UNIVERSITAT POLITÈCNICA DE CATALUNYA

MSc THESIS

A semantic analysis of predictive models in personalized medicine for Artificial Intelligence automated risk assessment

---

Author:
Marc Palomer I Cadenas

Supervisor:
Pol Solà I De Los Santos
Alexandre Perera Lluna

Submitted in fulfillment of the requirements
for the degree of MSc in Biomedical Engineering

B2SLab
CREB

July 20, 2023

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**Artificial Intelligence**
DL	**Deep Learning**
ML	**Machine Learning**
EU	**European Union**
US	**United Stated of America**
LLM	**Large Language Model**
ALTAI	**Assessment List For Trustworthy AI**
MDR	**Medical Devices Regulation**
IVDR	**In Vitro Di- agnostic Medical Devices Regulation**
ATM	**Automatic Term Mapping**
BERT	**Bidirectional Encoder Representations from Transformers**
LOESS	**Local Polinomial Regression**
PCA	**Principal Component Analysis**

List of Symbols

$\mathbb{R}^n$	Real mathematical space of n dimensions

Chapter 1

Abstract

AI systems present a promising future in terms of technical advances and societal benefit. However, such advances depend on the societal acceptance and adoption. Here, we pretends to deliver a reproducible open-source framework to evaluate the status of AI from multiple perspectives. Conrectelly, we focus on the healthcare vertical. To do so, we performed systemic analysis of the health-based AI science of PubMed.

First we show a remarkable increase on AI health-related science in the last lustrum. Similarly, we observed a preference of the authors to publish using AI techniques, specially deep learning techniques, and a tendency to abandon classical medical statistics. Although it is well-known that AI models tend to exhibit greater predictive capabilities, it comes to a cost of interpretability, explainability and privacy. As a consequence, health professionals might struggle to comprehend the decision-making process behind these complex algorithms leading to hindered adoption and potential accountability issues.

Second we characterize the competitive race among countries for AI dominance. Currently, China leads the race on AI-related health-science generation while the US and the EU show slower publication power. Data suggests that China will lead AI research and take the lead on edge-AI research. These might be increased by cultural differences, specially data access and privacy policies, which might have a negative impact western competitiveness in the field.

Third, through a semantic analysis based on a LLM embedding we evaluated the changes suffered by the AI-statistics health science. Our findings reveal that AI articles are semantically different from what was the norm of

medical statistics for the last two decades, indicating an increase of knowledge boundaries within health research. However, the introduction of a complete new field might provoke clinicians to feel excluded from AI-driven research topics, which, in turn, can potentially hinder the clinical-based innovation process.

In conclusion, this thesis provides insights into the evolution of AI in the healthcare and interprets them from the risk assessment lenses through the development of a set of reproducible open-sourced indicators. We expect this research, and its further advances to contribute to a better understanding of the current state of AI and a clearer awareness of the principal risks and challenges posed by this field.

Chapter 2

State of the Art

In recent years, there has been a remarkable and widespread surge in the development and implementation of artificial intelligence (AI) across various economical verticals, including healthcare. However, this advancement in AI comes entangled with risks and challenges. Ethical considerations, data privacy concerns, algorithmic biases and potential job displacement are some of the potential drawbacks of incorrectly handled AI.

2.1 AI overview

AI adoption has seen significant growth in recent times. Approximately 50 to 60% of organizations currently are utilizing AI technologies, being digital assistants the most prevalent solution, 42%, used by up to 72% of business executives [1]. In terms of AI type production, it is straightforward that generative AI has gained significant attention and public awareness lately. Prominent examples include text-to-image models such as DALL-E 2 and Stable Diffusion, text-to-video systems like Make-A-Video, and sophisticated chatbots like ChatGPT. All these trends combined signify the increasing recognition of AI's capabilities and its integration into diverse sectors. Geographically, China is currently holding the lead in AI journal, conference, and repository publications. Meanwhile, the United States retains its edge on AI conference celebrations although this lead is gradually diminishing. However, to the extend that is reported, the majority of large language models (LLMs) are produced by American institutions [2].

Economically, AI is a growing multi-billion market. In 2022, AI investments topped $6.1B in medical and healthcare, $5.9B in data management, processing, and cloud services, and $5.5B in Fintech highlighting the significance of AI as a innovation driver for near-to-all industry verticals [2].

Nevertheless, these fast adoption comes with an increase in legal cases. Since 2016, the number of legal AI-related cases recorded in US state and

federal courts has been multiplied by seven, topping 110 in 2022. These cases primarily revolve around civil, intellectual property, and contract law. In addition, ethical concerns, potential job displacement, and the need for transparency and accountability are critical aspects that require careful consideration. To that purpose, several ethical guidelines, responsible practices and regulatory reports have been issued by public and private institutions to serve as a guardrail for risk-less AI development.

2.2 European AI regulation

The European Union (EU) in its different institutions and chambers are leading the discussion on the impacts of emerging digital economy to civil and human rights. Similarly, the EU is leading the AI regulation pushing public and private actors in the pursue of the development of a framework that guides and regulates AI.

In 2018 the **High-Level Expert Group on AI (HLEG)** was constituted with a total of 52 AI experts. Their main purpose was to advise the European Comission on their mission to implementat an Strategy on Artificial Intelligence. As a result, key documents and reports were published by EU commission. In April 2019, the Commission released the HLEG ethics guidelines for Trustworthy AI [3]. These guidelines where a crystallization of the analysis of over 500 submissions from multi-diverse AI stakeholders. One key takeaway of the latter is the need for a AI design aligned with the EU values.

On 2020 the EU White Paper on AI [4] was published. This document was the first attempt for an objectivation and operationalization of an AI regulation. However, it was still open to public debate and feedback.

The EU AI Act [5], released on April 2021 and approved in these June 2023, is the materialization of all this mentioned efforts. The EU AI Act has two main objectives: 1) facilitate innovation, investment and development of AI through legal certainty and avoiding market fragmentation and 2) Ensure safe, lawful and trustworthy AI thorugh law enforcment of fundamental rights ans communitary values. Another key element of the AI act is a structured risk classification system for AI solutions. This classification system determines the degree of risk each AI system as a function of the vertical of implementation and the potential harm to humans, society & environment. In order to ensure a correct application of the AI act, an European AI Board will be constituted. Among its functions, this board will oversee the implementation of the regulation and ensure uniform application across the EU,

as well as generate opinion and recommendations on issues that might arise within national authorities.

In parallel to regulatory advances, efforts have been made to create a more operational approach as a complement meanwhile the full infrastructure to enforce the regulation is not in place.

2.3 AI development guidelines

The European public institutions and several private stakeholders developed guidelines on how to implement the advances in the regulatory requirements. Among others the Assessment List For Trustworthy AI (ALTAI) [6] was released by AI HLEG. The ALTAI pursues the development of AI aligned towards the EU values. These include human agency and oversight, technical robustness and safety, transparency, privacy and data governance, societal and environmental well-being, diversity, non discrimination and fairness, and accountability. In short:

- Human agency and oversight requirements account for the system-human interaction. For example, which is the role of the system in decision making (eg. lack of human supervision or loss of critical judgment).
- Technical robustness and safety requirements account for security, safety, accuracy and reliability. For example, issues in system underperformance, drifts, or uncorrectness of the predictions.
- Data privacy and governance accounts for the intersection of AI with the GDPR, correcting the use of DIPA assessments and prioritizing user rights over system performance.
- Transparency requirements account for the capacity to explain and interpret AI systems. Of special relevance are critical systems (ie. high-risk) where human judgment must be enhanced through explainable techniques.
- Societal and environmental well being requirements account for the assessment of the end-to-end environmental impact as well as the potential societal impact (eg. manipulation).
- Diversity, Non-discrimination and Fairness requirements account for a FAIR design of AI as well as to universal access to the technologies.

