Gregory MS

Annual Review of 2021

https://gregory-ms.com/

Table of Contents

Foreword

For Doctors, Researchers, and People Living with MS.

This document is not a prescription, a recommendation, or any type of suggestion for People Living with Multiple Sclerosis (PwMS). It was not written by someone with any sort of medical degree, formal or informal training.

This document is meant as a practical demonstration on how to apply Artificial Intelligence (AI) and Machine Learning (ML) to research a very specific subject. It should be seen as an attempt to expedite innovation in the field and improve Quality Of Life (QOL) for humankind.

The described analysis is both statistical and subjective. The subjective analysis should be considered wrong and incorrect, meant solely as a demonstration of possible deductions. A scientist or doctor will surely be able to make a more informed interpretation and deduction from the data made available.

If this is so incorrect from a scientific perspective, why should it matter to you? Because this is a possible solution to the information overload we face.

While search engines have shown to be able to deliver relevant results to our daily searches they are still unable to organize and classify those results based on a broader personal context. In this case, finding the most recent research that will improve the quality of life of PwMS.

In the same way, a search engine considers that a piece of content is relevant given how much of its content matches the search query and how high a domain name ranks in its algorithm. Here, our approach is to focus on the websites that have shown to be quality sources of scientific information.

Gathering results is not enough. Over the course of a year we collected over 6 400 articles and it would be very hard to read, select and summarise it. This is where Gregory begins to shine. Using a Natural Language Processing (NLP) software we are able to extract the relevant phrases in a title and cross reference articles based on them. This gives us a snapshot of what was published during 2021 and is able to guide us in selecting future paths of research.

There were a lot of published papers that may not be relevant to improving the life of PwMS and could be considered “noise”. So again, we took steps to build a Machine Learning model that predicts if an article is relevant based on what it learned from the human selection. This becomes important when we are trying to reach a consensus regarding a new protein, or disease modifying therapy. (DMT)

In his current state Gregory is able to share weekly digests of findings and provides users with real time dashboards for a number of different use cases.

2022 is showing a lot of promise in giving the medical and scientific community a valuable and free tool to research MS. All that risks going to waste if we are unable to reach more users and to connect with other relevant stakeholders.

I hope you find the document interesting enough to reach out so that we can discuss what it signifies, what it could mean, and how it can be more useful.

Summary

Researching a complex topic like Multiple Sclerosis (MS) is not an easy task. It requires formal training and dedication to a routine. We begin by presenting a methodology for automatic research and the database structure that is being used to save that information.

With the assistance of Machine Learning (ML) and Artificial Intelligence (AI) we then attempt to provide an analysis of the information that was gathered over the course of 2021, starting from February.

We identify some issues that need to be addressed in future research; such as the need for a better dataset from qualified professionals, the room for improving the automatic categorisation of articles, the caveat of not being able to identify that a medication should be disregarded due to serious side effects, and also the need to be very focused in training a ML model just for one kind of criteria.

With this we are able to demonstrate that a system like this is useful to reduce the human cost of conducting regular research around a topic and that AI tools are at our reach.

Methodology

The first step was selecting the websites we should query for information. This was done by interviewing healthcare workers, neurologists, physiatrists, and physiotherapists, mostly. With their help we were able to gather a total of 11 websites to search for articles and 3 to monitor for new clinical trials.

Sources for Articles

Source	Example	Observations
APTA	https://www.apta.org/search?Q="Multiple+Sclerosis"+OR+"autoimmune+encephalomyelitis"+OR+encephalomyelitis+OR+"immune+tolerance"+OR+myelin&searcharticletypes=8834&searchconditionandsymptoms=&searchloc=APTA	
BioMedCentral	https://www.biomedcentral.com/search?searchType=publisherSearch&sort=PubDate&page=1&query=Multiple+Sclerosis	
FASEB	https://faseb.onlinelibrary.wiley.com/action/showFeed?ui=0&mi=2h5krp8&type=search&feed=rss&query=%26content%3DarticlesChapters%26field1%3DAllField%26publication%3D15306860%26target%3Ddefault%26text1%3DMultiple%2BSclerosis%2BOR%2Bautoimmune%2Bencephalomyelitis%2BOR%2Bencephalomyelitis%2BOR%2Bimmune%2Btolerance%2BOR%2Bmyelin	
JNeuroSci	https://www.jneurosci.org/search/text_abstract_title%3AMultiple%2BSclerosis text_abstract_title_flags%3Amatch-phrase exclude_meeting_abstracts%3A1 numresults%3A50 sort%3Apublication-date direction%3Adescending format_result%3Astandard	

(Cont.)

Source	Example	Observations
MS & Rel. Disorders	https://www.msard-journal.com/action/doSearch?text1=Multiple+Sclerosis&field1=AbstractTitleKeywordFilterField&startPage=0&sortBy=Earliest	
Nature.com		Contains one article that was added manually because it appeared to be relevant.
PEDro	https://search.pedro.org.au/advanced-search/results?abstract_with_title=Multiple+Sclerosis&therapy=0&problem=0&body_part=0&subdiscipline=0&topic=0&method=0&authors_association=&title=&source=&year_of_publication=&date_record_was_created=&nscore=&perpage=20&lop=or&find=&find=Start+Search	
Sage Pub	https://journals.sagepub.com/action/doSearch?AllField=multiple+sclerosis&SeriesKey=msja&content=articlesChapters&countTerms=true&target=default&sortBy=Ppub&startPage=&ContentItemType=research-article	
Scielo	https://search.scielo.org/?q=Multiple+Sclerosis&lang=en&count=15&from=0&output=site&sort=&format=summary&fb=&page=1&q="Multiple+Sclerosis"+OR+"autoimmune+encephalomyelitis"+OR+encephalomyelitis+OR+"immune+tolerance"+OR+myelin&lang=en&page=1	
The Lancet	https://www.thelancet.com/action/doSearch?text1="Multiple+Sclerosis"+OR+"autoimmune+encephalomyelitis"+OR+encephalomyelitis+OR+"immune+tolerance"+OR+myelin&field1=AbstractTitleKeywordFilterField&startPage=0&sortBy=Earliest	

Sources for Clinical Trials

In the same way that we gather papers and scientific articles we can also collect information of new clinical trials. This allows Gregory to notify subscribers of new trials in real time. To begin we are using these three sources.

Source	Example
ClinicalTrials.gov	https://clinicaltrials.gov/ct2/results/rss.xml?rcv_d=14&lup_d=&sel_rss=new14&cond=Multiple+Sclerosis&count=10000
CUF	https://www.cuf.pt/cuf-academic-center/ensaios-clinicos?combine=&unidade=&estado=All&patologia=2346&especialidade=
Novartis	https://www.novartis.com/clinicaltrials/recruiting-trials?title=multiple sclerosis

Search terms

For consistency, articles were searched with the same set of keywords while clinical trials were filtered for Multiple Sclerosis. For all sources we used this list of search terms. We opted for this list with the criteria that it should be broad enough to gather as many papers as possible in the field of MS without going outside the scope of the disease.

Multiple Sclerosis, autoimmune encephalomyelitis, encephalomyelitis, immune tolerance, myelin

Clinical Trials, by nature, are easy to search based on “Multiple Sclerosis”.

Database

For the database we are using 4 tables: **Articles, Trials, Categories** and **Rel_articles_categories**. Each paper found is stored in the Articles table and can be assigned one or more categories. The rel_articles_categories table stores this many-to-may relationship.

