Organisation for Economic Co-operation and Development

NESTI 2018 - ROOM DOCUMENT

For Official Use **English - Or. English**

Identifying AI-based projects and expenditure in health R&D

An initial exploration based on NIH project funding data

Working Party of National Experts on Science and Technology Indicators
25-26 October 2018
OECD, Paris

This room document is presented to participants in the 2018 NESTI meeting under Item 6d of the draft agenda on measuring digitalisation. It also provides relevant background information for Items 7b and 8d on R&D statistics in the context of the Frascati Manual and work on R&D funding indicators and analysis.

For further information, please contact: Economic Analysis and Statistics Division.
Izumi Yamashita (Izumi.Yamashita@oecd.org); Fernando Galindo-Rueda (Fernando.Galindo-Rueda@oecd.org).

Table of contents

Tables

Figures

Identifying AI-based projects and expenditure in health R&D: An initial exploration based on NIH project funding data

1. Introduction and background

1. This document reports on the procedures and initial findings from a text-based analysis of project level funding data for health R&D to measure the extent of Artificial Intelligence-related R&D. This study seeks to assess whether it is possible to provide estimates of the proportion of R&D funding in a key area like health that is supporting work that draws on or contributes to the advancement of Artificial Intelligence (AI).

2. This paper uses a quantitative case study approach, applying a set of quantitative tools to identifying AI related research to a specific database for illustration purposes. The project level funding data of the US National Institutes of Health provide a useful and relevant demonstration ground as NIH funding is allocated to health R&D in Government budget R&D statistics and accounts for approximately 95% of all R&D under the responsibility of the US Department of Health and Human Services. These data are funder based, but the information on projects is provided by R&D performers as they submit applications. Project level data, provided it is comprehensive (i.e. not all R&D funding is channelled via projects leaving a microdata record trace), can be an extremely rich source of information. The purpose of this study is therefore twofold:

- To inform the discussion on the widening scope of AI R&D and its role enabling R&D in different policy areas.
- To support the ongoing NESTI project seeking to assess the feasibility of constructing a multi-country database on project funding for analytical purposes (Fundstat), including those concerning the identification of emerging domains.

3. This work reflects the widespread demand for data resources, tools and methods that help identify features of R&D funding in thematic areas that are not easily captured by pre-defined, often static taxonomies.

4. This project is part of the wider OECD efforts to measure Artificial Intelligence in scientific articles, patents, and other relevant data. It is also a case study to utilise text analysis enabled by AI-based techniques.

5. There is a wide and fast growing literature dealing with topic extraction on research fields drawing on several corpora, mostly publications, with some efforts looking at AI, as document in DSTI/CIIE/WPIA(2018)4. The increasing public availability of project level funding data is also enabling related efforts looking specifically at funding, and funders themselves as well commercial compilers of such data are also enabling semantic search functionalities [Drafting note: specific references to be added]. Some studies have looked in general at the classification of NIH funding (Park et al., 2016[3]; Talley et al., 2011[4]), NSF (Kawamura et al., 2018[5]; Freyman, Byrnes and Alexander, 2016[6]), and EU's FP7 (Kawamura et al., 2018[5]), but not on AI-related research specifically. A number of policy documents have drawn attention to the level of efforts made at advancing research on AI. A White House report indicated that the United States invested USD1.1 bn USD in "AI R&D" in 2015 and (NSTC, 2016[7]). The EU has spent around 13% of R&D budget in ICT

since 2014 (EC, 2018[8]). The Engineering and Physical Sciences Research Council (EPSRC) of the UK has allocated 300M GBP to fund research related to data science and AI (BEIS and DCMS, 2018[9]). The perspective on the use of AI to support and possibly transform R&D more broadly has been somewhat absent from the discussion.

6. This document is structured as follows. Section 2 describes the data used. Section 3 describes the methodology applied to identify AI related R&D funding and test its robustness. Section 4 presents the key results. Section 5 concludes with a series of possible next steps.

2. Data

7. This analysis deals with data on grant funding for projects in the NIH database (NIH RePORTER) between FY2001 to FY2017. NIH was estimated to have close to USD31bn worth of obligations for R&D, with an outlay of USD 28bn. The data were downloaded from the NIH ExPORTER website in csv format in June 2018. The data contain information provided in the project application or generated in the administrative process such as application ID and project number, project descriptions (project title, fiscal year, funding amount, abstract, project terms, NIH spending categories, and funding mechanism), and beneficiary information (PI ID, PI name, organisation name. Major attributes for the RePORTER data are as in **Table 1.1**. Founding amounts for 2017 correspond to close to USD 34 bn worth of R&D funding for intra and extramural projects extending over multiple years,

Table 1.1. Fields included in the NIH project data analysed

RePORTER Items	Descriptions
Application ID	Identification number for an application. Application is unit of observation in the database; a research project may contain multiple associated applications.
Project number	Identification number for a project.
Project title	Title of the project
Fiscal year	A 12 months period in which funding to the application is noticed.
Funding amount	The total amount of money allocated to the project by the government agency, in USD.
Abstract	Description of the contents of the project
Project terms	Terms that represent the characteristics of the project, automatically tagged after 2008.
NIH spending categories	Classification given by the NIH based on Research, Condition, and Disease Categorization (RCDC) taxonomy
Funding mechanism	The instrument through which the funding to the project is made
PI ID(s)	Identification number(s) for the principal investigator(s) of the project
PI Name(s)	Name(s) of the principal investigator(s) of the project
Organisation name(s)	The name(s) of organisation to which the PI(s) belong or the funding is distributed

8. The tagging of records is an important source of information. Yet the thematic classification item (NIH spending categories), which is based on the Research, Condition, and Disease Categorization Process is not an appropriate source of identifying information on AI because it does not have AI based categories. The set of project terms is also

insufficient on its own. The coverage of AI related terms does not seem fully comprehensive and tagging has not been consistent over the time period, as projects were manually tagged from 2001 to 2007 and automatically from 2008 to 2016). Furthermore, some documents do not include project terms.