- Accountability requirements account for the establishment of a traceability measures for forensic analysis in case of AI system failure, as well as a clear establishment of responsibilities.

Although major advances have been made in the development of regulatory documents as well as operational guidelines, some fields that are considered high risk require from extra care in the AI development and deployment. That is the case of healthcare environments. Healthcare is a zero-risk approach vertical that will generally deal with more strict requirements regarding AI, apart from those belonging to the medical essence of the product.

2.4 Health AI reports

Implementation of AI in healthcare raises social and ethical concerns of greater significance in comparison than when used on general purposes. This is due to the particularly high-risk environment and the direct impact of AI-driven decisions on people's health.

Previously mentioned documents are on general purpose AI, not specifically healthcare. Due to the specific characteristics of the healthcare field, the document PE 729.512 by the European Parliament was released in June 2022 to fill this gap. This document pretended to converge principles present in the more generic documents into the healthcare field, proposing mitigation measures and policy options to minimise risks and maximise benefits of medical AI. Based on the medical AI document PE 729.512 and on the auto-assessment principle, a concrete checklist for trustworthy AI in medicine has been developed by a network of European Commission funded research projects. Its name is FUTURE-AI, and its guidelines are organised according to six principles (fairness, universality, traceability, usability, robustness, explainability) and pretends to help AI designers, developers, evaluators and regulators to develop trustworthy and ethical AI solutions in medicine and healthcare.

The actual applicable regulations for medical AI tools in the EU are the 2017/745 Medical Devices Regulation (MDR) and the 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR). However, these regulations were formulated during the early stages of AI development, and as a result, they do not correctly manage various aspects specific to AI. For instance, they fail to regulate continuous learning of AI models, the identification of algorithmic biases, or to accurately define who is responsible of AI malfunction. It is

visible at first sight that some of the mentioned problems not accounted by the current legislation are closely related to the concepts that have been discussed in the previous chapters as robustness, fairness or liability, and that present special relevance in medical fields.

2.5 Open issue

The raise of AI in healthcare poses medical and scientific challenges for professionals and policymakers. Hence, it is of critical importance to objectify its current state through quantitative knowledge. Although efforts to take the pulse to AI have been made, these are case-specific, and both data and results are typically controlled by private corporations and, in consequence, they have a major lack of reproducibility.

We consider that the only way to have a continuous assessment of the state of AI is through an open approach, that allows stakeholders to reproduce the data and results. This is of major relevance in high risk verticals, as healthcare.

Chapter 3

Goals

3.1 Main goal

Our main objective is to generate a reproducible pipeline to measure the pulse to Artificial Intelligence in the healthcare field through a set of quantitative indicators.

3.2 Secondary goals

In addition we plan to,

1. Evaluate for the evolution of AI science in the healthcare field over the last two decades,
2. Evaluate the centralization of AI know-how by the world-economic powers,
3. Evaluate the differences between classical medical statistics and AI from the semantic perspective.

3.3 Expected contribution

By interpreting all the generated information from the risk-assessment lenses we expect this report to provide objective measures of the main risks and challenges that AI in healthcare might pose to society and thus contribute to the discussion on how to address AI risks through regulation and policies.

Chapter 4

Materials

4.1 PubMed

PubMed [7] is a free search engine specialised on life sciences and biomedical topics. It comprises more than 35 million publications of biomedical literature from MEDLINE, life science journals, and online books, see table 4.1. PubMed is one of the highest impact databases in biomedical research worldwide, which can be accessed both by its web domain or through an API.

	2022	2021	2020
PubMed Citations (Annual)	1,714,780	1,733,089	1,514,199
PubMed Citations Cumulative	34,693,538	33,136,289	31,563,992
PubMed Searches	2.58 Billion	2.57 Billion	3.3 Billion
Web/Interactive	1.283 Billion	1.186 Billion	1.076 Billion
Script/E-Utilities	1.303 Billion	1.391 Billion	2.2 Billion

TABLE 4.1: PubMed Statistics [8] in US Billions

Here we use the repository of PubMed at the date 14th of May of 2023, as reference data base to access scientific journals from which to create the corpus for the analysis.

4.1.1 API PubMed

An API or Application Programming Interface, is an structured manner for two computer programs to communicate. We utilize the PubMed API to access and download articles from PubMed in a structured manner. The variables retrieved from PubMed for each article are shown in the table 4.2.

Variable	Description
pmid	PubMed ID of the publication
doi	Digital object identifier
journal	Name of the journal of publication
author	List of authors
affiliations	University of affiliation of the authors
year	Year of publication
abstract	Abstract of the publication
query	Query or search term used for retrieval

TABLE 4.2: Variable description

4.1.2 PubMed Querying system

PubMed Query

In order to retrieve the desired articles from PubMed the user defines their search of criteria in a query. A query is a string concatenation introduced into the database through the searcher that serves for the database management system to identify and retrieve the demanded data from the database. The design of the query is dependant on the logic used by the search engine. In the case of PubMed, the query is defined as search terms (word or group of words) joined by logical operations. Search terms are chosen and tuned in order to modulate the results. All the set of available operations are defined in [9]. An example of a PubMed query is:

```
Hip arthroplasty guided surgery AND (standard practices OR good practices)
```

Through this query the database would retrieve all the articles that match the terms "hip arthroplasty guided surgery" and at the same time match the terms either "standard practices" or "good practices" .

Search Terms

Search terms is the name given to the word or group of words separated by logical operations in the query. For instance in the query example above, the first search term is "hip arthroplasty guided surgery". In PubMed search terms can be of two different types: MeSH terms or free text terms. These can be identified because MeSH terms are text followed by the specification"[MeSH]", while free text search terms are not, as in the example above.

MeSH or Medical Subject Headings is a specialised ontology for biomedical literature. MeSH terms are organized hierarchically, with broader terms

at the top and more specific terms below them. This structure allows for efficient and accurate retrieval of information in the PubMed database. Each term has a unique descriptor and multiple subheadings, which helps to organize medical and health-related concepts. This tree-like structure makes it easier for researchers to locate relevant articles and resources, as MeSH terms are manually mapped into newly uploaded articles on PubMed, as tags, and makes it easier to retrieve accurate information on the topic of interest. MeSH terms have been used in this project to retrieve articles with higher relation to the topic of interest. The MeSH ontology can be accessed online.

It's important to note that once a MeSH search term is used in a query, the database will retrieve all the articles that are tagged with that term or with terms under it in the hierarchical tree. On the other hand, if the search term is free text, PubMed first maps into a MeSH term through Automatic Term Mapping (ATM) algorithm. In the case of no coincidence, then free text is used to search within the database using the title and abstratcs as default corpus.

Here we map each desired search to MeSH terms manually.

Search term combinatorics

Several operations can be performed to increase the specificity of the query [9]. The main operations are listed below:

- **Boolean operators**: Logical operators (eg .AND, OR, NOT)) can be used between search terms, parenthesis must be used to bypass prioritization of boolean operations. For example,

  ```
  Arthrosis AND Ageing
  ```

  In this example, only the articles talking about both arthrosis and ageing are going to be retrieved.

- **Field tags**: Field tags lets you specify where the term must be searched. Default field tag is title and abstract. Field tags must be placed between brackets after the search term.

  ```
  COVID-19 AND yellow fever[tiab] AND 2000:2023[dp]
  ```

  In the previous example, `yellow fever` will be searched only in the title or abstract of the articles as defined by the field term `[tiab]` and

is forced by the field term `[dp]` to be published between `2000:2023`. It is important to note that when using a field term any ATM mapping is deactivated.