The Trials table lists the following information for every Clinical Trial:

1. discovery_date
2. title
3. summary
4. link
5. published_date
6. source

For the Articles we are listing the following data:

1. article_id
2. title
3. summary
4. link
5. published_date
6. source
7. relevant
8. ml_prediction gnb
9. ml_prediction lr
10. discovery date
11. noun phrases

Most of these fields are self explanatory, except for 7, 8, 9, and 11.

Relevant indicates a human selection that the article provides information that may be beneficial to improve the quality of life of PwMS.

ml_prediction_gnb refers to the result of a Gausian Naives Bayes Algorithm Machine Learning Algorithm.

ml_prediction_lr also stores the result of a Machine Learning Algorithm.

Noun Phrases is the result of applying natural language processing to the titles. The typical result is a list of phrases that functions in a sentence as subject, object, or prepositional object. This information is then used to cross reference papers among themselves. Further ahead we will describe these concepts in greater detail.

In order to avoid duplicate entries, the database was configured in such a way as not to allow two articles or clinical trials to have the same title. This allows us to cover the majority of cases but is not completely full proof as one character is enough to avoid this rule.

There is also a table called **categories** where we save topics of interest that are matched with the ML and AI ouput. For those we just store the category name and unique ID number.

Database maintenance

Gregory runs a series of database maintenance tasks meant to populate the database in the best possible way. For example, when an article is found without an abstract, it is saved in the database and re-visited at a later time to analyse the content in more detail. In a similar way, ML tasks require more processing and run at intervals to save the output in the database.

Articles are always stored in the database without a category assigned, that is also a maintenance task. Once we save the output of the ML and AI tasks, another task is executed to clear and rebuild the match between articles and categories.

There is a small change that a person may query the database at a time when these tasks are being executed and therefore some values may be blank.

Can computers think?

What is Machine Learning and Artificial Intelligence

Before we proceed it is important to clarify what Machine Learning (ML) and Artificial Intelligence are and how we use them in this context.

Both concepts are close to each other which leads to some confusion. Machine Learning is an algorithm that is trained with sample data and that outputs a prediction when we give it new values or new items to analyse.

Artificial Intelligence (AI) on the other hand goes one step further. My definition is that an AI is a system that is able to gather data on its own, act upon it, and communicate that information to humans. There is a blog post explaining this in more detail that I wrote in 2017.

Gregory uses both. There are ML algorithms that fetch information and others that were trained to identify relevant papers based on a previous human selection. There is also an Artificial Intelligence that summarises the contents of the database. With other modules we are making this information available to everyone, by email, and through twitter using the relevant hashtags: #CNSinfections, #EM, #EscleroseMultipl, #MS, #MultipleSclerosis, #NeuroTwitter

There are more complex systems that can be used for diagnosis and to find correlations between different sets of data. Those are out of our reach given the available resources.

Should you want to explore the possibilities of AI for research, you can look into some tools that are already available.

Connected Papers lets you visualize a network of citations.
Connected Papers | Find and explore academic papers

Elicit is a search engine that starts with your question and turns into a research tool. It can help you find randomised control trials (RCT) or papers that aggregate and review findings (review, systematic review, meta-analysis).
elicit.org

Research Rabbit is a tool that will learn from your searches as well as build graphs to visualize the network of authors and papers.
ResearchRabbit

Like Gregory, these tools are meant to assist your research and given their state of maturity I am confident in recommending them at this stage.

To the question "can computers think" all we can respond is no. They can be valuable allies in researching and summarising what we need to know, and they can automate the boring tasks for us. We should be looking at AI as the creation of agents, or assistants that free up our time and energy for the analysis and building of knowledge.

How we use Machine Learning

As mentioned before, Machine Learning is an algorithm that is trained with sample data and that outputs a prediction when we give it new values or new items to analyse.

In practice, what we did was read every title and abstract that Gregory was able to collect and tag the ones that showed promise in improving the quality of life for PwMS. Although 6 000 articles may seem like a lot, it is in fact a small training set. Of those, only about 3 to 4% seemed relevant, and to be on the cautious side, very few were tagged as not being relevant at all.

Gregory contains two algorithms, a Gaussian Naive Bayes (GNB) and a Logistic Regression model. We tested both and the GNB model proved to be the most accurate. However, we are keeping both of them and saving their output in the database under the fields **ml_prediction_gnb** and **ml_prediction_lr** respectively.

It is possible to download the entire database in excel format to analyse just how close or far the human tagging of articles was.

https://gregory-ms.com/downloads/

To improve the Machine Learning algorithm we would have to be assisted by a neurology team able to agree on a criteria for relevancy and tag more items as relevant. The new model could then be trained and tested to provide more accurate predictions.

The volume of research

How much do we publish and where?

As I write this line, there are **6 466** papers in Gregory's database and more may appear during the last day of 2021. As the world goes on to celebrate the end of 2021 he will be running the same research tasks that began 310 days ago. This means that everyday he has found an average of 20 new papers. Imagine it takes a human 1 hour per day to run the search across the different websites, and 10 to 15 minutes to read each abstract, and save it to a spreadsheet. This would signify more than 21 700 hours of work. We are publishing too much for a single person to be able to keep up.

Let's take a look at the pace and place where we found papers matching our search criteria.

Articles published over time

For clarity we are using "Discovery date" for the timeline. We observed that in the majority of cases the published date and discovery date can be considered almost identical, with maybe one or two days difference. The small spike in results seen in the beginning of March reflects new sources added. Otherwise, the database grew at a steady pace.

December 31st Gregory had indexed **6 494 articles**.

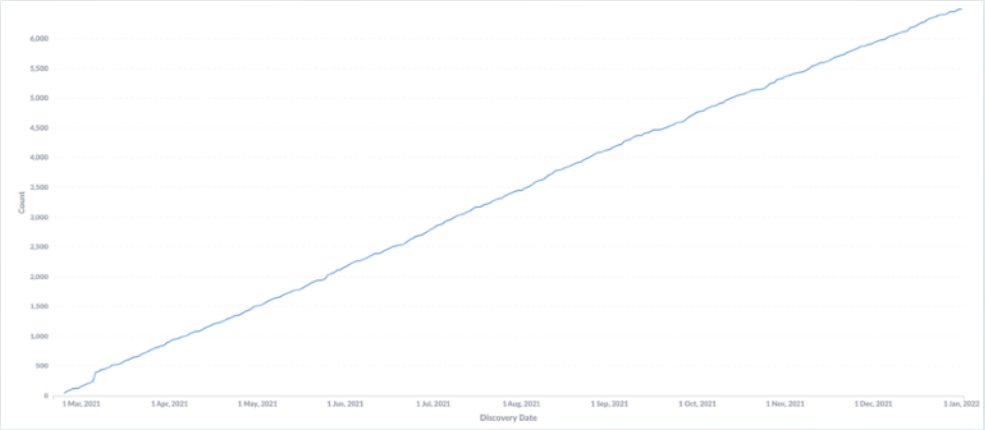


Figure 1.

Source: https://metabase.gregory-ms.com/question/30-full-articles-table-timeline

Distribution of articles per source

It comes as no surprise that PubMed contains the most papers and the same time, of all the sources we use, it is the one that allows more access to information. Not only through a syndication feed (RSS) but also through a number of open source tools and datasets made available on GitHub.