9. The NIH MeSH (Medical Subject Headings) is a hierarchically organised set of keywords set managed by one of the NIH institutes (U.S. National Library of Medicine). It contains a heading for "**Artificial intelligence**", which defines it as "theory and development of computer systems which perform tasks that normally require human intelligence. Such tasks may include speech recognition, learning; visual perception; mathematical computing; reasoning, problem solving, decision-making, and translation of language". **Table 2.2** provides a description of the hierarchical structure in which AI appears and the subheadings underneath. Within MeSH, AI features in two separate domains, under "Mathematical concepts" and "Information science"

Table 2.2. MeSH tree structure for Artificial intelligence

Mathematical Concepts [G17]
Algorithms [G17.035]
Artificial Intelligence [G17.035.250]
Machine Learning [G17.035.250.500]
Supervised Machine Learning [G17.035.250.500.500]
Support Vector Machine [G17.035.250.500.500.500]
Unsupervised Machine Learning [G17.035.250.500.750]
Information Science [L01]
Computing Methodologies [L01.224]
Algorithms [L01.224.050]
Artificial Intelligence [L01.224.050.375]
Computer Heuristics [L01.224.050.375.095]
Expert Systems [L01.224.050.375.190]
Fuzzy Logic [L01.224.050.375.250]
Knowledge Bases [L01.224.050.375.480]
Biological Ontologies [L01.224.050.375.480.500]
Gene Ontology [L01.224.050.375.480.500.500]
Machine Learning [L01.224.050.375.530]
Supervised Machine Learning [L01.224.050.375.530.500]
Support Vector Machine [L01.224.050.375.530.500.500]
Unsupervised Machine Learning [L01.224.050.375.530.750]
Natural Language Processing [L01.224.050.375.580]
Neural Networks (Computer) [L01.224.050.375.605]
Robotics [L01.224.050.375.630]

Source: U.S. National Library of Medicine NIH. Extracted on 28 September 2018 from https://meshb.nlm.nih.gov/record/ui?ui=D001185.

10. Subject subheadings include Biological Ontologies, Computer Heuristics, Expert Systems, Fuzzy Logic, Gene Ontology, Knowledge Bases, Machine Learning (including Supervised Machine Learning and Unsupervised Machine Learning), Natural Language Processing, Neural Networks (Computer), Robotics and Support Vector Machine. As it can be noted, this structure does not provide alone a comprehensive source of all potentially

relevant AI terms but provides a basic structure for the categorisation of research activity and outputs in the health domain.

11. Unfortunately, the MeSH research tagging system is not yet systematically applied to funding applications and therefore such information is not available within the RePORTER data system. MeSH is principally applied to scholarly publications listed in MEDLINE, PUBmed and related sources, while the funding data do not contain readily available information on publication outputs associated to the funded projects.

3. Methodology

3.1. Method selection

12. Because of the lack of systematic tagging information within the RePORTER project-level data relevant to the objectives of this study, the analysis relies entirely on the text information (title, abstract and project terms[1]). This helps demonstrate its potential application to other project funding databases.

13. After consideration of several options, the analysis takes a keyword matching approach for text mining and identification of AI-based research projects. Topic modelling was discarded after an initial attempt, because it proved particularly challenging to identify entirely AI-based topics in the data. Automatically generated topics were ambiguous with respect to their AI content as the semantic weight of the health objectives of the research dominated over the scarcer information about AI-based methods used in the research. Topic identification from internal or external linkages to other data (e.g. citations) was not possible either.

14. Keyword matching presents a number of challenges since, as previously explained, there is at present no standard set of keywords available to represent AI-based health research and such standard is bound to vary over time. Failing to capture all possible relevant keywords poses an underestimation risk due to false negatives. This problem can also arise in the case when the title, abstract and terms of a project application do not contain detailed information on the research methodology to be used in the project. This is an underlying data problem which may be particularly acute when the abstracts focus on outlining the expected outcomes and AI plays an enabling role within a project. Ideally, the corpus should contain a "methods" section to facilitate a more effective data mining process as well as a better understanding of the role played by AI in health R&D.

15. Conversely, keyword matching from a predefined menu does not assure that the research is AI-based (a risk of false positives), as some potential AI-keywords may also describe other research domains – the most common being the term "neural networks" that was adopted by the AI field from the neurology domain. The use of terms relating to

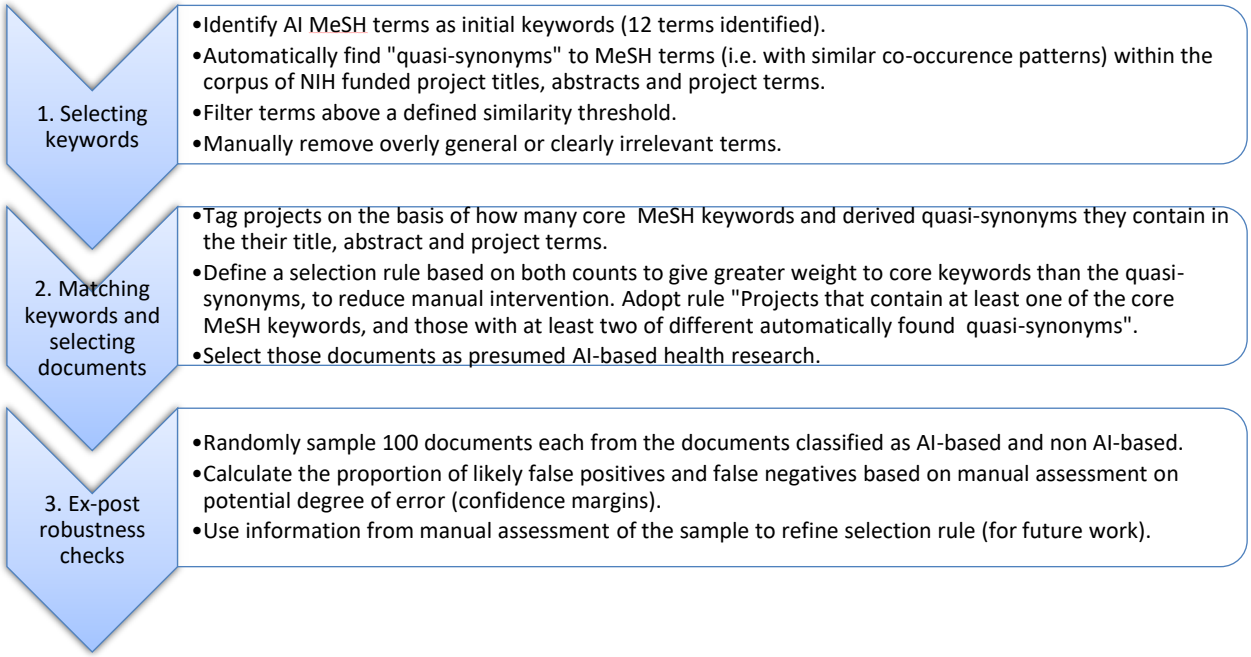
[1] Project terms could be systematic tagging information, which is not the case for the RePORTER data as mentioned before; they are used as one of the text information in the analysis.

statistical analysis tools known more many decades and recently popularised for AI applications (e.g. Bayesian analysis) may result in non AI-based projects being captured.

