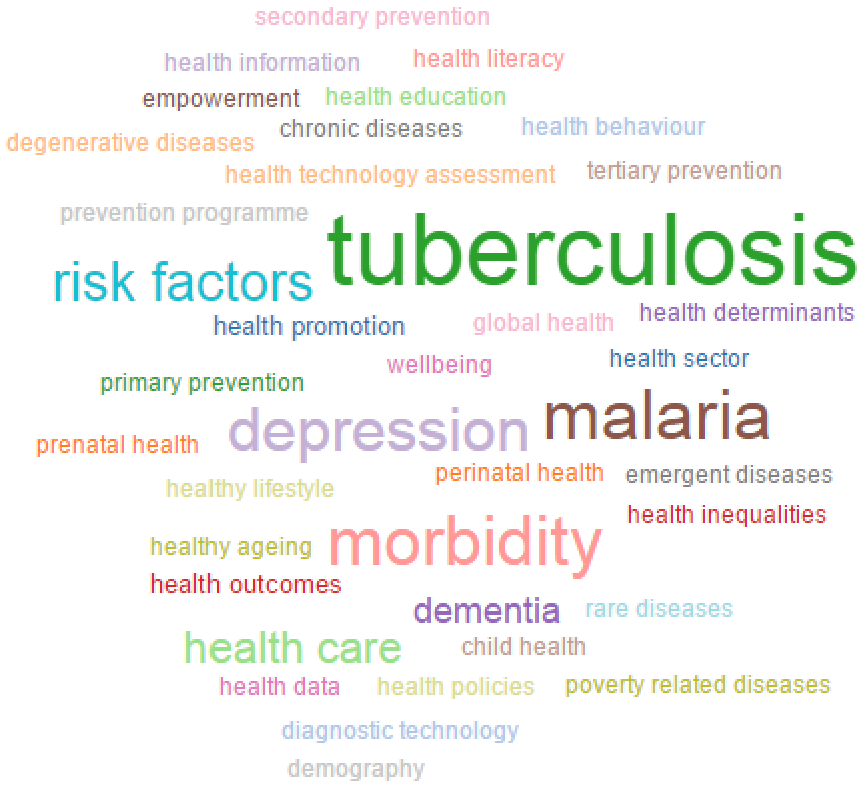


Figure 91 - ‘*Public Health’* wordcloud

3.2.25 Topic ‘*Regenerative Medicine’*

The CTM model has classified 112.896 papers into ‘*Regenerative Medicine*’.