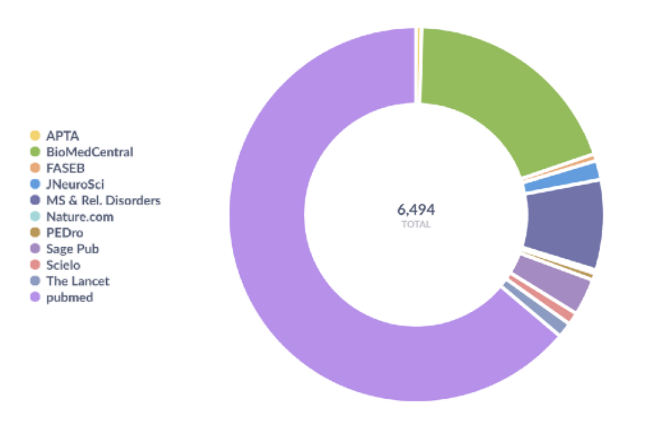


Figure 2.

Source: https://metabase.gregory-ms.com/public/dashboard/5b82d03f-5d6c-4ccc-b059-f30b280fa7d5?date_filter=2021-01-01~2021-12-31

Source	Number of Articles
APTA	13
BioMedCentral	1273
FASEB	21
JNeuroSci	92
MS & Rel. Disorders	502
nature.com	1
PEDro	21
Sage Pub	193
Scielo	52
The Lancet	68
pubmed	4258

Distribution of articles per source

By relevant we mean to answer if a paper provides information to improve the quality of life of PwMS. By itself this statement contains a lot of caveats. The most obvious is possible human error and lack of knowledge to fully understand the academic papers and their potential. The second one is about what constitutes a better quality of life.

For some patients this includes better mobility, for others this would be repairing the central nervous system. We tried to focus on these two and anything else that could improve the remyelination process. Relevancy is not a simple metric and in this case it was the result of a subjective analysis of titles and abstracts. In the future this metric could be made more reliable with the support of a qualified neurology research capable of classifying the collected information and train the ML algorithm to do a better job.

Accepting the previous argument for lack of a better one, the next question would be if different sources provide a higher percentage of relevant articles.

As of this writing, the number of relevant articles, either by ML or Human selection, was 240. In a total of **6494** papers we are seeing 3.7% of all literature pointing towards better quality of life for patients.

Regarding accordance, the human and the algorithm agreed on 66% percent of the cases.

Selection	Number of Articles	Percentage
Human	**221**	**100%**
Machine Learning GNB Model	**165**	**74,6%**
Human and ML GNB Model	**146**	**66,0%**
Data from December 31 2021		

While 66% might sound like a low number, we must take into account the small sample of data available to us and that the margin of error is good enough for Gregory to be autonomous in gathering and distributing information.

And how does relevancy correlate with sources? We are able to ascertain that the human and the ML algorithm have different opinions regarding APTA, for example. While the human didn't consider any information relevant, the algorithm found about 30% of relevant articles.

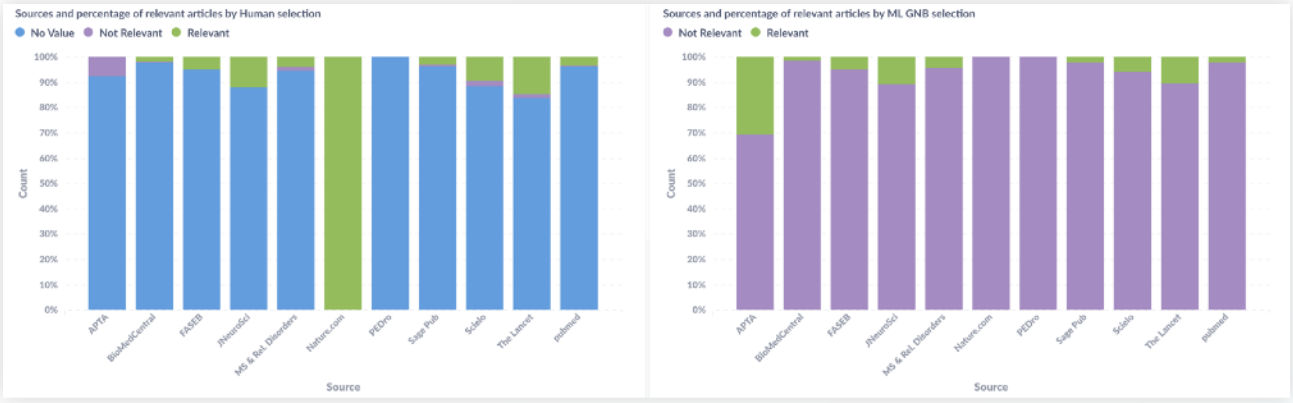


Figure 3.

Source:https://metabase.gregory-ms.com/public/dashboard/5b82d03f-5d6c-4ccc-b059-f30b280fa7d5?date_filter=2021-01-01~2021-12-31

Figure 3 is showing Nature.com as relevant by the human selection and not relevant by the algorithm. This happened because it was a manual insertion in the database that can be found at https://www.nature.com/articles/s41586-021-03892-7. In the same way, there was only one article discovered on FASEB and it showed to be relevant. (https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002465R)

Summarising over 6400 articles

As we mentioned before, Gregory contains a Natural Language Processing (NLP) module, it's called spaCy and can be found at https://spacy.io. The features offered by spaCy help us turn sentences into something that a computer can understand and quantify. For example, spaCy can look at the sentence "Apple is looking at buying U.K. startup for $1 billion" and identify Apple as an organisation, UK as a geo-political entity, and the monetary value that is $1 bilion.

Our use of this NLP was very straightforward. We wanted a way to cross reference papers based on the information of their titles. To do this we used the Noun Phrases, or Noun chunks as described by the documentation:

> **Noun chunks are "base noun phrases" – flat phrases that have a noun as their head. You can think of noun chunks as a noun plus the words describing the noun – for example, "the lavish green grass" or "the world's largest tech fund".**

Source: https://spacy.io/usage/linguistic-features#noun-chunks

Each time there are new articles, Gregory will run a maintenance task to identify noun phrases and populate this field.

This proved to be quite effective, as we can see from the queries below:

```
sqlite> select count(article_id) from articles where noun_phrases like '%ocrelizumab%';
69
sqlite> select count(article_id) from articles where title like '%ocrelizumab%';
73
sqlite>
```

What we did was query the database to give us a count of all articles that had ocrelizumab in the noun_phrases field, returning a count of 69. For the same query using the title field, we had a count of 73.

Some examples on titles that returned the search term in the noun phrases field.

Challenges of persons with multiple sclerosis on ocrelizumab treatment during COVID-19 pandemic

Noun Phrases in Title

- Challenges
- persons
- multiple sclerosis
- ocrelizumab treatment
- COVID-19 pandemic

CD19 B cell repopulation after ocrelizumab, alemtuzumab and cladribine: Implications for SARS-CoV-2 vaccinations in multiple sclerosis

Noun Phrases in Title

- ocrelizumab
- alemtuzumab
- cladribine
- SARS-CoV-2 vaccinations
- multiple sclerosis

Longitudinal humoral response after SARS-CoV-2 vaccination in ocrelizumab treated MS patients: To wait and repopulate?