- **Truncating terms**: Truncation is a modification of the search term by adding a "*" at the end of the last word. This method is used to indicate the database that you want all the articles matching the search term and with its variations from the letter where you put the "*". It is important to notice that when using truncating terms any ATM mapping is deactivated.

```
Support Vector Machine* AND Ictus prediction
```

In the previous example, not only the results matching Support Vector Machine will be retrieved, but also plurals as Support Vector Machines. In consequence, when using truncation we are working with free text search terms, not MeSH.

- **Quoted terms**: Activates free text search, deactivates ATM and thus MeSH in order to retrieve only articles that exactly match the search term.

```
"convolutional neural network"
```

In this query the only articles that will be retrieved are those that exactly have in their title or abstract (default free text search) the sentence "convolutional neural network". It is important to note that if an article has in its title a plural synonym as `"convolutional neural networks"` it will not be retrieved since the match must be exact.

4.2 Python and R

Python programming language [10] in its version 2.7 and the interactive web-based platform Jupyter notebook [11] have been used for all this project. Secondly, R programming language [12] and its famous distribution R Studio [13] have been used mainly for data visualization. The main libraries used in this thesis are tabulated at table 4.3.

Language	Package	Description
Python	Numpy [14]	Numerical library
	Pandas [15]	Data processing library
	Sentence Transformer [16]	NLP pretrained models
	plotly [17]	Plotting library
R	ggplot [18]	Plotting library
	plyr [19]	Split-apply-combine data

TABLE 4.3: List of libraries used in the project

4.3 BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [20] is a fine-tuned BERT NLP algorithm [21]. BioBert's fine tunning is performed in the PubMed's database. As a consequence, it is reported that BioBERT significantly outperforms BERT on the following three representative biomedical text mining tasks: biomedical named entity recognition , biomedical relation extraction and biomedical question answering [22]. Hence, we use it in our analysis. BioBert is easily accessible through Python and the `Sentence Transformer` library[16]. BERT is the core model used in BioBERT. It has been trained by google [21] in 2020 with the general-purpose corpus of wikipedia. It has been used in several NLP tasks as text prediction (eg. when writing an e-mail) or text question answering. BERT uses bidirectional architecture shown in Figure 4.1 to understand the relationships between words in a text. Before processing word sequences, BERT replaces around 15% of the words with a [MASK] token and tries to predict the original values based on the surrounding words in the sequence. Predicting the masked values and disregarding predictions for non-masked words causes the model to focus on understanding word context [23].

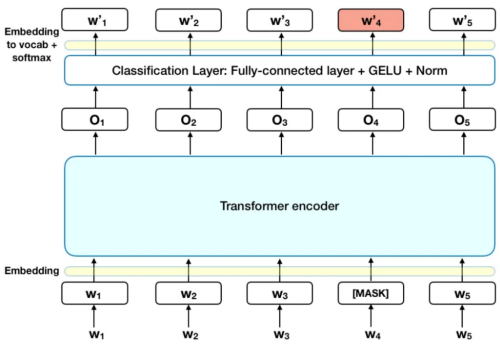

FIGURE 4.1: BERT pipeline [23]

Chapter 5

Methodology

5.1 Query fields

We generated the queries combining three groups of search terms that attempt to cover the following questions: i) what is the field of study of the AI solution, ii) what AI algorithm is being used?, and iii) what type of analysis is being performed. Per each search field, a set of MeSH terms has been manually assigned.

Field of the study

This term attempts to cover for the field of study on the article is published. The following table 5.1 relates the different selected fields of study with its corresponding MeSH terms.

TABLE 5.1: Field of study table

Field of study	**MeSH translation**
Biomedical research	("Biomedical Research" [MeSH] OR "Research Design" [MeSH] OR "Medicine in Literature" [MeSH])
Drug discovery, development and evaluation	("Drug Discovery" [MeSH] OR "Drug Development" [MeSH] OR "Drug Evaluation" [MeSH] OR "Clinical trials as topic" [MeSH] OR "Pharmaceutical preparations" [MeSH] OR "Biomarkers, Pharmacological" [MeSH] OR "Pharmacovigilance" [MeSH] OR "Pharmacology" [MeSH] OR "Pharmacokinetics" [MeSH] OR "Drug Administration Schedule" [MeSH] OR "Therapeutic drug monitoring" [MeSH])
Biomarker research	("Etiology" [MeSH] OR "Causality" [MeSH] OR "Biomarkers" [MeSH])

Healthcare resource optimization	("Crew Resource Management, Healthcare" [MeSH] OR "Health Care Rationing" [MeSH] OR "Resource allocation" [MeSH] OR "Health care costs" [MeSH] OR "Health care reform" [MeSH] OR "Health Care Economics and Organizations" [MeSH])
Patient prioritization and fast treatment	("Time-to-Treatment" [MeSH] OR "Health care costs" [MeSH] OR "Hospital Rapid Response Team" [MeSH] OR "Health Services Accessibility" [MeSH] OR "Severity of Illness Index" [MeSH] OR "Triage" [MeSH] OR "Emergency Medical Services" [MeSH])
Treatment decision-making	("Precision Medicine" [MeSH] OR "Therapeutics" [MeSH])
Surgical procedures optimization	("Surgical Procedures, Operative" [MeSH] OR "Surgical Equipment" [MeSH] OR "Surgical Instruments" [MeSH] OR "Minimally Invasive Surgical Procedures" [MeSH])
Clinical Analysis and pathological anatomy	("Clinical Laboratory Techniques"[Mesh])
Hospitalary audits	("Clinical Audit"[Mesh] AND "Quality of Health Care"[Mesh] AND "Prescription Drug Monitoring Programs"[Mesh] AND "Management Audit"[Mesh])
Disease detection	("Predictive value of tests" [MeSH] OR "Early diagnosis" [MeSH] OR "Diagnosis" [MeSH] OR "Diagnostic techniques and Procedures" [MeSH] OR "Mass screening" [MeSH])
Disease course prediction	("Disease Progression" [MeSH] OR "prognosis" [MeSH] OR "Predictive value of tests" [MeSH])
Disease monitoring	("Disease monitoring (KPIs monitoring)" [MeSH] OR "Monitoring, Physiologic" [MeSH] OR "Point-of-Care Testing" [MeSH])

Fields of study are derived from the ones proposed by Amisha et. al. [24].

Algorithm type

This term attempts to cover the high-level architecture of the model,. The following table 5.2 relates the different selected type of algorithm or model with its corresponding MeSH terms.

TABLE 5.2: Algorithms type table

Algorithm type	**MeSH Term**
Linear Models	"Linear Models"[Mesh]
Logistic Models	"Logistic Models"[Mesh]
Decision Trees	"Decision Trees"[Mesh]
Random Forest	"Random Forest"[Mesh]
Support Vector Machine	"Support Vector Machine"[Mesh]
Neural Networks, Computer	"Neural Networks, Computer"[Mesh]
K-means	"K-mean*" [tiab]
Hierarchical Clusterings	"Hierarchical Clustering*" [tiab]
Principal Component Analysis	"Principal Component Analysis"[Mesh]
Factor Analysis, Statistical	"Factor Analysis, Statistical"[Mesh]
Convolutional Neural Networks	("Convolutional Neural Networks" [tiab] OR "Convolutional Neural Network" [tiab] OR "Convolutional Neural Net" [tiab] OR "Convolutional Neural Nets" [tiab] OR "CNN" [tiab] OR "CNNs" [tiab])
Recurrent Neural Networks	("Recurrent Neural Network*" [tiab] OR "RNN" [tiab] OR "RNNs" [tiab] OR "Recurrent Neural Net*" [tiab])
Long Short-Term Memory Networks	("Long Short-Term Memory Network*" [tiab] OR "LSTM*" [tiab] OR "Long Short-Term Memory Net" [tiab] OR "Long Short-Term Memory Nets" [tiab])
Generative Adversarial Networks	("Generative Adversarial Network*" [tiab] OR "GAN" [tiab] OR "GANs" [tiab] OR "Generative Adversarial Net" [tiab] OR "Generative Adversarial Nets" [tiab])
k-nearest neighbours	("KNN" [tiab] OR "KNNs" [tiab] OR "k-nearest neighbour*" [tiab])
Naive Bayes	"Naive Bayes" [tiab]
Gaussian Mixture Models	"Gaussian Mixture Model*" [tiab]
Reinforcment learning	"Reinforcment learning" [tiab]
t-distributed stochastic neighbor embedding	("t-SNE" [tiab] OR "t-distributed stochastic neighbor embedding" [tiab])

Association rule	"Association rule" [tiab]
Genetic algorithms	"Genetic algorithm*" [tiab]
Instance based	"Instance based" [tiab]
Synthetic data	"Synthetic data" [tiab]

The algorithm types where based on the divisions defined by Isaac Kofi Nti et. al. at [25].