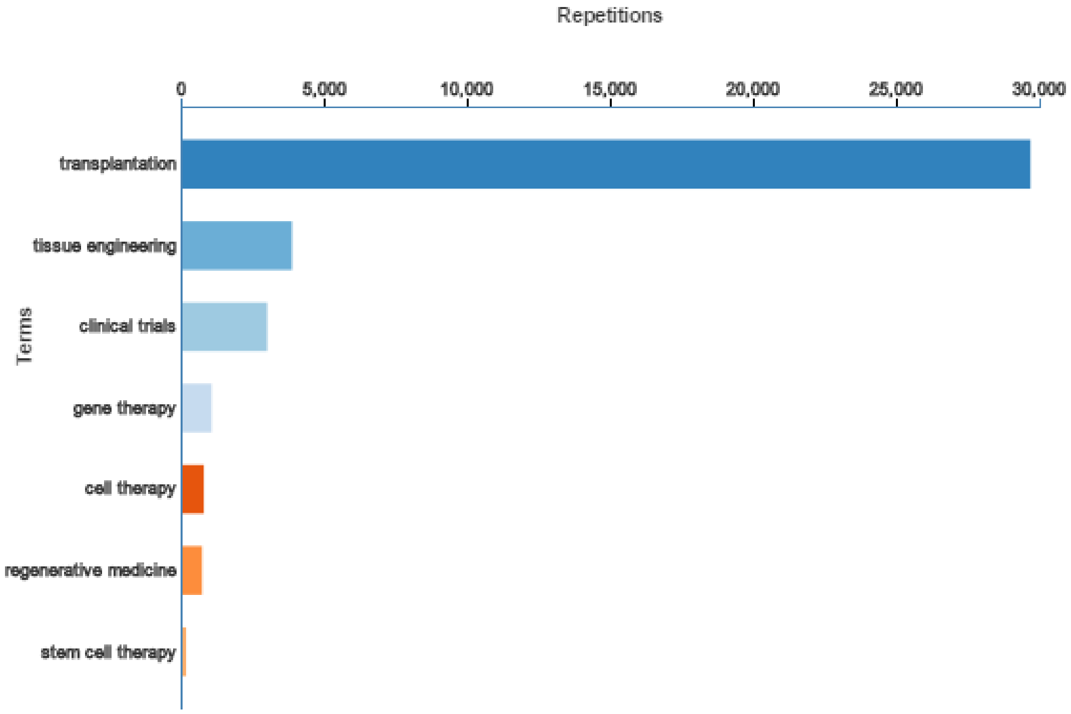


Figure 92 - Ten of the most frequent terms in *‘Regenerative Medicine’*

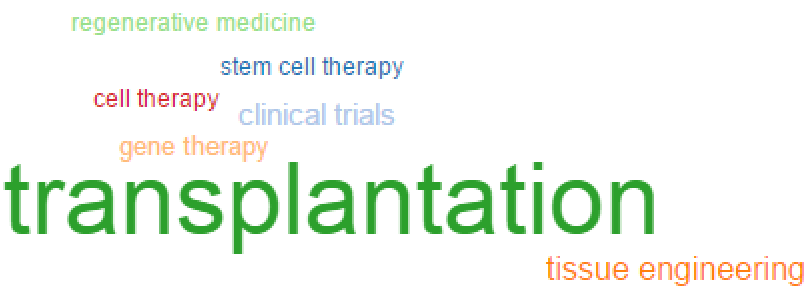


Figure 93- ‘*Regenerative Medicine’* wordcloud

3.2.26 **Topic ‘*Rheumatology’***

The CTM model has classified 315.496 papers into ‘*Rheumatology*’.