3.2. Method description

16. The text mining process adopted is outlined in **Figure 3.1**. The process consists of 1. Selecting keywords, 2. Matching keywords and selecting documents that are AI-based, and 3. Ex-post checks.

Figure 3.1. Text mining process applied to NIH grant application data



17. As a starting point of the process 1, we consider the MeSH-AI terms presented in **Table 2.2**, not as tags (the documents are not tagged with them) but as reference for the first stage of text mining. Two terms (Biological Ontologies and Computer Heuristics) are dropped of the reference as they do not appear in the corpus; Neural Networks (Computer) is changed to Artificial Neural Networks to limit its meaning as neural networks is much used in neurological sense. Other lists may be considered in the future.

18. Following initial basic cleaning of the corpus to convert the text into lower case, remove punctuation, stopwords and processing hyphenated terms, an un-supervised machine learning methodology (word embedding) was used to identify additional AI keywords as those with high co-occurring patterns, as potential quasi-synonyms. This process generates vector representations of terms in a group of documents based on the term co-occurrence patterns with using a neural network algorithm. The analysis then produces cosine similarities measures for each pair of selected terms.[2][3]

[2] Cosine similarity is an index that ranges between -1 (completely opposite) to +1(identical).

[3] The texts have been analysed using Word2vec functions in Python's Gensim package for word embedding after being tokenised by Parses function in the same package. For creating a Word2vec

19. The analysis sets a minimum threshold of cosine similarity as +0.6. This is plausible baseline considering the ability of the algorithm to identify quasi-synonyms; as the cosine similarity drops, the likelihood of retrieving terms that have irrelevant meanings (noise) rises. The highest cosine similarity found was less than +0.8, implying that similarities implied by co-occurrence are not strong in a majority of cases. This arguably reflects that AI is not yet mainstreamed into projects (only 0.92% of the documents include the initial MeSH terms) and the weakness of the underlying signal in the text descriptions due to the absence of detailed data on methods. The approach does nonetheless allow to identify a list of particularly relevant terms in the health research context.

20. The automated procedure yields a total of 100 terms. 25 duplicated terms are removed or merged. For example, "machine learning techniques" (duplication with "machine learning") is removed; "bioinformatic" and "bioinformatics" are merged. A total of 7 abbreviations are also removed to avoid duplication. Then 32 terms found to have overly too general meanings were manually removed as listed in the **Table A.2**. Finally 36 terms are identified as additional AI keywords as listed in the **Table A.1**.

21. The 36 quasi-synonyms are penalised as having half less impact than MeSH keywords when selecting funding documents to represent AI-based health research. The documents that meet one of the following criteria are identified as ones for AI-based health research projects. The title, abstract, or project terms must therefore include (a) at least one MeSH keywords or (b) at least two quasi-synonyms.

4. Results

4.1. AI keywords and their distribution

22. The full list of derived keywords is available in **Annex 1**. **Table 4.1** shows the list of presumed AI keywords that appear in more than 1 000 documents corresponding to granted applications to NIH. It shows that the "bioinformatic" is by far the highest incidence followed by "data acquisition" and "robot". Only "machine learning" and "robotics" are in the list from the original MeSH keywords. Despite their significant co-occurrence patterns with AI core terms, some terms are clearly not unambiguously related to AI, such as multi-dimensional or data acquisition, indicating that there is a significant risks of overstating the estimated share of AI R&D in the data if the requirement of having at least two of such derived keywords is not strong enough.[4][5] The robustness analysis described below will provide an indication of the size of the potential upward bias.

---

model, the vector size has been set to 100 using terms that at least appear twice and setting the window size (the range to learn term co-occurrence) as 5. CBOW is used for the training algorithm. Only initial applications to each project are used for building the model to avoid duplications, as following applications are often similar in their contents and may distort results.

[4] It may be also questioned whether all bioinformatics is necessarily related to AI although it often deals with big data sources, and as noted above, Bayesian is a generic statistical term.

[5] The terms manually removed from the automatically generated list of quasi-synonyms are presented in **Annex 2**.

Table 4.1. Frequently appeared AI keywords

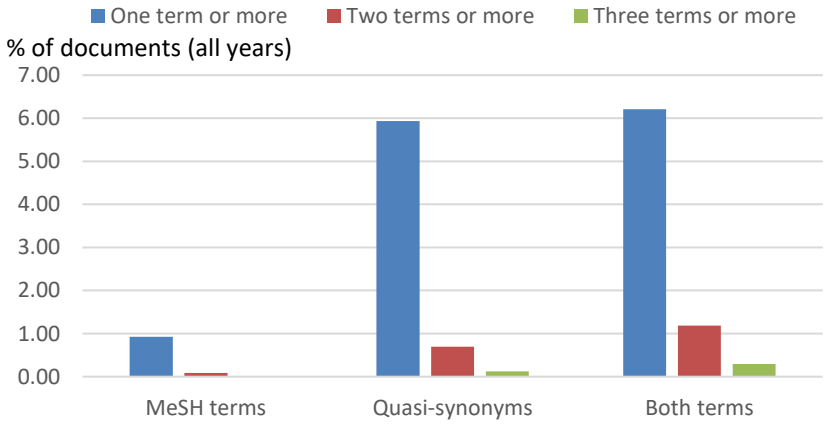
Keywords	number of documents with AI keyword incidence
bioinformatic	40,399
data acquisition	7,817
robot	5,347
machine learning	5,165
pattern recognition	4,206
electronic medical records	4,158
robotics	3,892
bayesian	3,769
data mining	3,614
ontologies	2,763
decision support	2,169
high dimensional	2,101
big data	1,641
multi dimensional	1,593

Source: OECD calculations based on NIH RePORTER data.

23. The first set of results is an overview of the distribution of AI-related keywords. **Figure 4.1** shows the distribution of AI keywords (for MeSH terms and related co-occurring terms) in the documents analysed. Less than 1% of the documents contain MeSH terms. There are very few documents that contain more than one MeSH terms. Around 6 % of the documents contain the related terms (quasi-synonyms) identified by the text analysis; less than 1% of the documents contain more than two quasi-synonyms.

Figure 4.1. Distribution of AI keywords: MeSH and related terms

As percentage of documents corresponding to granted applications



Source: OECD calculations based on NIH RePORTER data.