Noun Phrases in Title

- Longitudinal humoral response
- SARS-CoV-2 vaccination
- ocrelizumab
- MS patients

Evidence of extensive cellular immune response after SARS-CoV-2 vaccination in ocrelizumab-treated patients with multiple sclerosis

Noun Phrases in Title

- Evidence
- extensive cellular immune response
- SARS-CoV-2 vaccination
- ocrelizumab-treated patients
- multiple sclerosis

It would be desirable to have more tests with a larger volume of data, but for the time being it proves enough efficacy for us to use noun phrases as a way to determine categories, or tags, in the database.

Tools like spaCy bring us the possibility to do a lot more, like what we see from search engines like Elicit that can take a simple english question and return academic results that lead us to an answer.

A simpler approach would be to gather the abstracts of every article mentioning Ocrelizumab and summarise them. Kamal Khumar as published an example on how this can be done, and we plan exploring it during this year

For this year we opted for a simpler and more familiar approach to summarise the current state of MS research. Figure 4 shows a word cloud generated from the results of the noun phrases analysis.

We could have gone further in this analysis and in building the word cloud, but given that this is the first year review and the resources available, this was the most realistic output. The words were taken from the noun phrases mentioned early in order to remove any common english words and give a more accurate result.

Figure 4.

Word cloud of the noun phrases for all the articles discovered by Gregory.

Further analysis could select the most important noun phrases and track their development over time, and configure real time alerts for new articles based on them.

There are also other tools that can be explored and put into practice to help the medical field. The biggest barrier to making them a part of the research method is the divide between IT Developers and professionals in the medical field. We could also argue that there are other obstacles such as lack of resources, and difficulty in finding a way to maintain these new tools. Further on, I plan to demonstrate what building Gregory required and thus refute this argument.

https://github.com/medspacy/medspacy

https://github.com/explosion/healthsea

Putting it to practice

Tracking topic across the MS field of research

Close to the end of the year we also used the noun phrases to populate a category field, where each article gets attributed one or more categories. The selection of categories was made based on what was posted on the MS Society page and that matched our relevancy criteria. In some cases we made exceptions, such as for COVID-19, and Tolebrutinib based on an article for phase 2 trials discovered.

This resulted in a list of 11 categories, but 2 of them did not return any results: Simvastatin and Tcelna. Double checking the full database, none of these terms were present in the title.

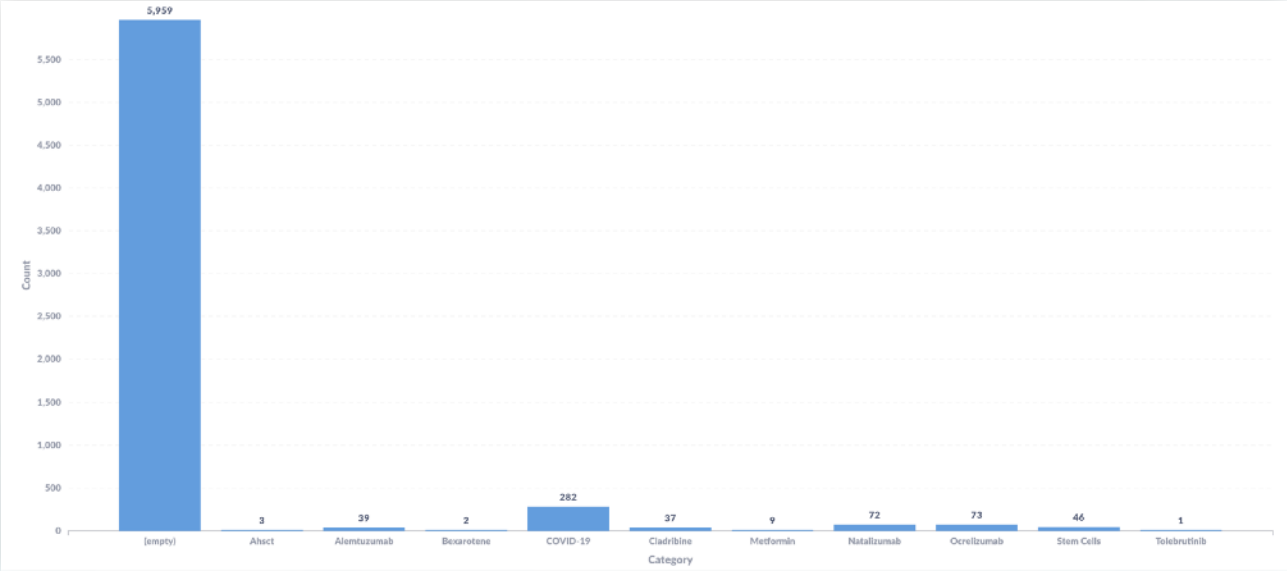


Figure 5.

Source https://metabase.gregory-ms.com/public/dashboard/5b82d03f-5d6c-4ccc-b059-f30b280fa7d5?date_filter=2021-01-01~2021-12-31

We also can't ignore the fact that over 90% of the articles did not match any of the categories. In the future we can calculate the frequency of words in noun phrases to suggest new categories.

With the information we have now, it is also possible to look at which categories show more relevant articles.

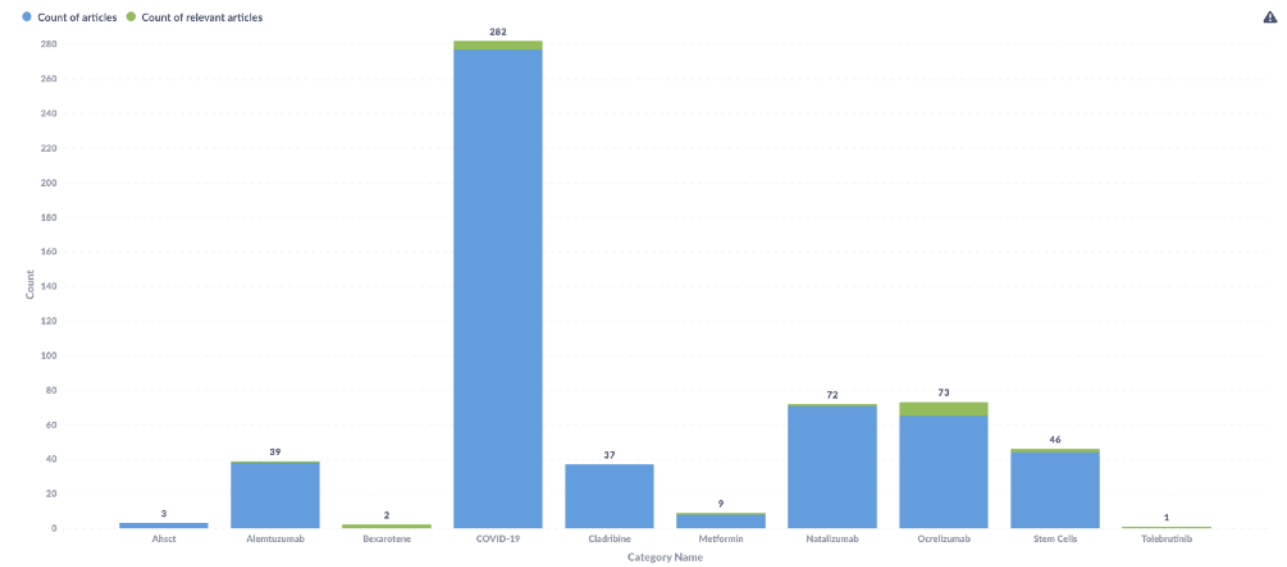


Figure 6.