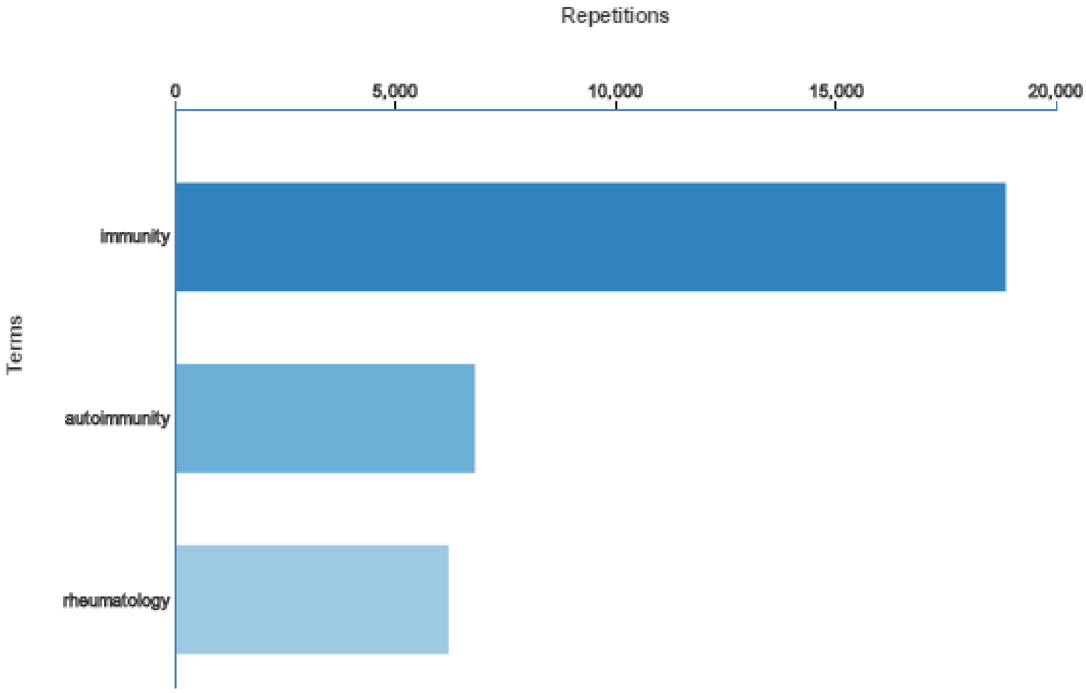


Figure 94 - Ten of the most frequent terms in *'Rheumatology'*

rheumatology
autoimmunity
immunity

Figure 95 - '*Rheumatology'* wordcloud

3.2.27 Topic '*Surgery'*

The CTM model has classified 236.828 papers into '*Surgery'*.