4.2. Volume of AI-based health research measured by number and funding amount

24. The volume of AI-based health research is assessed based on the number of observations (granted applications for the NIH database) and the amount of funding allocated to them. **Table 4.2** summarises the volume of AI-based health research from 2001

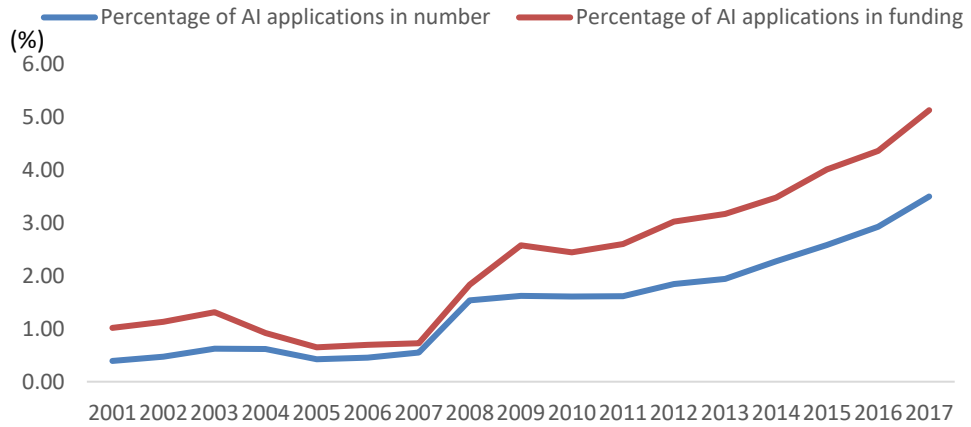
to 2017. The estimated share of AI-based health research in number has increased from 0.39% in 2001 3.50% in 2017. The share in funding amount represents a similar tendency, from 1.02% in 2001 to 5.13% in 2017. Figure 4.2 presents the growth profile over time, which shows and acceleration in 2007 to 2008.

Table 4.2. Volume of estimated AI-based health research

Year	Number of granted applications - AI	Number of granted applications - ALL	Percentage of AI grants (% of counts)	Funding amount - AI (USDm)	Funding amount - ALL (USDm)	Percentage of AI-related funding
2001	306	78,033	0.39	180	17,735	1.02
2002	379	79,912	0.47	221	19,618	1.13
2003	363	58,049	0.63	259	19,744	1.31
2004	458	74,367	0.62	200	21,917	0.92
2005	328	77,050	0.43	150	23,260	0.65
2006	345	76,195	0.45	161	23,309	0.69
2007	448	81,697	0.55	216	29,817	0.73
2008	1,223	79,576	1.54	545	29,829	1.83
2009	1,453	89,583	1.62	915	35,494	2.58
2010	1,353	84,195	1.61	896	36,689	2.44
2011	1,183	73,320	1.61	816	31,376	2.60
2012	1,267	68,610	1.85	925	30,623	3.02
2013	1,290	66,451	1.94	921	29,074	3.17
2014	1,492	65,540	2.28	1,040	29,919	3.48
2015	1,700	65,821	2.58	1,207	30,095	4.01
2016	1,960	66,944	2.93	1,394	31,993	4.36
2017	2,387	68,235	3.50	1,729	33,712	5.13

Note: These figures are results of keyword matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NIH RePORTER data.

Figure 4.2. Percentage of NIH funded applications identified as AI-based



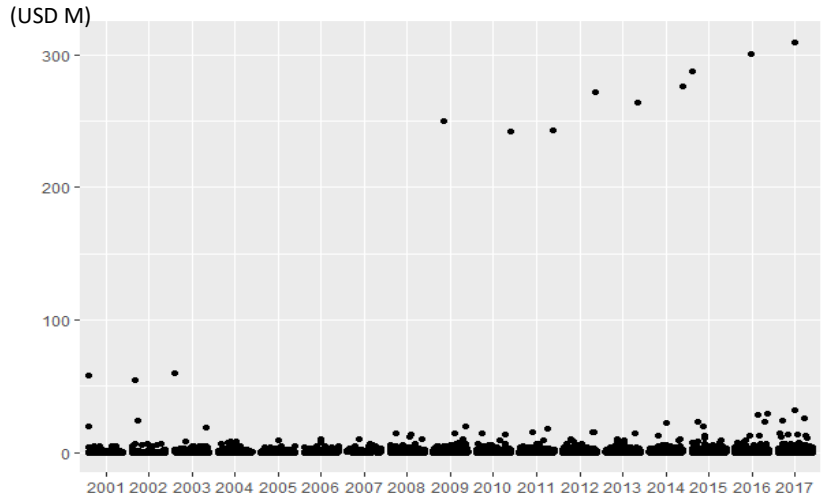
Source: OECD calculations based on NIH RePORTER data.

4.2.1. Accounting for potential outliers

25. As it appears in the **Figure 4.2**, the percentages of the number and funding amount greatly differ in some period. **Figure 4.3** plots funding amount per application to better understand the trends depicted in **Figure 4.2**. It shows that there are exceptionally big

applications in their funding amount after 2009. Applications with more than USD 100 million funding are examined, as they have high impacts on estimated AI funding shares.

Figure 4.3. Funding amounts per application



Source: OECD calculations based on NIH RePORTER data.

26. The extreme values in excess of USD 100 million are accounted for by an infrastructure project with multiple applications assigned to them, namely "National Biomedical Information Services". After examining the project description, it was decided to retain it. The project is an intramural within one of the NIH institutes, the National Libraries of Medicine. It contributes to build and provide various information systems normally utilise theory or system on AI, such as natural language processing and large bibliographic databases. The project description is available in **Annex 3**.

27. The results do nonetheless confirm the sustained growth in AI-related funding over the period with a brief hiatus from 2004 to 2007. Growth is particularly fast both in terms of numbers of documents and funding in the 2010s, with no signs of deceleration in the final years for which data are available.