Analysis type

This term attempts to cover the type of analysis that is being performed. The following table 5.3 relates the different selected analysis types with its corresponding MeSH terms.

TABLE 5.3: Analysis type table

Analysis type	**MeSH Terms**
Regression Analysis	"Regression Analysis"[Mesh]
Classification	"Classification"[Mesh]
Cluster Analysis	"Cluster Analysis"[Mesh]
Survival Analysis	"Survival Analysis"[Mesh]
Generative	"Generative" [tiab]
Dimensionality Reduction	"Dimensionality Reduction" [tiab]

In this case, the types of analysis were derived from personal research and validated by peers at research group where this thesis is being performed.

5.2 Query composition

To extract the information of PubMed a query is required. Once query fields are defined its composition is required. Composition is the task of merging a Field of Study, with an algorithm type and a specific analysis through the logic operators. For example:

```
("Biomedical Research"[MeSH] OR "Research Design"[MeSH] OR
"Medicine in Literature"[MeSH]) AND "Linear Models"[Mesh] AND
"Regression Analysis"[Mesh]
```

The above example is the first query generated by combination of the first MeSH translations of the three fields.

If composition was generated in an all-with-all strategy, some of the compositions would lack of technical sense and might thus retrieve no articles. For example: `"Hierarchical Clustering*" [tiab] AND "Generative" [tiab]`.

5.3 Data retrieving

5.3.1 Data download and storage

Once the set of queries is defined, we use a Python Code 9.1 to retrieve the data using the PubMed API. Essentially, a query is sent to PubMed through the API which it returns a set of data containing every article matching the query. In this set of data, matching algorithms are instances, and the variables of interest are columns. This process is iterated through all the queries. Once all the data is retrieved, results are automatically stored into a CSV file. Since this process is requires intensive interaction with an API there is the risk that PubMed cancels the process by identifying it as an attack. Some preventive measures as pauses throughout the loop where established.

5.3.2 Data preprocessing

First, integrity of the variables is assessed through adjusting its format. PMID, DOID, and the year are formatted as integers, while affiliation, authors, journal, abstract, and query are formatted as strings. Second, redundancies are deleted. Note that an article might appear in multiple queries. The latter case is not considered a redundancy. Third, missingness is assessed. Observations with empty abstract are completely removed from the study. Finally, to correct by the publication review bias (ie. drop in the number of publications because manuscripts are still in the process of acceptance) years 2022 and 2023 are removed from the study.

5.3.3 Data transformations

Some of the query fields are sparse, to simplify the interpretation and the analysis we performed some feature aggregation. For example, the types of algorithms are categorized on three groups: i) "Classical statistics" for articles using descriptive statistics, linear regression, or logistic regression; ii) "AI group" for articles utilizing machine learning (ML) or deep learning (DL) algorithms; iii) and "others" for articles using any other types of algorithms.

5.3.4 Data enrichment

In order to expand the information of the set of data, we used some variables as a proxy for a feature of interest. For example, we extracted the country of publication by using the affiliation of the first author as a proxy. In addition, we defined a new variable "SuperPower" that clusters countries in an economical power basis as US, EU, China or Others.

5.4 Data analysis

5.4.1 AI evolution analysis

In order to estimate the evolution of the AI in the last decades, we propose two indicators:

- **Publications per year:** Evaluate the number of publications generated per year per the AI and the classical statitics field. In order to visualize the information a barplot is used. We generated the bar plot using the `ggplot2` R package.
- **Deep learning publications per field of study per year:** Evaluate the number of deep learning publications generated per year per each field of study. Here we focus on DL techniques since we want to gain insight on the potential risk of poor explainability. In order to visualize the information a scatter plot is used. We generated the scatter plot using the `ggplot2` R package. In addition, we add linear regression to estimate the field growth per lustrum. Also we perform a bar plot of the regression coeficients to quantify the growth of the publication number per field.

5.4.2 AI centralization analysis

In order to estimate the centralization of the AI geographically and by the economical powers, we propose the following indicators:

- **Publications per continent:** Evaluate the number of publications per year and continent. As in the latter case, a stacked plot can be used for visualization. In order to correct for the continent size, we estimate the relative number of publications. Here we use a a relative stacked plot to illustrate the contribution of each continent to the overall publication count of each year.

- **Publications per economic powers:** Evaluate the number of publications generated by the main economic powers (US, EU, China). Similarly, both the absolute and relative publications are evaluated. To visualize trends we use a scatter plot and a tendency line using a local regression technique (LOESS) per each economic power.

5.4.3 AI semantic analysis

To estimate the semantic drift imposed by the introduction of AI we project each abstract towards a semantic embedding using BioBert pre-trained algorithm and the SentenceTransformer library [16]. With this step we can charactherise each abstract to a $\mathbb{R}^n$ space. Concretelly, BioBert embedding is an $\mathbb{R}^{768}$ space, implying that each abstract is converted to a 1 x 768 vector. Once each abstract is projected to the embedding a matrix is obtained. This matrix allocates articles as rows and dimensions of the embedding as columns. Although abstracts are in the $\mathbb{R}^n$ space and thus are prepared for mathematical analysis, its high dimensionality makes it hard to extract any conclusions. To that purpose we applied a dimensionality reduction technique named Principal Component Analysis. The first principal components are retained. In order to evaluate the semantic drift, we propose the following metric:

- **Abstract embedding analysis:** Use a scatter plot of the first and second PCs for an unsupervised analysis of the embedding aggregated as Classiscal Statistics, DL and ML.

Chapter 6

Results

6.1 Data preparation

A total of 1656 queries where composed. These queries where then used to retrieve articles from the PubMed API. A total of 455798 articles where retrieved from the PubMed database. In table 6.1 it can observed an example of the raw data retrieved from PubMed for a given query:

TABLE 6.1: Field of study table

Variables	Example
pmid	19207933
doi	'10.1111/j.1439-0388.2008.00759.x',
journal	'Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie',
abstract	'The present study aimed at assessing the status of the Chilika buffalo population of eastern India employing cytogenetic and molecular...
authors	"['Mishra', 'Kataria', 'Bulandi', 'Prakash', 'Kathiravan', 'Mukesh', 'Sadana']",
affiliations	"['National Bureau of Animal Genetic Resources, Karnal, Haryana, India. bpmishra$_1$*hotmail.com*']",
year	2009
query	'("Clinical Laboratory Techniques"[Mesh]) AND "Cluster Analysis"[Mesh] AND "Factor Analysis, Statistical"[Mesh] AND 2009[dp]',
Obj. MeSH Translation	'("Clinical Laboratory Techniques"[Mesh])',
AlgorithmType	'"Factor Analysis, Statistical"[Mesh]',
AnalysisType	'"Cluster Analysis"[Mesh]',
AlgorithmAcronym	'"Factor Analysis, Statistical"[Mesh]',

The total time of to download the data was 4 hours and 29 minutes. Once preprocessed, we removed all the retrieved articles that lacked from an abstract representing a 2.52% of the total. Once curated, we expanded the data set with feature aggregation and enrichment.