Count of articles per category and relevancy according to either Human Selection or the Machine Learning Algorithm.

Let's take this further. For each of these categories we are able to produce their own dashboard.

Mesenchymal and Autologous Hematopoietic Stem Cells

This particular topic may generate confusion between people without medical training, such as me. In our perspective there is only one type of stem cells. The peril here begins if a person researches the topic with the keywords "benefits of stem cells" only to see a myriad of good news and optimism regarding their use, unaware if it applies to their condition or what sort of therapy it implies.

The same thing easily happens here, where a search for stem cells can return a lot of relevant articles but refer to two different therapies.

From the MS Society website, regarding Autologous Hematopoietic Stem Cells (AHSCT):

> **AHSCT is a type of bone marrow transplantation that attempts to reset the immune system, which is responsible for damaging the brain and spinal cord in MS. The stem cells used during the procedure are 'autologous' meaning they are derived from your own body.**

Looking at Gregory's database we find the following results.

Figure 7.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Ahsct

Given the potential of this therapy it seems odd to see such a low number of articles, reviewing the database we came to the conclusion that the system was being too strict focusing on the abbreviation of the therapy. We added the category "Stem Cells" and were able to collect more data.

Figure 8.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Stem Cells

Both charts can then be combined for a more complete analysis of what is the volume of research regarding stem cell therapies.

Figure 9.

Data from the category “Stem Cells” combined with the category aHSCT.

But what is really important is how many of these articles are relevant in this context. Those have been tagged with green lines in both charts. We can also look into the accordance between the ML algorithm and the human.

Relevant article by human selection Relevant articles by Machine Learning Relevant articles where ML and Human agree

Relevant article by human selection	**9**
Relevant articles by Machine Learning	**2**
Relevant articles where ML and Human agree	**2**

There are two hypothesis for these results. Either the algorithm is working and the human is biased towards considering this therapy effective, or the algorithm requires a bigger data set to serve as a training model.

Both can also be true, and as we have been stating, the more data we have the better the model will work. To help the reader draw their own conclusions we are sharing this subset of the database in the file **Stem Cells+aHSCT query result 2022-01-01T12 10 16. 634252Z.xlsx.**

Stem Cells+aHSCT query_result_2022-01-01T12_10_16.634252Z.xlsx

Alemtuzumab

> "Alemtuzumab, sold under the brand names Campath and Lemtrada among others, is a medication used to treat chronic lymphocytic leukemia and multiple sclerosis. In CLL, it has been used as both a first line and second line treatment. In MS it is generally only recommended if other treatments have not worked." Wikipedia

Given the low number of results we were able to gather, it doesn't feel right to provide any sort of analysis for Alemtuzumab.

Figure 10.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Alemtuzumab

Relevant article by human selection	**2**
Relevant articles by Machine Learning	**1**
Relevant articles where ML and Human agree	**1**

Bexarotene

> "Bexarotene is used to treat skin cancer. It was also tested as a treatment for MS in a Phase 2 clinical trial. The results of this trial showed bexarotene could repair myelin in humans. But trial participants experienced some serious side effects so bexarotene will not be taken forward into a Phase 3 study." MS Society.

The search for Bexarotene poses an important question, although the system identified it as being relevant, in one of the two articles discovered it was made clear that the medication had serious side effects and should not be pursued.

With access to the abstract and using NLP, can we identify that a medication is not recommended as treatment ?

Looking at tools such as Elicit, it seems that this is not impossible. Their system already provides search results together with possible counter arguments to its conclusions. Although Gregory could be used to conduct such an analysis we will not be putting effort into developing this feature without solid support from a team of neurologists.

Figure 11.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Bexarotene

Relevant article by human selection	**2**
Relevant articles by Machine Learning	**2**
Relevant articles where ML and Human agree	**2**

Cladribine

> "Cladribine is a disease modifying therapy (DMT) for relapsing MS. Its brand name is Mavenclad and you take it as a tablet." MS Society.

This is already a medication that is approved and used to treat MS, and yet over the course of the year we didn't find any articles that could be marked as relevant to improve symptoms or QOL for PwMS.

We then need to be honest and say that maybe our criteria for relevant is not the most appropriate, and that the human understanding of the 37 articles discovered was not enough to identify the benefits.

Figure 12.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Cladribine

Relevant article by human selection	**0**
Relevant articles by Machine Learning	**0**
Relevant articles where ML and Human agree	**0**

Covid-19

While not a topic presented in the MS Society website, COVID has been an important subject for everyone. Here the criteria was skewed to include information on the effectiveness of the vaccines for PwMS.

T-cell responses to SARS-CoV-2 in multiple sclerosis patients treated with ocrelizumab healed from COVID-19 with absent or low anti-spike antibody titers - PubMed

Mesenchymal stem cell-based therapy and exosomes in COVID-19: current trends and prospects - Stem Cell Research & Therapy

Oral administration of methylprednisolone powder for intravenous injection dissolved in water to treat MS and NMOSD relapses during COVID-19 pandemic in a real-world setting

The prevalence of COVID-19 infection in patients with multiple sclerosis (MS): a systematic review and meta-analysis - PubMed

The impact of COVID-19 home confinement on neuromuscular performance, functional capacity, and psychological state in Spanish people with Multiple Sclerosis

While this might feel like a worthy endeavour, given the access to the database and the ML tools, it may actually have been prejudicial to the training of the algorithm. By tagging content related to covid-19 we may have added noise to our training data. Like a human, an AI's work is better when it's focused on a single task and mindset.

Figure 13.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=COVID-19

Relevant article by human selection	**5**
Relevant articles by Machine Learning	**5**
Relevant articles where ML and Human agree	**5**

We need to take into account an important detail, this category was used to find articles where the noun phrases contained COVID-19. Other articles where the noun SARS-CoV-2 was used did not qualify to be included. This is a caveat that can happen regarding other nouns, such as Ocrelizumab and Ocrevus, or Natalizumab and Tysabri. This is an issues that we are not able to resolve at this time but will be added to the issues list on GitHub.

Metformin

> "Metformin is used to treat type 2 diabetes. It is now being tested in combination with clemastine (a hay fever drug) as a myelin repair treatment for MS." MS Society

This is a clear example on how Gregory can be used to track the potential of a medication. Over the course of 2021 Gregory was able to find 9 mentions of the drug and the dots show us how far apart in time they were discovered. A good use case would be to provide an email alert in real time, together with the abstract and the ML prediction on whether it is relevant or not. A mere 9 articles published over the course of a year do not warrant much effort in keeping track, but for Gregory it is just another day in its daily routine.

Figure 14.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Metformin

Relevant article by human selection	**2**
Relevant articles by Machine Learning	**1**
Relevant articles where ML and Human agree	**1**

Natalizumab

Also known as TYSABRI.

"(...) is a prescription medicine used to treat relapsing forms of multiple sclerosis (MS), to include clinically isolated syndrome, relapsing-remitting disease, and active secondary progressive disease. TYSABRI increases the risk of progressive multifocal leukoencephalopathy (PML). When starting and continuing treatment with TYSABRI, it is important to discuss with your doctor whether the expected benefit of TYSABRI is enough to outweigh this risk." Tysabri

For the second time we are seeing a case of a drug that is already recognised treatment for MS and however, by itself, Gregory wasn't able to find relevant articles.