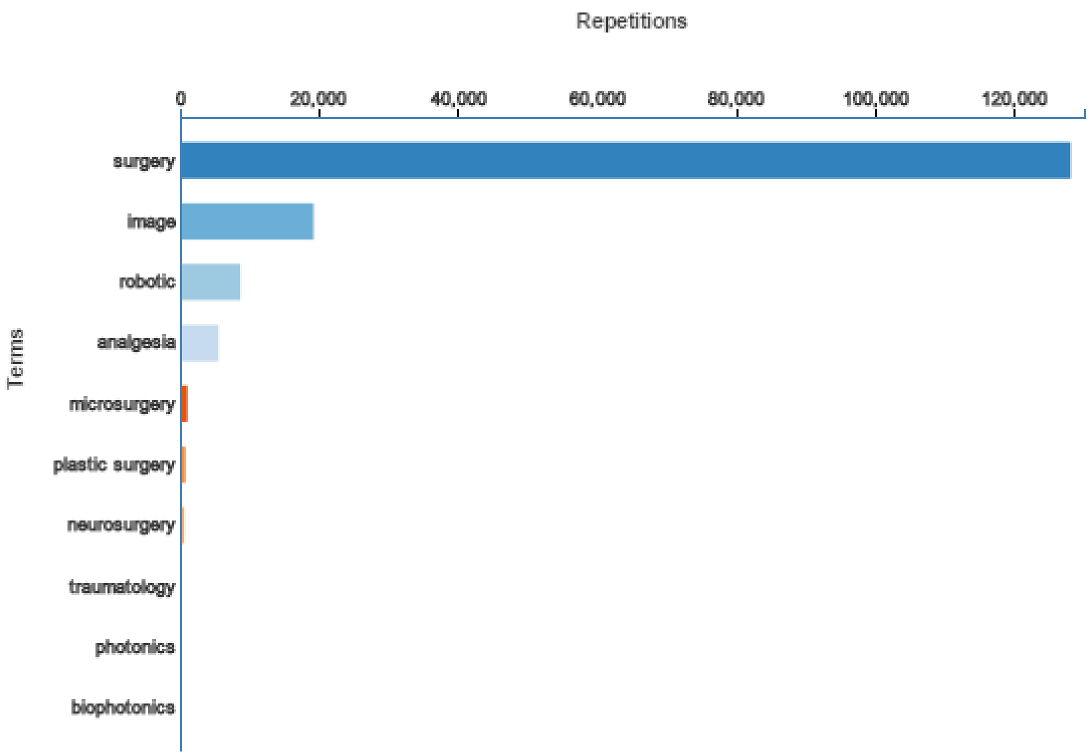


Figure 96 - Ten of the most frequent terms in *'Surgery'*

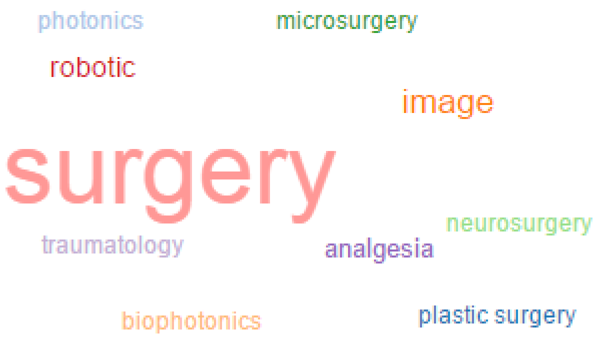


Figure 97 - '*Surgery'* wordcloud

3.2.28 **Topic '*System Biology'***

The CTM model has classified 917.091 papers into '*System Biology*'.

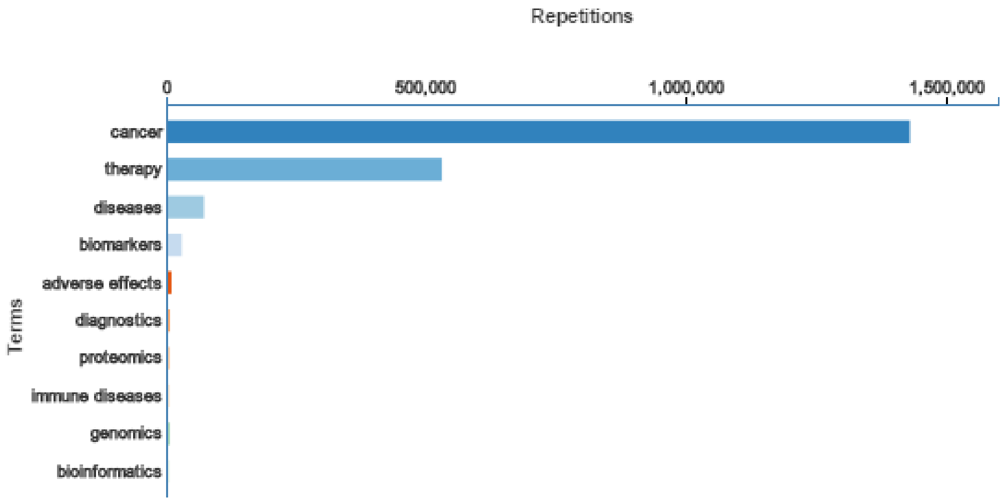


Figure 98 - Ten of the most frequent terms in *'System Biology'*

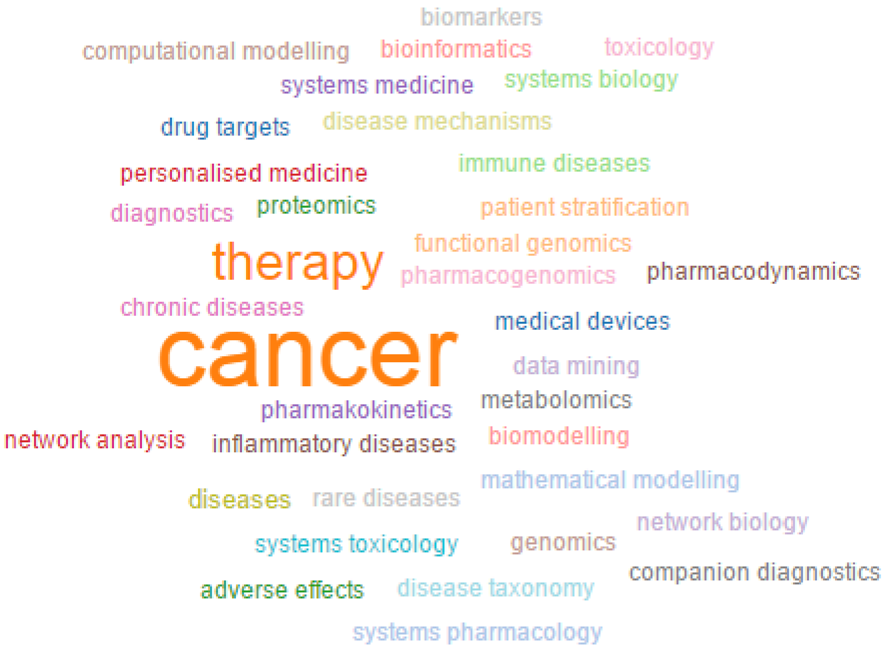


Figure 99 - '*System Biology'* wordcloud

3.2.29 Topic '*Toxicology'*

The CTM model has classified 121.354 papers into '*Toxicology*'.