6.2 Data analysis

6.2.1 AI publications per year

In the following figure, we represent the number of abstracts that are published per year stratified in two groups: i) AI, that is a broad term for ML & DL related algorithms and, ii) Statistics, representing classical methods used in the healthcare space (eg. Linear regression).

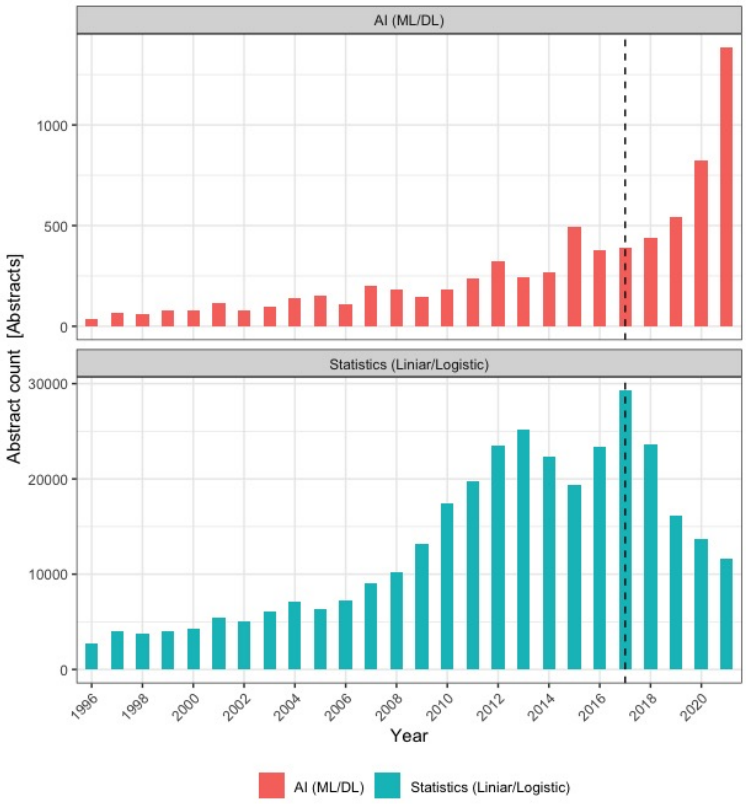

FIGURE 6.1: Plot of abstract count stratified per AI and Medical Statistics field

It can be observed, specially since 2017, an tendency towards the use of AI based algorithms which comes at a cost of a disuse of classical statistics.

6.2.2 DL publications per field of study per year

In the following figure 6.2 we focus on the use of deep learning techniques. Here we represent the absolute number of DL articles published each year. In addition, we use an aggregation of all the study fields to avoid over characterization. This aggregation is "biomarker research, disease detection, prognosis & monitoring".

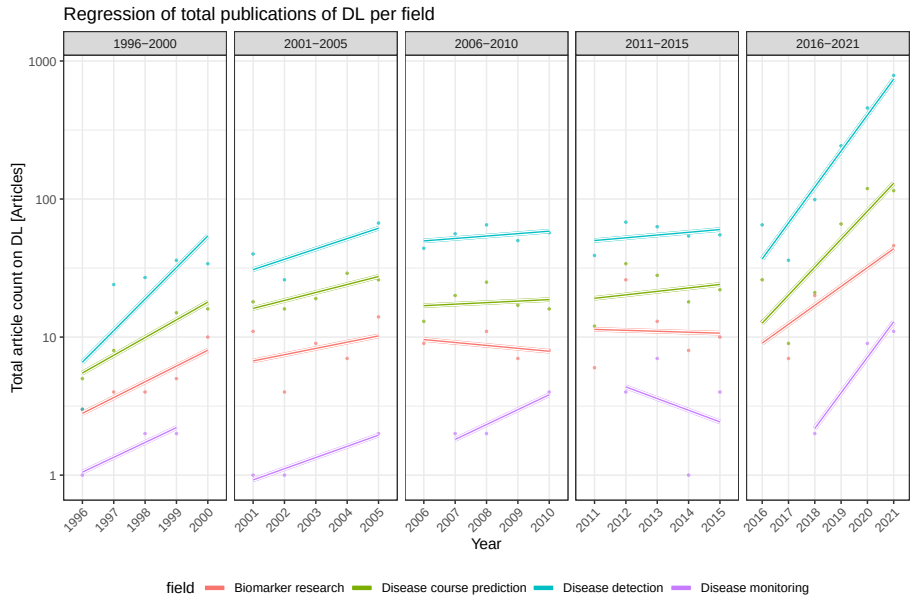

FIGURE 6.2: DL Growth per objective

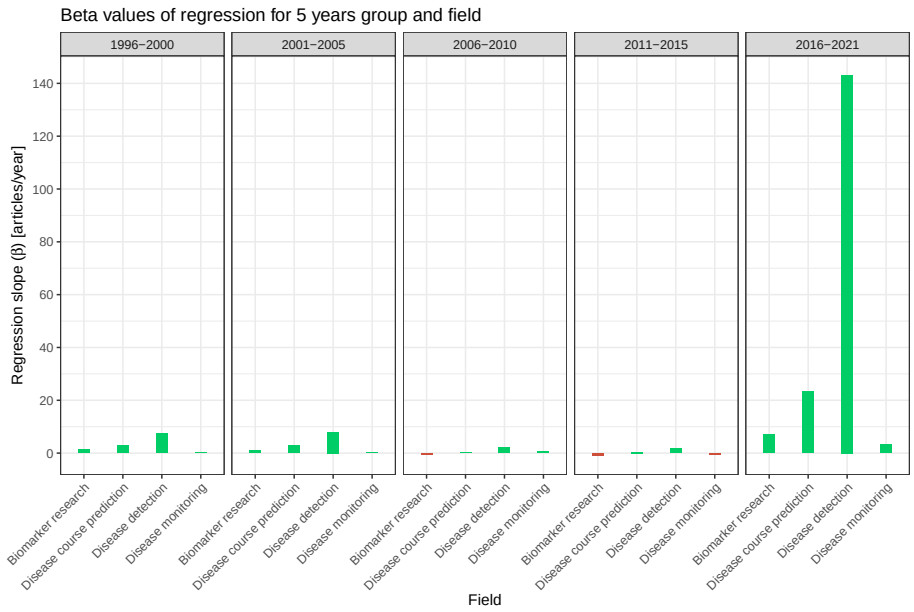

FIGURE 6.3: Slope of regression per objective

It can be observed a small but continuous increase of publications for all the four strata from 1999 to 2015 and a clear acceleration from 2016 until 20201.

This increase or decrease in the absolute number of publications is quantified in figure 6.3, where through a bar plot of the coefficients of the linear regressions.A clear tendency towards the generation of deep learning techniques for disease detection can be observed in the last lustrum.