In this case, our hypothesis is that the articles present in the findings do not mention improvement of QOL for PwMS and therefore didn't make into the ML selection. To facilitate the confirmation of this hypothesis we are attaching to this report the file **Natalizumab query result 2022-01-01719 00 45.162831Z.xlsx** that contains the articles used to build the charts below.
Natalizumab query_result_2022-01-01T19_00_45.162831Z.xlsx

Figure 15.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Natalizumab

Relevant article by human selection	**1**
Relevant articles by Machine Learning	**1**
Relevant articles where ML and Human agree	**1**

Ocrelizumab

Also known as Ocrevus.

> "Ocrevus is indicated for the treatment of adult patients with early primary progressive multiple sclerosis (PPMS) in terms of disease duration and level of disability, and with imaging features characteristic of inflammatory activity (see section 5.1)." EMA

Figure 16.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Ocrelizumab

Relevant article by human selection	**6**
Relevant articles by Machine Learning	**8**
Relevant articles where ML and Human agree	**5**

This is an interesting case, where the ML algorithm believes the human should have considered 3 other papers as being relevant. We have listed them below. One of them refers to the response of the medicine with the COVID vaccines, while the two others focus on safety of this medication.

There are many hypothesis on why this happened. On the human side, one could have been fatigue from seeing too much information regarding covid or simple distraction. Regarding the safety of the medication, the human may have disregarded them for being too focused on what improves the QOL. To this we can argue that a safe medication with an administration time of once every six months should be considered as an improvement in QOL.

Safety of Ocrelizumab in Patients With Relapsing and Primary Progressive Multiple Sclerosis - PubMed

Longer-term Safety of B-Cell Therapy With Ocrelizumab in Multiple Sclerosis - PubMed

Evidence of extensive cellular immune response after SARS-CoV-2 vaccination in ocrelizumab-treated patients with multiple sclerosis - Neurological Research and Practice

Given the volume of results for this medication, we are attaching the corresponding database export to this report with the filename **Ocrelizumab query result 2022-01-01721 21 33.915845Z.xlsx**

Ocrelizumab query_result_2022-01-01T21_21_33.915845Z.xlsx

Tolebrutinib

The only information we could find was a paper discovered by Gregory on August 19, published on The Lancet.

According to the information on that link:

> **"Tolebrutinib is an oral, CNS-penetrant, irreversible inhibitor of Bruton's tyrosine kinase, an enzyme expressed in B lymphocytes and myeloid cells including microglia, which are major drivers of inflammation in multiple sclerosis. We aimed to determine the dose-response relationship between Tolebrutinib and the reduction in new active brain MRI lesions in patients with relapsing multiple sclerosis."**

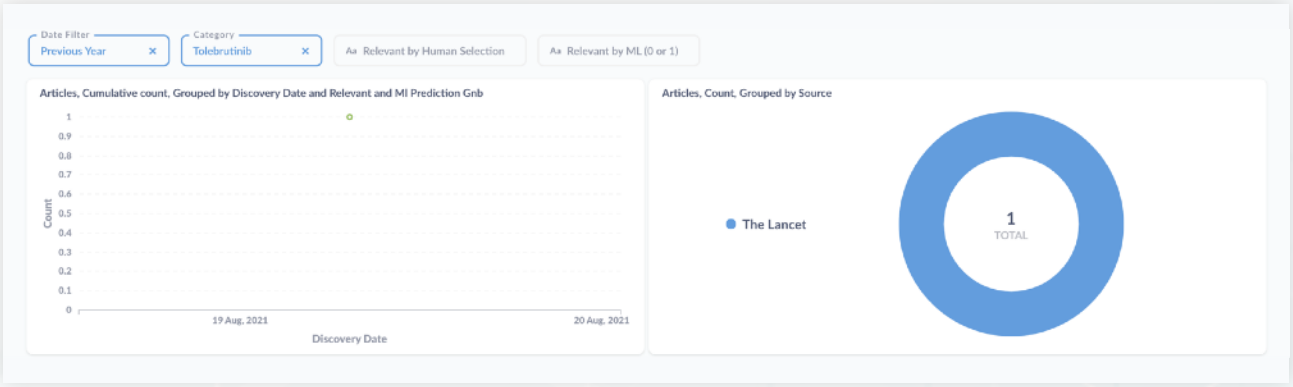


Figure 17.

Source: https://metabase.gregory-ms.com/public/dashboard/4588cf11-c805-4d16-bb5a-f479aa2bcb04?date_filter=2021-01-01~2021-12-31&category=Tolebrutinib

Relevant article by human selection	**1**
Relevant articles by Machine Learning	**1**
Relevant articles where ML and Human agree	**1**

There is also a news article published October 2021, in Neurology Today that refers to a clinical trial that took place. However at this moment news articles fall outside the scope of Gregory's search.

What Gregory MS means

for the scientific and medical community

We have going step by step in how Gregory works. How the database is structured, what sources it uses, which keywords. We have seen that it applies Natural Language Processing to categorize findings, and Machine Learning to try to identify what is relevant.

One thing that we didn't mention strongly was its ability to send research digest by email to external stakeholders.

We have also gone step by step through the different charts and dashboards that can be built with the information on the database.

In this path we have also identified a number of issues that will need to be addressed in future developments of the tool.

1. The training dataset was built by a person without medical training

2. Synonyms for medication aren't identified

3. Our AI tools aren't capable of telling if a drug is relevant to treat MS but has serious side effects

4. Different topics of analysis require different ML algorithms, as was the case when we tried to tag articles from COVID-19 as being relevant or not

And while there may also be other issues, we believe to have shown that Gregory does a good job in cutting down the human cost of finding and summarising the current knowledge on this field.

Not only that, but in the way that we have applied current AI and ML tools, it does not interfere with the scientific method, given that all the information can be reviewed by a human before making a decision.

There is also potential to include more sources of information to make Gregory more relevant for MS research. A clear example is adding more sources for clinical trials. There is also room to gather news articles into the database, provided that we are able to discern between those and the scientific journals.

For example:

- CEIC
- Champalimaud Foundation
- EMA
- FirstWord Pharma
- RNEC

The biggest advantage comes when a system like this is focused on a single topic or goal, as is the case for MS.

We are still scratching the surface of what can be done to put AI in service of medicine. Given proper guidance and user feedback, a system like this can grow to become a qualitative and quantitive view of what is being done to treat MS.

Therefore our obstacle is to make it known to the largest possible number of neurologists and to form a community that is willing to provide feedback and assist in improving the data sets for the algorithm.

But most important of all, we hope to have proven that AI is not an idea from science fiction. The current AI technology is at our reach and it's just a matter of figuring out what is the best way to use it.

And as good as Gregory is, it does not compare to better tools that were built by people who are more skilled in AI and data science. Tools like Connected Papers, Elicit, and Research Rabbit.

Connected Papers | Find and explore academic papers

elicit.org

ResearchRabbit

It is not just exciting to test these tools with a real life problem, it is also a duty to look into anything that can be useful to improve the lives of patients with chronic conditions.

And it is for that reason that Gregory MS was built as an open source project, allowing anyone to install it on their server and tailor it to their specific research needs.