4.3. AI research across different health research domains

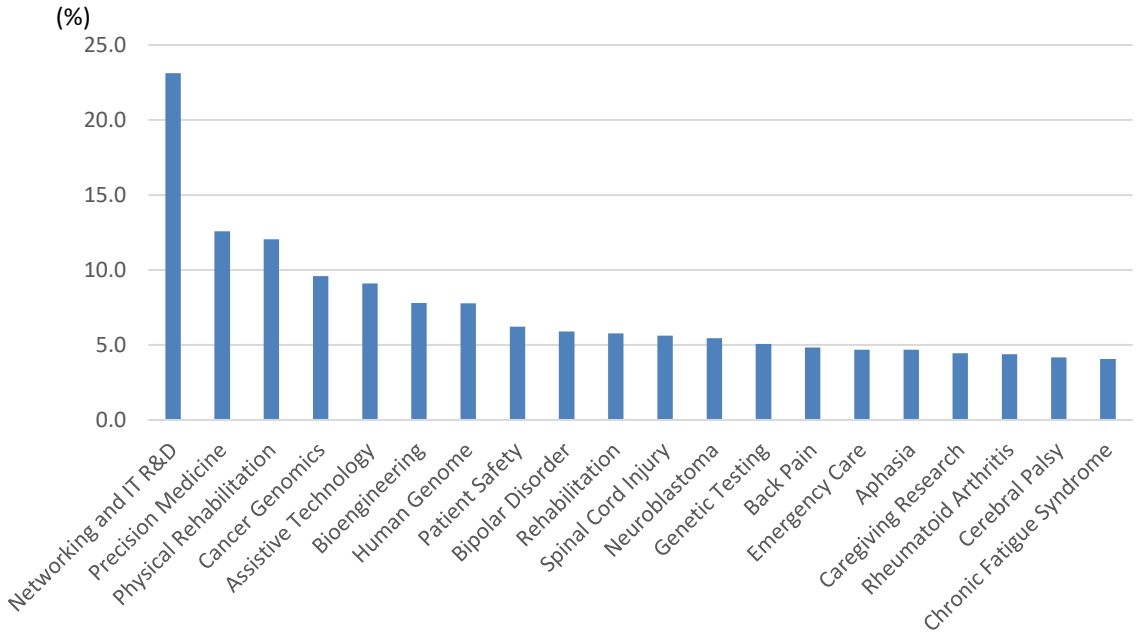
28. Equipped with a tentative assessment of which documents are likely to be AI-based, it is possible to analyse in more detail how these are distributed across the NIH funding portfolio. **Figure 4.4** shows the incidence of AI-based research within each NIH spending category. As expected, the Networking and IT R&D category[6] is the highest with 23.1% of research identified as AI-based over the entire period for which NIH spending category data is available (from 2008 to 2016). Among others, it is possible to note that categories related to gene analysis are prominent such as Precision Medicine (12.6%), Cancer Genomics (9.6%), Human Genome (7.8%), Bioengineering (7.8%) and Genetic Testing (5.1%). This is consistent with the importance of AI-based tools for examining the genetic

[6] "Networking and IT R&D" is the only category represents information technology, which is closely related to AI; most of the 285 categories are based on disease or technologies utilised in health research. Full list of the categories is found in the RCDC website.

code, the vast amount of information contained and predicting phenotypical expression data from genetic code data, to cite some examples. Genetic research has been a major testing ground for AI and data science in general. Categories involving mechanical development such as Physical Rehabilitation (12.1%), Assistive Technology (9.1%) and Rehabilitation (5.8%). This appears to reflect the role of robotics in developing smart assistive devices and the role of AI in supporting brain-machine interfaces.

Figure 4.4. NIH spending categories with largest estimated shares of AI-based R&D

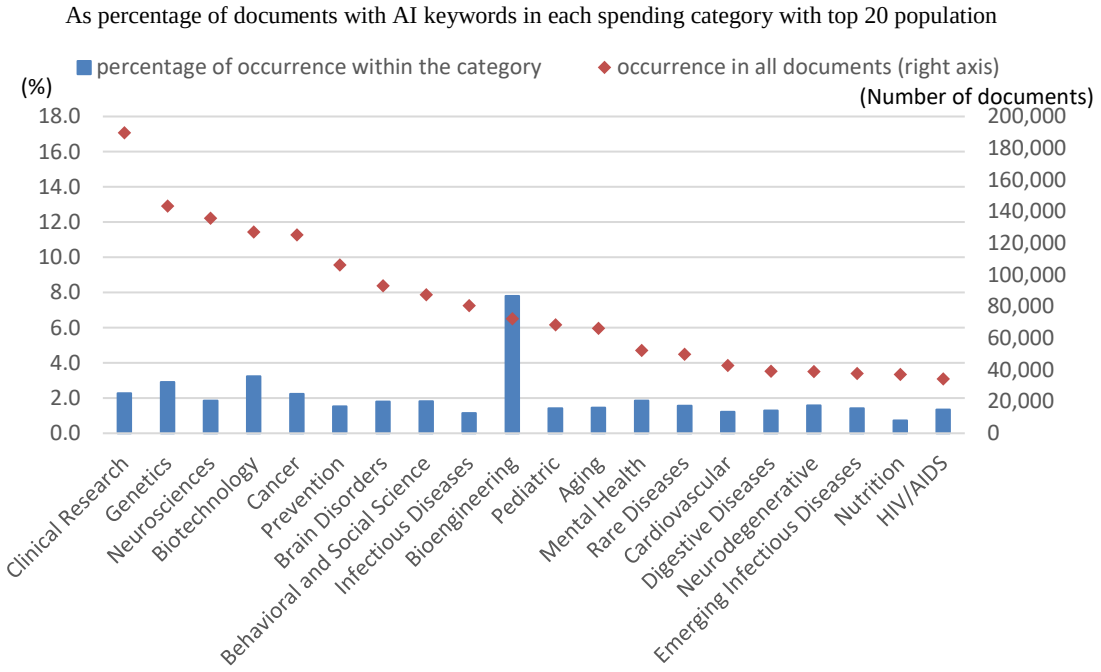
Percentage of AI-related documents within each spending category



Note: Only data between 2008 and 2016 are available. A document may have multiple spending categories, which are separately counted; sum of the percentages is not 100%.
Source: OECD calculations based on NIH RePORTER data.