6.2.3 Publications per continent

Here the AI scientific production per continent is assessed. The total amount of accepted articles per continent is represented as well as its relative percentage to the total year publications is shown in figure 6.4

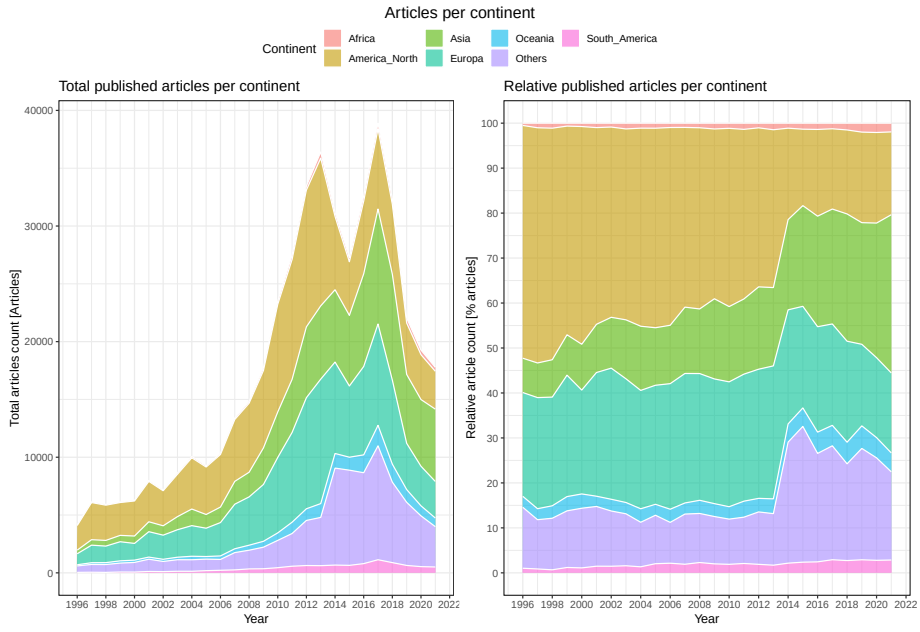

FIGURE 6.4: Publications per Continent

In the previous figure several interesting insights can be observed. First, a sudden decrease of publicatios in year 2013, since it is a global effect, we hypothesise it could be explained by the financial crisis of 2008 and its consequent funds freezing that might have lead to a period of poor publishing power that was reverted several years later once funds were unfrozen. Second, a steady growth of AI publication on raising economies as South America and Africa. Finally, a raise of Asia in the last lustrum, when at the same time US and EU have loss relevancy.

6.2.4 Publications per economic powers

Of specific interest is the analysis of the economic super powers. In the following figure 6.5 we represent the yearly scientific production of AI (DL+ML) for US, EU and China in the total and relative amount of published articles.

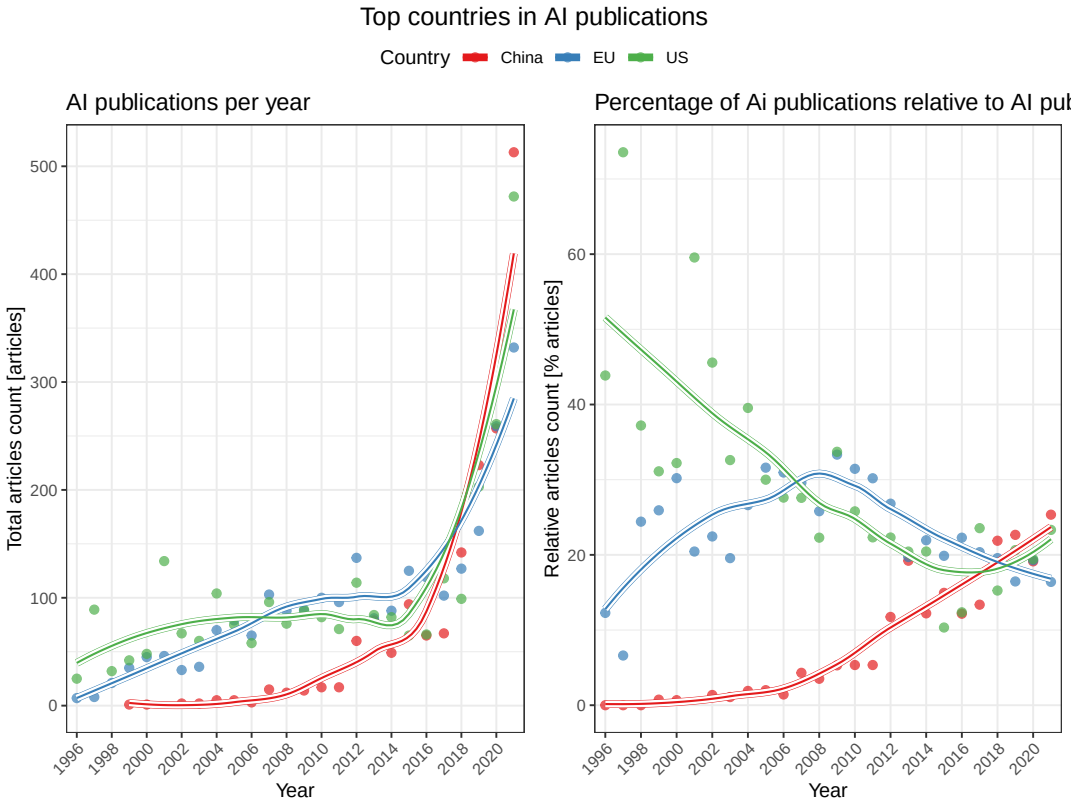

FIGURE 6.5: Tendency of the superpowers

It can be observed an exponential-like tendency in the last lustrum on AI publication. Although the three super powers have similar tendencies it is clear to the eye that the growth rates are not the same, being China the fastest. This is clear when observing the relative percentage. This observation projects a future with clear dominance of China in the AI space.

6.3 Semantic analysis

We used BioBert to project all the abstracts retrieved from PubMed to its embedding. Once projected, we used Principal Components analysis to analyse the trend of the publications. The following figure 6.6 represents a scatter plot of the first and second principal components of the embedding per each year.