Report and Accounts

It may feel a bit strange for us to include our activities and accounts in this review. The reasoning behind it is that it is important to provide transparency on how Gregory came to be and how it is managed in terms of costs and development efforts.

Gregory MS is the result of a one man team, whose credibility can rightfully be put into question.

There is no goal of making a profit out of keeping the site active and the database accessible. Gregory exists for the common good and to help drive innovation in the pursuit of new treatments for MS.

At the same time we can't expect people to trust us if we are not willing to be transparent, which is why it makes sense to keep a transparent report of activities, sponsorships of any kind, and expenses.

What we did in 2021

Gregory's development began around January 2021 but he only "woke up" in February. The first version was a black and white list of titles and links that did very little to get anyone's attention.

Development kept going at a steady pace and eventually a version 2 was ready. It was a fully redesigned site with sections to communicate it's value and purpose for the different stakeholders.

- Doctors
- Researchers
- Patients
- Physical Therapists

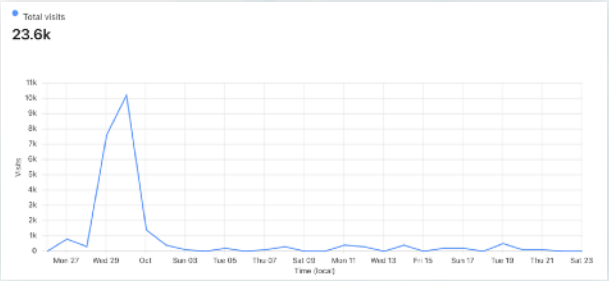


Figure 18 .

Visits to gregory-ms.com during September 2021. Source: CloudFlare analytics.

With this refreshed version it was time to put some effort into publicising the site in relevant networks, online and offline.

On Twitter, @brunoamaral activated a content plan to communicate the site's purpose and new features. An email was also sent out to relevant parties, while the link was also shared within Reddit, Telegram, and other places where there were groups discussing MS.

HackerNews turned out to be the biggest driver of traffic, bringing **16 000 visits**

https://news.ycombinator.com/item?id=28691044

Over the course of this year we have not been walking alone.

1. CloudFlare

CloudFlare became the first company to support Gregory through their Project Galileo. This meant that we have access to their enterprise offering at no cost and thus make sure the site is faster and more secure for visitors. We are also using their Analytics solution to monitor visits. CloudFlare analytics puts the user's privacy as a priority, and we agree. Although that means we have access to less data it also means a better user experience for the user, who doesn't need to consent to cookies, and it means traffic data is not shared with ad networks that could indirectly breach patient confidentiality.

2. Sociedade Portuguesa de Esclerose Múltipla (SPEM)

The Portuguese Society for MS was the first to give us feedback on what we were doing, and were kind enough to give us a spotlight in their Annual Congress. They have been doing an amazing work in Portugal, providing patients with guidance and access to physical therapy focused on these specific needs. Some of their professionals also provided suggestions for sources of information that we managed to implement.

The slides used for the congress can be seen in this link
XVI National Congress on Multiple Sclerosis

3. OneOverZero

Not a company or organisation of any sort, OneOverZero is a group of IT professionals and enthusiasts that have been following Gregory since day one. Their help was valuable to overcome obstacles in development. It is to them that I owe the suggestion and guidance in implementing the Machine Learning and Artificial Intelligence modules.

Accounts

Up until September we were using a shared server at no cost, the same that Bruno uses to host his website. As the site grew it became necessary to provide a dedicated hosting solution, which was done in the middle of the month. For hosting services we are using DigitalOcean. Payments were made in USD and values were calculated with the exchange rate of January 2021.

Over the course of the year we also required some graphic design services from a Freelancer. For this we requested the services of Margarida Gomes. Her assistance was specially useful so that we could have a proper branding for Gregory. In the beginning we were using an illustration from the movie Big Hero 6 and thus infringing copyright and any concept of fair use.

Costs for 2021

Name	Category	Total value in USD	Value attrib. to Gregory MS in USD	Value in €	Invoice	Description
Margarida Gomes	Design			€250,00		(Note 1)
DigitalOcean September	IT Services Hosting	30,20 $	3,84 $	€2,65		(Note 2)
DigitalOcean October	IT Services Hosting	53,00 $	24,00 $	€21,20		
DigitalOcean November	IT Services Hosting	53,00 $	24,00 $	€21,20		
DigitalOcean December	IT Services Hosting	54,91 $	25,67 $	€22,08		
Godaddy 2021	IT Services Domain			€9,63		
Total				**€326,76**		

Note 1.

10h of graphic design that were used to develop the illustration for Gregory and the presentation slides used for the National Congress on MS organised by the Portuguese MS Society.

Note 2.

Dedicated Server to host Gregory-MS.com and also an occasional server to handle the heavy task of compiling a Virtual Machine for one of the modules.

Development costs

We have no way to estimate how many hours went into programming and testing to develop Gregory. We do now that during 2021 there were **650 commits** of new code to add features, fix bugs, or edit content.

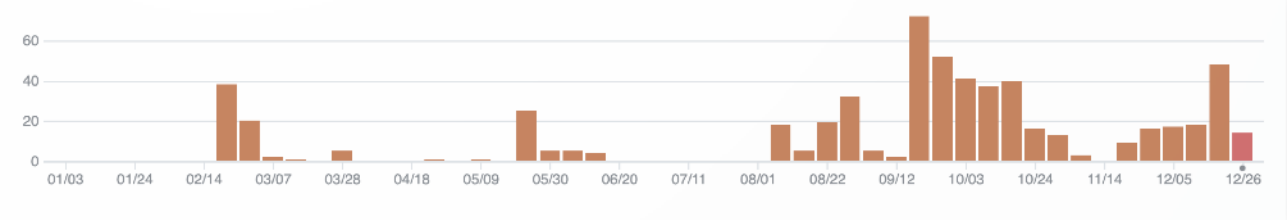


Figure 19.

Code commits over time. Source: https://github.com/brunoamaral/gregory/graphs/commit-activity

There were also **21 issues identified** on GitHub, with **17 resolved** during 2021.
https://github.com/brunoamaral/gregory/issues?q=is%3Aissue

Our roadmap for the future

The development roadmap we had for 2021 was ambitious and although it didn't come to completion we are very happy with the results. A more detailed roadmap can be found on GitHub.
Issues · brunoamaral/gregory

1. **Find articles as they are published** (done)
2. **Identify articles relevant to improve Quality of Life (QOL) or ameliorate symptoms** (done)
3. **Track new therapies in the observatory** (done)
4. **Send a free newsletter to relevant stakeholders** (done)
5. **Inform of related clinical trials** (in development)

Although we are providing this information on the "Observatory" section of the website, listing clinical trials whose title matches the category, it needs to be integrated with the database. It's is a task of medium difficulty and we hope to conclude by June 2022.

6. **Allow everyone to browse and analyse the database** (in development)

We are using a free software for building charts and dashboards from Gregory's database. It's called Metabase and can be found at https://metabase.com/. It allows registered users to build their own queries and dashboards, and create subscriptions to receive regular updates by email. However, this isn't available to users without an account and at the moment we don't see a secure way of allowing everyone access.

While we work on a solution, anyone can download the full database from the Downloads Page.