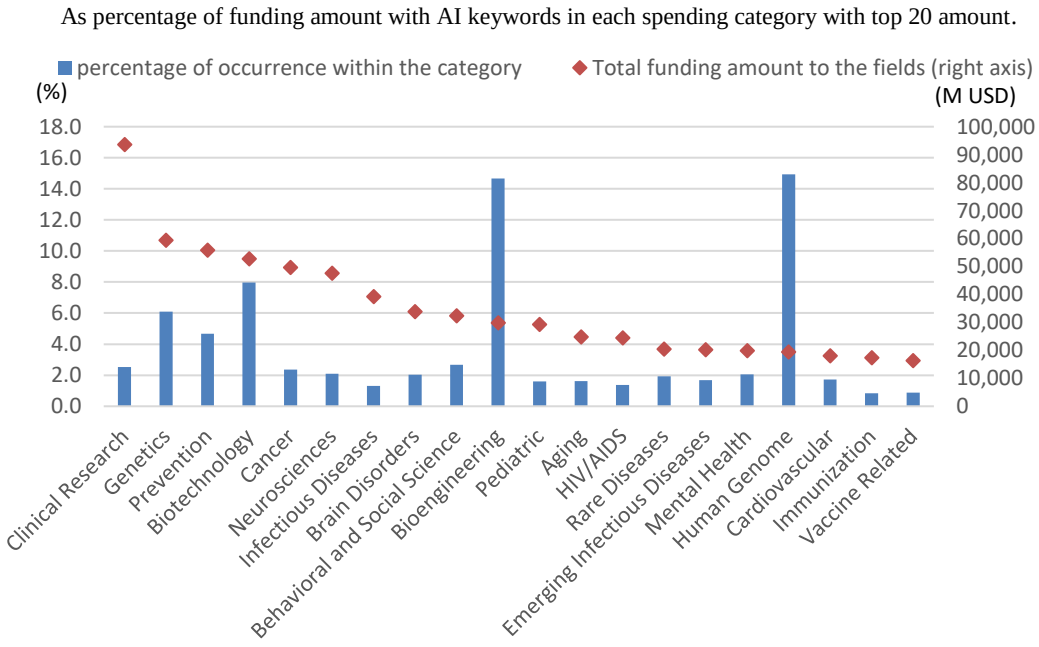
29. **Figure 4.5** and **Figure 4.6** show the occurrence of AI-based health research in the NIH spending categories sorted by share of the documents in the whole corpus measured by number of grant application documents (**Figure 4.5**) and by funding amount (**Figure 4.6**). Clinical Research is especially high in its number of documents and funding, The estimated share of AI-based health research within this area for the entire period is close to 2.5%. Genetics, Neurosciences, Biotechnology, Cancer and Prevention are major areas with significant use of AI. Comparing those figures, it is noticeable that the percentage of AI-based research is higher in funding amount than number of documents in most of the categories, especially in the area of Bioengineering and Human Genome.

Figure 4.5. Extent of AI-based health R&D projects in the NIH categories with more projects



Note: Only data between 2008 and 2016 are available. A document may have multiple spending categories, which are separately counted; sum of the percentages is not 100%.
Source: OECD calculations based on NIH RePORTER data.

Figure 4.6. Extent of AI-based health R&D funding in the largest NIH spending categories



Note: Only data between 2008 and 2016 are available. A document may have multiple spending categories, which are separately counted; sum of the percentages is not 100%.
Source: OECD calculations based on NIH RePORTER data.

4.4. Robustness checks

4.4.1. False positives

30. In order to assess the robustness of the results, especially in light of the presence of somewhat ambiguous keywords that may somewhat biases these estimates, an initial batch of 100 documents were selected at random from the 17,935 documents categorised as AI-based health research. They were manually classified into the four categories shown in the **Table 4.3**, following inspection of their title, abstract and the project terms.

- 32 documents mention the use or development of AI theories or systems[7] [8](A).
- 38 documents mentions that the project works on tasks normally achieved by using AI theories or systems[9] (B).
- 28 are likely to be classified as A or B, but there is not enough information to either confirm or deny it (C).
- 2 have contents not related to AI and are not likely to be classified as A or B (D).

31. If false positives are defined as those documents in the category C and D, the implied 95% confidence interval of false positive rate is from 0.21 to 0.40; if it is defined as proportion in the category D, the interval is from 0.00 to 0.07. Examples of the project descriptions of each category are shown in the tables in **Annex 4**.

Table 4.3. Classification for estimating false positive rate

Categories	Definitions	Number (and percentage) of documents in the categories
A	Mentions the use or develop of AI theories or systems.	32
B	Is not classified as A, but mentions that the project works on tasks normally achieved by using AI theories or systems.	38
C	Might be possibly classified as A or B, but there is not enough information to either confirm or deny it.	28
D	Is not likely to be classified as A or B.	2

Note: AI theories and systems refer to the ones in the definition of artificial intelligence in the section 2. .
Source: OECD calculations based on NIH RePORTER data.

32. In both of the cases classified in the category D, AI keywords are only appeared in project terms and the project title and abstract do not mention topics related to AI nor support the allocation of the project terms. It happens because the project terms are automatically allocated with analysing text data including those out of the database. It shows the limitation of the keyword matching approach that it only well works when the documents mention AI-based methodologies or tasks.

33. For the cases classified in the category C, enough information is not available in the document to confirm AI-based nature. For example, in the document titled “The Future

[7] AI theories and systems refer to the ones in the definition of artificial intelligence in the section 2.

[8] Those theories or systems include “natural language processing (NLP) methods” and “machine learning and computer vision algorithms”.

[9] Those tasks include “gene ontology analysis”, “automated image analysis” and “to develop algorithms for discovering community structures in biological networks”.

of Genomics Medicine in Patient Care: Contributions from CHOP", the abstract contain an AI-keywords "electronic medical records" and the project terms contain "data mining". However, since there is no description on applying data mining to electronic medical records nor on what types of data mining are used, it is not possible to confirm it as AI-based.

4.4.2. False negatives

34. To assess the potential margin of error associated with using an incomplete list of keywords, a similar manual detection process is followed for documents not selected as AI related. 100 documents were selected at random from the 1,235,643 documents categorised as non AI-based health research. Among those 100, 1 document each were classified in the category A and B (AI-based health research). 26 were categorised in the category C and 72 are classified in the category D. If false negative rate is defined as the rate of category A and B, the 95% confidence interval of false negative rate is from 0.00 to 0.07; if it is defined as proportion in the category A to C, the interval is from 0.19 to 0.38.

35. The document categorised into A includes descriptions "neural network" and "training" used in the meanings related to AI. Since both terms are used in multiple meanings, it has not been included into the keyword list. The document categorised into B uses "classification" in the context that "the project develops signal-processing algorithms for 3-D detection, imaging, classification, and monitoring". As it mentions that the algorithm classifies 3-D images, the project may involve the use of AI theories or systems.