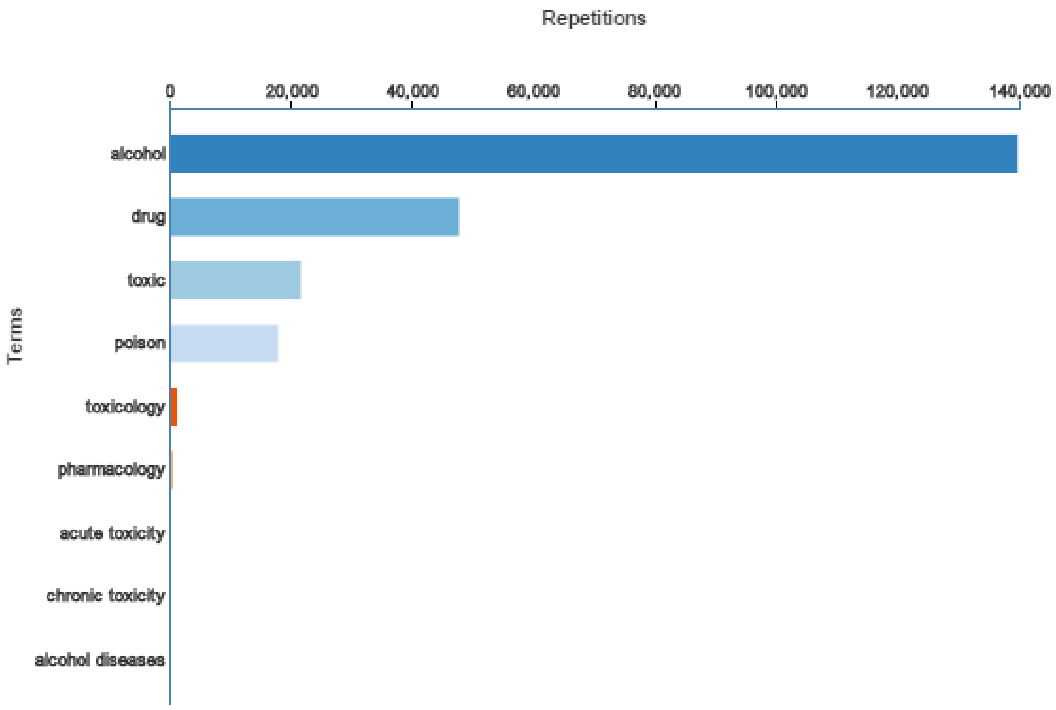


Figure 100 - Ten of the most frequent terms in ‘*Toxicology*’

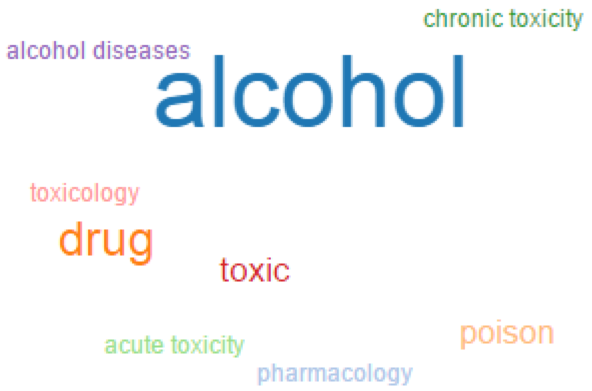


Figure 101 - ‘*Toxicology*’ wordcloud

3.3 Examples of abstracts classification

Finally, after the data cleaning process and the topic modelling, it is possible to see the complete process: from the original abstract to its topic.