7. **Spread the word about clinical trials** (done)

We found a way to be a little more effective regarding sharing this information. Gregory has a twitter account and it is being used to share Clinical Trials as they are published on the listed sources. The account can be followed at https://twitter.com/gregoryms_ or by following the hashtags:
#CNSinfections #EM #MS #MultipleSclerosis #NeuroTwitter #Neurology #EscleroseMultipla.

8. **Track acceptance of new therapies and drugs** (to do)

We haven't even began looking into this milestone. It feels more difficult for lack of access to the information and to the people that can in fact provide insight on the best way to obtain it. For this year we will keep this milestone in our sights but it won't be our primary focus.

9. **Help researchers optimize their efforts** (to do)

It may seem easy but the lack of feedback we have been getting from the medical community online and offline is the biggest obstacle. What we have built so far is based on common sense and on a stone-by-stone logic. We make sure that each new feature allows for more than just one use and that each hour spent coding is towards something that won't go to waste.

Right now our focus is building a flexible solution and to share it with as many healthcare professionals as possible in hope of building a community that will guide Gregory's development.

Cost estimates for 2022

	Category	Value in €
Domain cost per year	IT Services | Domain	€8,57
Server cost per year	IT Services | Hosting	€211,97
Backups cost per year	IT Services	€42,36
Total		**€262,90**

Last Remarks

In a year Gregory MS became the biggest public database focused on MS research, and the features developed have the potential to give us a clear view of the landscape where we are. We hope to have demonstrate the sort of analysis that can be done.

There are of course problems and difficulties in building a tool such as this, especially given our limited resources. Over the course of this document we were able to identify quite a few, and in some cases point in the direction of the solution.

And given that this is an Open Source project, available on GitHub, there is room to have the participation of other stakeholders and to see it develop into variants for other goals.

And as we have made it a point to mention, there are tools online that are better than the one we have built, in the sense that they have a better database and return a good quality of results despite being generic to every subject matter.

Gregory MS does not make the statement of being perfect, but it does stand as the only Public Database of MS research that uses AI and ML to assist and motivate the innovation in this field.

We plan to keep developing the software and maintain access to it free. And because our roadmap is not set in stone, we are open to discussing what would be a useful feature for those who plan on using it.

References

1. aHSCT in MS (Autologous Hematopoietic Stem Cell Transplantation). National Multiple Sclerosis Society. Accessed December 31, 2021. https://www.nationalmssociety.org/Treating-MS/aHSCT

2. Alemtuzumab. In: *Wikipedia*. ; 2021. Accessed January 1, 2022. https://en.wikipedia.org/w/index.php?title=Alemtuzumab&oldid=1053709377

3. Alemtuzumab (Lemtrada) | Multiple Sclerosis Society UK. Accessed January 1, 2022. https://www.mssociety.org.uk/about-ms/treatments-and-therapies/disease-modifying-therapies/alemtuzumab

4. Amaral B. Algorithms, Artificial Intelligence, and Us. brunoamaral.eu. Accessed December 30, 2021. https://brunoamaral.eu/post/algorithms-artificial-intelligence-and-us/

5. Russell S, Norvig P. Artificial Intelligence: A Modern Approach, Global Edition. 4th edition. Pearson; 2021.

6. Cladribine (Mavenclad) | Multiple Sclerosis Society UK. Accessed January 1, 2022. https://www.mssociety.org.uk/about-ms/treatments-and-therapies/disease-modifying-therapies/cladribine

7. Cladribine (Mavenclad) | Multiple Sclerosis Society UK. Accessed January 1, 2022. https://www.mssociety.org.uk/about-ms/treatments-and-therapies/disease-modifying-therapies/cladribine

8. Explore treatments in trials | Multiple Sclerosis Society UK. Accessed December 31, 2021. https://www.mssociety.org.uk/research/explore-our-research/emerging-research-and-treatments/explore-treatments-in-trials

9. Alpaydin E. Introduction to Machine Learning, Third Edition. MIT Press; 2014.

10. Khumar K. *Kamal2230/Text-Summarization*.; 2021. Accessed December 31, 2021. https://github.com/kamal2230/text-summarization/blob/6d50ed46fbab53388c9be1fd41e435712d8d8514/Summarisation_using_spaCy.ipynb

11. *Node-RED*. Node-RED; 2022. Accessed January 1, 2022. https://github.com/node-red/node-red

12. ocrevus-epar-product-information_en.pdf. Accessed January 1, 2022. https://www.ema.europa.eu/en/documents/product-information/ocrevus-epar-product-information_en.pdf

13. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation. Accessed January 1, 2022. https://scikit-learn.org/stable/

14. Honnibal M, Montani I, Van Landeghem S, Boyd A. SpaCy: Industrial-Strength Natural Language Processing in Python.; 2020. doi: 10.5281/zenodo.1212303

15. khumar K. Text summarization using spaCy. Analytics Vidhya. Published August 2, 2020. Accessed December 31, 2021. https://medium.com/analytics-vidhya/text-summarization-using-spacy-ca4867c6b744

16. Hurley D. Tolebrutinib Appears Promising Against Multiple Sclerosis in Phase 2 Trial: Study Now Moves to Phase 3. *Neurology Today*. 2021;21(19):12. doi: 10.1097/01.NT.0000797972.51189.e7

17. Types of Stem Cell | Stem Cells | University of Nebraska Medical Center. Accessed January 1, 2022. https://www.unmc.edu/stemcells/educational-resources/types.html

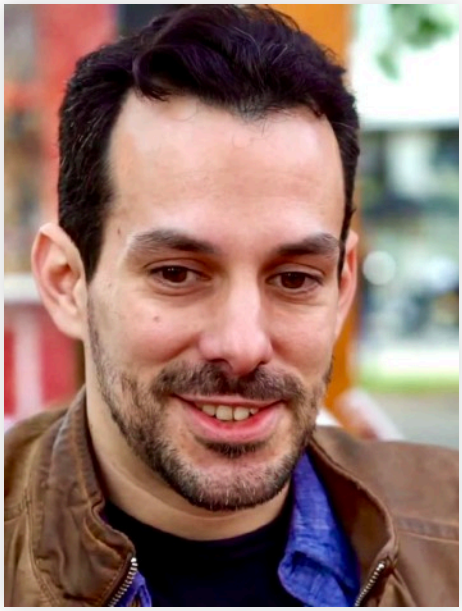
18. Why Choose TYSABRI® (natalizumab)? Accessed January 1, 2022. https://www.tysabri.com/en_us/home/about/www.tysabri.com/en_us/home/about/choosing-tysabri.html

Bruno Amaral

I am a Digital Strategist with over 15 years of experience in both Communication with some knowledge on programming.

During my career I had the opportunity to work in projects for major brands in Portugal. EDP, where I currently work, Unicer, Siemens, Porto Business School, Unilever and a few others.

I also teach Communication Strategy in the University of Lisbon.



+351 912 875 856

bruno@gregory-ms.com

https://brunoamaral.eu/

Bruno Amaral - Digital Intelligence and Strategy

Thank you to

Gregory MS didn't come to be without the help of people I have found along the way.

The Portuguese MS Society, SPEM. Especially, **Paulo Gonçalves** and **Alejandro Carrabs**.

André Correia from **Cruzamento, a Podcast on innovation and healthcare**.

Margarida Gomes, for her professionalism in delivering the best possible graphic design.

Everyone from **OneOverZero**, and especially Manfred Macx.