Table 4.4. Classification for identifying false negatives

Categories	Definitions	Number of documents in the categories
A	Mentions the use or develop of AI theories or systems.	1
B	Is not classified as A, but mentions that the project works on tasks normally achieved by using AI theories or systems.	1
C	It may be classified as A or B, but there is not enough information to either confirm or deny it.	**26**
D	Unlikely to be AI related.	72

Note: AI theories and systems refer to the ones in the definition of artificial intelligence in the section 2.
Source: OECD calculations based on NIH RePORTER data.

4.4.3. General assessment

36. After examining 200 documents for estimating false positive / negative rate, tere are some potential lessons for refining the keyword matching approach. Firstly, methodologies that utilise AI theories or systems (such as "machine learning", "natural language processing" and "deep learning") clearly show that the projects are AI-based. The accuracy of the keyword matching would be improved by expanding the set of such unambiguous keywords. Secondly, keywords related to health research tasks supported by AI (such as "bioinformatics", "gene ontology" and "gene annotations") are useful to detect AI-based health research with health specific descriptions (without keywords in AI theories and systems). Those keywords may contribute to reduce the number of documents classified in the category C. Finally, some keywords should more be penalised in the classification criterion as they are weak to represent AI-based research: "robotics, robotic and robot" are sometimes used simple automations that may not utilise AI; "electronic medical records" are not always subject to natural language processing.

37. The sensitivity test based on an assessment of random samples of the corpus reveals the inconclusiveness of a large segment of documents in both the group that is presumed by the text mining to be AI related and that presumed not to be AI related. This points to the fundamental challenge of data quality. At an aggregate level, the text mining approach delivers plausible results albeit with some potential upward bias after taking into consideration very naïve estimates of type 1 and type 2 errors.

5. Conclusions and next steps

38. This document has carried out a text-based analysis of project level funding data in the area of health R&D. The quantitative case study, undertaken on funding data from the US National Institutes of Health, one of the major R&D funding organisations in the world in this area, reveals for the potential for using project level data descriptions to carry out in depth analysis of the content and methods of R&D with such type of funding data.

39. The findings on the subject of interest, AI-based health R&D, appear to conclusively demonstrate an upward trend in AI-related research funding and a higher rate of adoption in domains that appear most likely to benefit from AI applications, namely in connection with the analysis and processing of large amounts of genetic data, enabling mobility through robotics and brain-machine communication devices. Working with R&D funding data, it is possible to put a monetary figure on AI estimates (USD 1.7bn) as opposed to just using counts of highly heterogeneous entities like documents or projects. The estimated percentage of AI-related R&D funding in 2017 is higher than by projects (5% of funding versus 3.5% of grant funded projects).

40. This case study brings challenges such as finding the right balance between mechanical procedures and personal judgements, the impact of large projects with high funding levels and multiple components, some of which AI related. Data-based classification decisions on topics like AI, but also any other form of enabling technology, mechanically or manually implemented, require ultimately a view of when such applications or procedures cease to be so as they become part of the “new normal”. To be operationalised in a given corpus, definitions of AI are necessarily context dependent.

41. The study reveals that unstructured funding microdata contains valuable information that can shed insights on the structure of R&D efforts by governments. It also reveals that such efforts are highly dependent on comprehensive data infrastructure and exhaustive description of the projects. Text mining of data containing superficial descriptions will fail to identify methodological features of the projects that can be highly relevant to some purposes. In the case of health R&D project abstracts, the information content about the application is likely to prevail unless access to more detailed methodological descriptions can be secured.

42. A number of possible next steps and extensions to the work may be envisaged, depending on priorities to be agreed by delegates. Within this particular case study, it may be possible to reducing the extent of manual intervention, and assessing the extent to which other methods can be applied, including comparing and linking to other corpora.

43. One potential development would be to investigate the possibility of using data from other US-based funding organisations that actively seek to advance the state of the art on AI as a means to enrich the vocabulary on AI related research and establish potential linkages. A starting point could be the NSF funding database[10].

44. There is a clear interest in evaluating the possible of this work more globally to examine AI trends. This raises issues of data availability and consistency. In the case of health R&D, different countries organise their funding of health R&D in different ways, and the micro, project-level data are not always openly available. The question for NESTI delegates is whether it would be possible to make project level funding data available to OECD, and whether OECD should study the possibility of working with commercial providers of such data. A questionnaire on the availability of data on funding at a project level for analytical purposes such as those described in this paper will be the subject of a forthcoming survey to countries, as agreed at the December 2017 meeting.

[10] See https://www.nsf.gov/awardsearch/download.jsp

References

BEIS and DCMS (2018), *Industrial Strategy Artificial Intelligence Sector Deal*. [9]

EC (2018), *Artificial Intelligence for Europe; Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions; COM(2018) 237 final*. [8]

Freyman, C., J. Byrnes and J. Alexander (2016), "Machine-learning-based classification of research grant award records", *Research Evaluation*, Vol. 25/4, p. rvw016, http://dx.doi.org/10.1093/reseval/rvw016. [6]

Kawamura, T. et al. (2018), "Funding map using paragraph embedding based on semantic diversity", *Scientometrics*, pp. 1-18, http://dx.doi.org/10.1007/s11192-018-2783-x. [5]

NITRD (2018), *The Networking and Information Technology Research and Development Program Supplement to the President's FY2018 Budget*, https://www.nitrd.gov/pubs/2018supplement/FY2018NITRDSupplement.pdf (accessed on 21 June 2018). [10]

NSTC (2016), "Preparing for the Future of Artificial Intelligence", https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (accessed on 21 June 2018). [7]

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264239012-en. [11]

Omar, M. et al. (2017), "Global mapping of artificial intelligence in Google and Google Scholar", *Scientometrics*, Vol. 113/3, pp. 1269-1305, http://dx.doi.org/10.1007/s11192-017-2534-4. [1]

Park, J. et al. (2016), *Analyzing NIH funding patterns over time with statistical text analysis*. [3]

Talley, E. et al. (2011), "Database of NIH grants using machine-learned categories and graphical clustering", *Nature Methods*, The article itself is too short. While a website introduced in it is worth looking at.
http://nihmaps.org/index.php, pp. 443-444, http://dx.doi.org/10.1038/nmeth.1619. [4]

van den Besselaar, P. and L. Leydesdorff (1996), "Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence", *Journal of the American Society for Information Science*, Vol. 47/6, pp. 415-436, http://dx.doi.org/10.1002/(SICI)1097-4571(199606)47:6<415::AID-ASI3>3.0.CO;2-Y. [2]

Annex. Background information

Annex.1. List of keywords identified

Table A.1. List of keywords identified after removing duplications

As Quasi-synonyms with cosine similarity higher than 0.6.