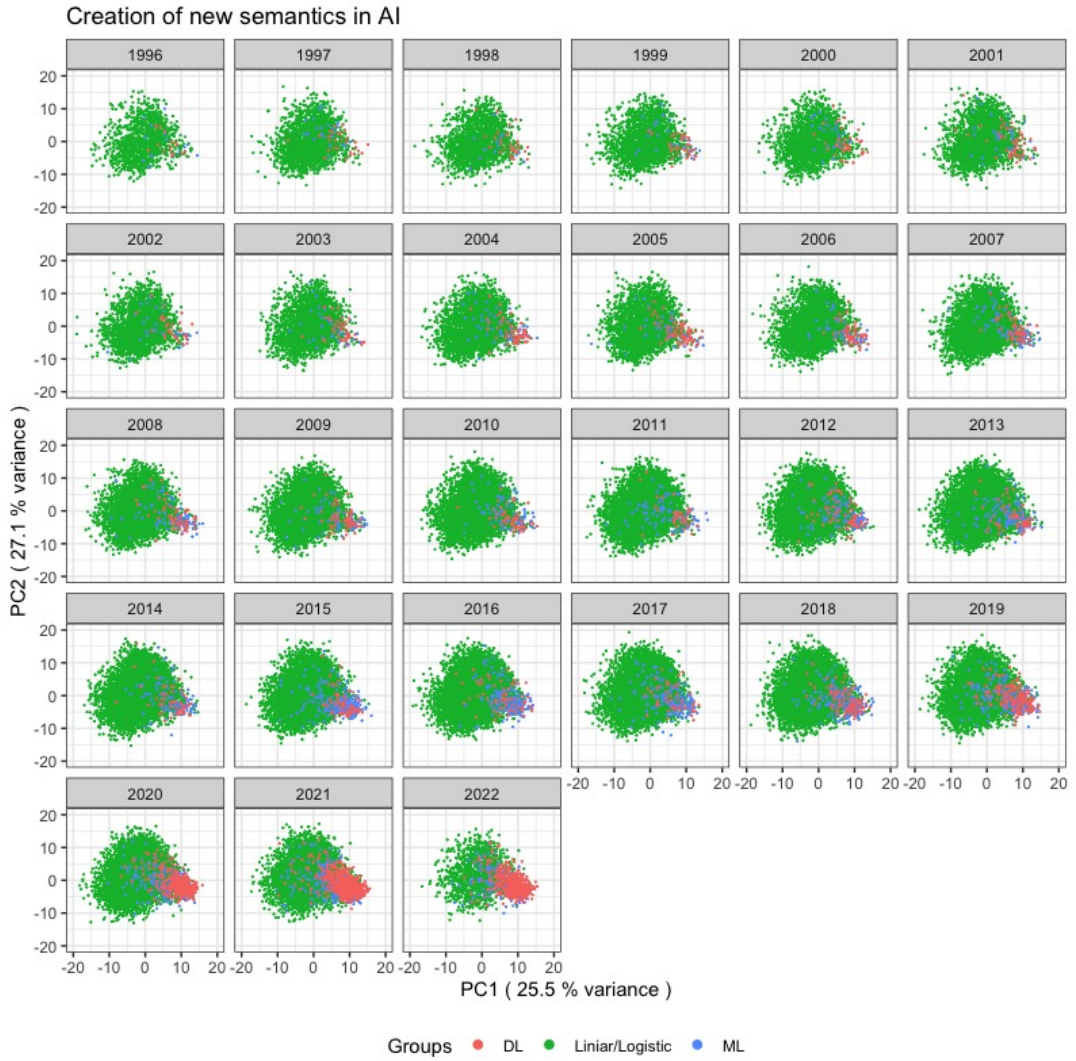

FIGURE 6.6: PCA representation of BioBERT embedding of the abstracts

First it can be seen that, AI was already present over two decades ago but that it has not been until recently that it's dominance is becoming clear. Second, it can be observed that during near to two decades, the *scientific semantic space* (ie. trated topics and how they are written) are quite constant. Finally, it is clear that AI field is a new science with a very concrete semantic space that differs of what has been dominating for two decades. This might pose a challenge to AI adoption (eg. communication issues).

Chapter 7

Conclusions

7.1 Evolution of AI in the health space

Interest in AI has raised exponentially in the last years. It is possible to observe this in the amount of scientific publications in the field. We showed that in the last lustrum there is a consistent bias of authors and journals towards AI in detriment of classical statistics. Similarly, the use of DL is increasing on all healthcare verticals. These results are indicative of the entrance of complex algorithms in the health space. This complex modeling allows to increase the performance of the predictive systems but it comes to the cost of a loose in interpretability. While linear and logistic models are more easily interpretable by health professionals, deep insights of advanced AI models like DL models are, sometimes, not even understood only by professionals specialised in AI. Furthermore, in AI interpretability is not always possible. This implies that the receiver of the output of the model can not infer what factors determine the decision or prediction taken by the model and thus might experiment difficulties in the consequent decision-making processes.

From the lenses of risk assessment, multiple risks arise. First, if the mechanisms used by the AI system to make decisions can not be understood and are not transparent, it might generate mistrust and poor adoption in the clinical practice. Second, poor interpretability and complexity of the AI systems might induce automation bias. Automation bias occurs when the human over-relies in the AI system and does not use its professional judgment. For example, the tendency of the medical professionals to believe in the output of the models as ground truth, without taking further consideration in all the relevance the contextual information has. This can be critical for patient safety. The latter has a secondary effect which is the loss of professional abilities.

Hence, the more complex the algorithms are the less interpretable. The less interpretable the more risk of poor adoption and risks as automation

bias, and in consequence potential loose of expert habilities. Being observed the exponential growth towards complex AI systems one must think the need to exponentially increase our habilities to add transparency to the AI decision making process.

7.2 Centralization of AI science

Competition between countries happens across multiple fields, from political alliances to competing for resources, military superiority, or financial and technological dominance. Similarly countries have historically acknowledged that technological dominance is one of the most important keys to global hegemony, as has historically happened with the steam engine in the industrial revolution or the Manhattan project in the second world war. We are now witnessing a running race between countries for AI dominance.

In this report we have shown that Asia is currently leading the race of AI scientific content generation in the health space with over 35% of the share of publications. In contrast, the publishing growth of North America and Europe in 2021 seems to be slowing down.

This results support the hypothesis of a shift in global AI research leadership from the US and EU to China as the new AI superpower. Some risks can be derived, first there is a cultural risk regarding the performance-privacy trade-off. Chinese culture towards data privacy highly differs from the western world, and it might be the case of highly accurate systems wanting to enter the western market but at a cost of large privacy invasion.In addition, China will gain a competitive edge in AI development due to their more relaxed regulation. Since, advances on AI are as compound interest, this might lead to total AI dominance by China in the short term.

On the other hand, highly complex systems as LLMs, cost bilions and a large amount of computational resources to train. As it happened with the atomic bomb, countries are now rushing to train LLM systems once seen the potential derived from ChatGPT. All these LLMs are controlled by a few private actors that will centralize the information of all its users. Some of these LLMs have long term memories allowing to store in its weigths private information. In the case of healthcare, this could imply that private medical information could be send in mass to a few private companies controlled by a few states or corporations. Similarly, the existance of only a few massive predictive models controlled by a reduced set of priviledged actors opens the door to massive societal manipulation.

Currently US holds the lead on LLMs and would be the principal actor that might gain control over clinical data in the mid term if no actions are performed.

7.3 AI semantic analysis

The use of LLM embeding has provided us with an unbiased approach to evaluate the semantic drift of the field. We observed an emerging cluster in the last lustrum that was differentiated from the classical statistics and representative only of the AI articles. This result implies that AI is semantically different to all the data science that has been used in the health space for the last two decades. While this phenomenon demonstrates that the introduction of AI in health is expanding the knowledge boundaries it also poses certain risks to the integration of AI-driven research into clinical practice.

First, there is a growing concern that clinicians may find themselves excluded from AI-driven research topics. This exclusion has the potential to diminish the real-world impact of AI-driven research and confine its significance to theoretical realms. Without active clinician involvement, research topics may lack the necessary practical insights, resulting in a gap between AI-generated knowledge and its application in clinical settings. This divide can hinder the translation of AI advancements into meaningful improvements in patient care and outcomes.

To address the risks associated with the generation of this new semantic space, several measures can be considered. First, foster interdisciplinary collaboration between AI researchers and clinicians is crucial. This collaboration promotes a holistic understanding of research topics, ensures the integration of clinical insights, and facilitates the alignment of AI-generated knowledge with real-world medical practice. Second, efforts should be made to enhance the interpretability and explainability of AI models, enabling clinicians to better understand the reasoning behind AI-generated recommendations and increasing trust. Finally, establishing robust mechanisms for ongoing validation and verification of AI-driven research findings is essential to ensure the reliability of AI systems.

Chapter 8

Further work

- Explore new forms of semantic extraction from free textsclusters, such as cluster-to-cluster centroid distance, intracluster mean distance or cluster centroid displacement over years.
- Publish the findings in a journal to acheve open and reproducible goals.