The figure below shows three examples of abstracts before and after the application of the dictionaries and their final HRCS category assignment:

Original Abstract	Cleaned Abstract	Topic
Cholinesterase inhibitors (ChEIs) are used for symptomatic treatment of Alzheimer's disease. These drugs have vagotropic and anti-inflammatory properties that could be of interest also with respect to cardiovascular disease. This study evaluated the use of ChEIs and the later risk of myocardial infarction and death. The cohort consisted of 7073 subjects (mean age 79 years) from the Swedish Dementia Registry with the diagnoses of Alzheimer's dementia or Alzheimer's mixed dementia since 2007. Cholinesterase inhibitor use was linked to diagnosed myocardial infarctions (MIs) and death using national registers. During a mean follow-up period of 503 (range 0-2009) days, 831 subjects in the cohort suffered MI or died. After adjustment for confounders, subjects who used ChEIs had a 34% lower risk for this composite endpoint during the follow-up than those who did not [hazard ratio (HR) 0.66, 95% confidence interval (CI) 0.56-0.78]. Cholinesterase inhibitor use was also associated with a lower risk of death (HR: 0.64, 95% CI: 0.54-0.76) and MI (HR: 0.62, 95% CI: 0.40-0.95) when analysed separately. Subjects taking the highest recommended ChEI doses (donepezil 10 mg, rivastigmine >6 mg, galantamine 24 mg) had the lowest risk of MI (HR: 0.35, 95% CI: 0.19-0.64), or death (HR: 0.54, 95% CI: 0.43-0.67) compared with those who had never used ChEIs. Cholinesterase inhibitor use was associated with a reduced risk of MI and death in a nationwide cohort of subjects diagnosed with Alzheimer's dementia. These associations were stronger with increasing ChEI dose.	alzheimer inflammatory cardiovascular myocardial dementia alzheimer dementia alzheimer dementia myocardial donepezil alzheimer dementia	**Mental Health**
The RET receptor tyrosine kinase is crucial for normal development but also contributes to pathologies that reflect both the loss and the gain of RET function. Activation of RET occurs via oncogenic mutations in familial and sporadic cancers - most notably, those of the thyroid and the lung. RET has also recently been implicated in the progression of breast and pancreatic tumours, among others, which makes it an attractive target for small-molecule kinase inhibitors as therapeutics. However, the complex roles of RET in homeostasis and survival of neural lineages and in tumour-associated inflammation might also suggest potential long-term pitfalls of broadly targeting RET.	tyrosine oncogenic cancer thyroid lung breast pancreatic tumour homeostasis neural tumour	**Cancer**
As a result of collaborative efforts with international organizations and the salt industry, many developing and developed countries practice universal salt iodization (USI) or have mandatory salt fortification programs. As a consequence, the prevalence of iodine deficiency decreased dramatically. The United States and Canada are among the few developed countries that do not practice USI. Such an undertaking would require evidence of deficiency among vulnerable population groups, including pregnant women, newborns, and developing infants. Government agencies in the United States rely heavily on data from NHANES to assess the iodine status of the general population and pregnant women in particular. NHANES data suggest that pregnant women in the United States remain mildly deficient. This is important, because the developing fetus is dependent on maternal iodine intake for normal brain development throughout pregnancy. Professional societies have recommended that pregnant and lactating women, or those considering pregnancy, consume a supplement providing 150 Î¼g iodine daily. The United States and Canada collaborate on the daily recommended intake and are also confronted with the challenge of identifying the studies needed to determine if USI is likely to be beneficial to vulnerable population groups without exposing them to harm.	woman newborn woman woman fetus maternal brain pregnancy woman pregnancy	**Reproduction**

Figure 102 - Examples of topic assignment

4 ANNEX

4.1 Original abstract – transformed abstract – topic

The following excel includes a single sheet within the topic assignment for each cleaned abstract and its original text:

D03.03.Abstracts
classification_v1.0.xls