Chapter 9

Appendix

9.1 Code for data retrieving through PubMed API

```
def abstract_download(queries=None):
    import requests
    from xml.etree import ElementTree
    import pandas as pd
    import numpy as np
    from itertools import chain
    import time
    import traceback
    from BB_functions import printProgressBar,search_pubmed


    if queries is None:
        query_list = pd.read_csv('AA_Randomised_queries.csv')
            #canviar

        query_list = list(query_list.loc[:,'queries'])
    else:
        query_list=queries

    print(query_list)
    length = len(query_list)


    saving=1
    start = 0
```

```
    #canviar abans d'executar
num_savings=10

n_articles = []
abstract_list=[]
query_abstract =[]
iterator_saving = []


printProgressBar(0, length, prefix = 'Progress:', suffix = '
   Complete', length = 50)
for i in range(0, length):
    for year in range(2000,2024):

        if year == 2000:
            year=' AND 0000:2000[dp]'
        else:
            year = ' AND '+str(year)+'[dp]'


        query = str(query_list[i]+year)

        print(query)
        abstract, num_articles = search_pubmed(query, retmax
           =100,retstart=0)



        n_articles.append(num_articles)
        abstract_list.append(abstract)
        query_abstract.append(query_list[i])

        if num_articles>99:
            for j in range(1, int(num_articles/99)):
                    #time.sleep(5)
                abstract, num_articles = search_pubmed(query,
                   retmax=100,retstart=100*j)

                n_articles.append(num_articles)
```

```
            abstract_list.append(abstract)
            query_abstract.append(query_list[i])
            time.sleep(1)



    if (i==(int(saving*length/num_savings)) or (i == length-1)
       ):


       try:

           abstract_df = pd.DataFrame(list(chain.from_iterable
              (abstract_list)))
           narticles_df = pd.DataFrame([query_abstract,
              n_articles]).transpose()
           iterator_saving.append([saving, i])
           iterator_savings = pd.DataFrame(iterator_saving)
           iterator_savings.columns=['saving','i_iteration']


           narticles_df.to_csv('arxives/n_articles_'+str(start
              +saving)+'.csv')
           abstract_df.to_csv('arxives/abstracts_'+str(start+
              saving)+'.csv')
           iterator_savings.to_csv('arxives/iteration_'+str(
              start+saving)+'.csv')

           abstract_list=[]
           query_abstract=[]
           n_articles=[]
           iterator_saving=[]

           print('saved'+str(saving))
           saving = saving+1

       except Exception:
           print(f'unsuccessful saving, iteration: {i}')
```

```
            traceback.print_exc()


        printProgressBar(i + 1, length, prefix = 'Progress:',
            suffix = 'Complete', length = 50)




import requests
from xml.etree import ElementTree
import pandas as pd
import numpy as np
from itertools import chain
import time
import traceback



def search_pubmed(query, retmax=1000,retstart=0):

    abstracts = []
    try:
        base_url = 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils
            /'
        search_url = base_url + 'esearch.fcgi'
        fetch_url = base_url + 'efetch.fcgi'
        params = {
            'db': 'pubmed',
            'term': query,
            'retmax': retmax,
            'retstart': retstart
        }
        r = requests.get(search_url, params=params)
        root = ElementTree.fromstring(r.content)
        total_results = int(root.find('.//Count').text)
        id_list = [e.text for e in root.findall('.//Id')]
        for i in range(0, len(id_list), 100):
```

```
params = {
    'db': 'pubmed',
    'id': ','.join(id_list[i:i+100]),
    'rettype': 'xml',
    'retmode': 'xml'
}
r = requests.get(fetch_url, params=params)
root = ElementTree.fromstring(r.content)
for article in root.findall('.//PubmedArticle'):
    pmid = article.find('.//PMID').text
    doi_elem = article.find('.//ArticleId[@IdType="doi
       "]')
    doi = doi_elem.text if doi_elem is not None else ''
    abstract_elem = article.find('.//AbstractText')
    abstract = abstract_elem.text if abstract_elem is
       not None else ''
    title_elem = article.find('.//Title')
    title = title_elem.text if title_elem is not None
       else ''
    year_elem = article.find('.//Year')
    year = year_elem.text if year_elem is not None else
        ''
    authors = [a.text for a in article.findall('.//
       AuthorList/Author/LastName')]
    affiliations = [a.text for a in article.findall
       ('.//AffiliationInfo/Affiliation')]
    abstracts.append({
        'pmid': pmid,
        'doi': doi,
        'journal': title,
        'abstract': abstract,
        'authors': authors,
        'affiliations': affiliations,
        'year':year,
        'query': query
    })
#print(f'Retrieved {len(abstracts)+retstart} of {
   total_results} abstracts. query: {query}')
```

```
        return abstracts, total_results
    except:
        total_results = -1
        abstracts.append({
                    'pmid': None,
                    'doi': None,
                    'journal': None,
                    'abstract': None,
                    'authors': None,
                    'affiliations': None,
                    'year':None,
                    'query': query
                })
        return abstracts, total_results

def printProgressBar (iteration, total, prefix = '', suffix = '',
     decimals = 1, length = 100, fill = '', printEnd = "\r"):
    """
    Call in a loop to create terminal progress bar
    @params:
        iteration - Required : current iteration (Int)
        total - Required : total iterations (Int)
        prefix - Optional : prefix string (Str)
        suffix - Optional : suffix string (Str)
        decimals - Optional : positive number of decimals in
            percent complete (Int)
        length - Optional : character length of bar (Int)
        fill - Optional : bar fill character (Str)
        printEnd - Optional : end character (e.g. "\r", "\r\n") (
            Str)
    """
    percent = ("{0:." + str(decimals) + "f}").format(100 * (
        iteration / float(total)))
    filledLength = int(length * iteration // total)
    bar = fill * filledLength + '-' * (length - filledLength)
    print(f'\r{prefix} |{bar}| {percent}% {suffix}', end =
        printEnd)
    # Print New Line on Complete
```

```
    if iteration == total:
        print()
```

Bibliography

[1] "The state of AI in 2022—and a half decade in review". In: (2022).

[2] Erik Brynjolfsson John Etchemendy Katrina Ligett Terah Lyons James Manyika Helen Ngo Juan Carlos Niebles Vanessa Parli Yoav Shoham Russell Wald Jack Clark Nestor Maslej Loredana Fattorini and Raymond Perrault. *The AI Index 2023 Annual Report*. eng. Report. Stanford: Stanford University, Apr. 2023.

[3] High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. eng. Report. Brussels: European Commission, Apr. 2019.

[4] "EU White paper on AI". In: (2020).

[5] "EU AI Act". In: *European Comission* (2021).

[6] Comisión Europea and Contenido y Tecnologías Dirección General de Redes de Comunicación. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020. DOI: doi/10.2759/002360.

[7] "PubMed". In: ().

[8] "MEDLINE PubMed Production Statistics". In: (2023).

[9] *Help - PubMed*. 2023.

[10] G. van Rossum. *Python tutorial*. Tech. rep. CS-R9526. Amsterdam: Centrum voor Wiskunde en Informatica (CWI), 1995.

[11] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87 –90.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.

[13] RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC., 2020.

[14] Charles R. Harris et al. *Array programming with NumPy*. Sept. 2020. DOI: 10.1038/s41586-020-2649-2.

[15] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134.

[16] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Nov. 2019.

[17] Plotly Technologies Inc. *Collaborative data science*. Montreal, QC, 2015.

[18] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[19] Hadley Wickham. "The Split-Apply-Combine Strategy for Data Analysis". In: *Journal of Statistical Software* 40.1 (2011), pp. 1–29.

[20] Pritam Deka and Anna Jurek-Loughrey. *Unsupervised Keyword Combination Query Generation from Online Health Related Content for Evidence-Based Fact Checking*. 2021.

[21] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[22] Jinhyuk Lee et al. "Data and text mining BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: (2021). DOI: 10.1093/bioinformatics/btz682.

[23] "BERT Explained: State of the art language model for NLP". In: (2021).

[24] Amisha et al. "Overview of artificial intelligence in medicine". In: *Journal of Family Medicine and Primary Care* 8 (7 2019), p. 2328. DOI: 10.4103/JFMPC.JFMPC_440_19.

[25] Isaac Kofi Nti et al. "Applications of artificial intelligence in engineering and manufacturing: a systematic review". In: *Journal of Intelligent Manufacturing 2021 33:6* 33 (6 Apr. 2021), pp. 1581–1601. DOI: 10.1007/S10845-021-01771-6.