MeSH Terms	Quasi-Synonyms (Highly co-occurring terms)	Cosine similarity
machine learning	bayesian	0.757499
natural language processing	data mining	0.713131
machine learning	high dimensional	0.700497
natural language processing	big data	0.681247
machine learning	pattern recognition	0.678257
natural language processing	information retrieval	0.677777
machine learning	deep learning	0.674633
robotics	robot	0.669463
knowledge bases	ontologies	0.663008
machine learning	maximum entropy	0.661907
machine learning	variable selection	0.659154
natural language processing	information extraction	0.656357
machine learning	pattern classification	0.654707
machine learning	graph theory	0.646934
machine learning	multi dimensional	0.643441
machine learning	dimension reduction	0.640443
machine learning	computer vision	0.637981
knowledge bases	genomic datasets	0.636857
knowledge bases	genome annotations	0.634838
natural language processing	text mining	0.626507
gene ontology	hierarchical clustering	0.621874
natural language processing	electronic medical records	0.621627
natural language processing	decision support	0.619277
support vector machine	k nearest neighbor	0.619062
machine learning	radiomics	0.613235
machine learning	graphical models	0.612647
machine learning	feature selection	0.612587
support vector machine	recursive partitioning	0.61227
gene ontology	bioinformatic	0.611629
support vector machine	discriminant analysis	0.611443
robotics	data acquisition	0.605373
machine learning	classifier	0.604499
support vector machine	diffusional kurtosis imaging	0.601916

Note: Cosine similarity is an index that ranges between -1 (completely opposite) to +1(identical).
Source: OECD calculations based on NIH RePORTER data.

Annex.2. List of terms excluded from the keyword list

Table A.2. List of terms excluded from the keyword list

As terms with overly general meanings.

MeSH Terms	Quasi-synonyms	Cosine similarity
machine learning	statistical	0.701009095
machine learning	algorithms	0.693119526
machine learning	computational	0.684298635
robotics	hardware	0.684267342
machine learning	multivariate	0.68085587
machine learning	mathematical	0.665338814
machine learning	sophisticated statistical	0.664831877
knowledge bases	datasets	0.662302434
knowledge bases	data sets	0.660583138
natural language processing	informatics	0.659093618
robotics	software	0.652282357
robotics	computer	0.64769882
machine learning	algorithm	0.645343542
machine learning	probabilistic	0.636729896
machine learning	statistics	0.636088848
machine learning	analytic	0.629539728
robotics	high performance	0.628646731
robotics	workstation	0.627742946
knowledge bases	databases	0.622641981
machine learning	statistical modeling	0.622220159
gene ontology	annotation	0.617037952
machine learning	crowdsourcing	0.61396265
machine learning	statistical methods	0.613721073
gene ontology	annotations	0.612591505
robotics	electronics	0.612070322
machine learning	modeling	0.610540867
machine learning	heuristic	0.608321071
machine learning	computation	0.60525012
machine learning	mutual information	0.603126228
machine learning	semantic	0.602818727
robotics	hardware software	0.601902366
knowledge bases	free text	0.600069165

Note: Cosine similarity is an index that ranges between -1 (completely opposite) to +1(identical).
Source: OECD calculations based on NIH RePORTER data

Annex.3. Description of large project identified as potentially AI-related

Table A.3. Description of an exceptionally big project

Project title	National Biomedical Information Services
Abstract	Delivering Biomedical Information Services In FY 2009, NLM expanded the quantity and range of high quality Information available to researchers, health professionals, and the general public. Among the NLMs intramural programs that contribute to its National Biomedical Information Services are the following examples: PubMed/MEDLINE: PubMed, which incorporates MEDLINE, is NLMs premier bibliographic database with over 19 million references to Biomedical journal articles. MEDLINE articles are indexed by experts using the Medical Subject Headings (MeSH) controlled vocabulary, updated annually. In FY2009, more than 700,000 new indexed citations were added. PubMed Central: The PubMed Central archive of over 1.8 million full-text journal articles is central to the NIH effort to make accessible the published results of research it supports. In FY09, NLM enhanced the NIH Manuscript Submission system (NIHMS) by which authors can submit articles to PubMed Central in compliance with the NIH Public Access Policy. NLM has also made PMC software available to archiving organizations in the UK and Canada. MedlinePlus and MedlinePlus en español: MedlinePlus and the Spanish language MedlinePlus en español provide access to consumer health Information on more than 800 diseases and conditions. Recent enhancements included improved search capabilities, addition of summary Information, and expansion to include Information in more than 40 languages. Clinical Trials: ClinicalTrials.gov contains Information on more than 79,000 clinical research studies in more than 169 countries, with hundreds added each week. It was significantly expanded to respond to new clinical trial registration and results reporting requirements established by the FDA Amendments Act of 2007 (PL 110-85). In FY09, some 32,000 trials were registered, raising the total to 79,000. Since NLM implemented the results database required by the law in September 2008, summary results of more than 880 clinical trials have been added. In September 2009, the previously optional adverse events module of the results database became mandatory. Toxicology and Environmental Health: The Toxicology Data Network (TOXNET) is a primary reference for toxicologists, poison control centers, public health administrators, physicians and other environmental health professionals, and includes databases such as Hazardous Substances Data Bank, TOXLINE, GENE-TOX, and the Toxic Release Inventory. Influenza Virus Resource: The Influenza Virus Resource contains influenza virus sequences and enables researchers to compare different virus strains, identify genetic factors that determine the virulence of virus strains, and look for new therapeutic, diagnostic and vaccine targets. Updated daily, it includes over 90,000 influenza sequences and more than 2,000 complete genomes. Disaster Preparedness and Response: NLMs Disaster Information Management Research Center facilitates access to disaster Information, promotes effective use of libraries and disaster Information specialists in disaster management efforts, and supports initiatives to ensure uninterrupted access to critical health Information resources when disasters occur. A collaboration with the Bethesda Hospital Emergency Preparedness Partnership provides backup communication systems and develop tools for patient tracking, Information sharing and access, and responder training, and to serve as a model for hospitals across the nation. Molecular Biology, **Bioinformatics**, and Human Genome Resources: NCBI Information resources include molecular biology databases and **bioinformatics** software tools such as GenBank, Entrez, BLAST, RefSeq, UniGene, LocusLink, annotation and assemply of complete genomes, and the NCBI software toolkit. NCBI also produces the Information retrieval systems for the PubMed, PubMed Central, and the NCBI Bookshelf. Some areas of emphasis in FY09 augmenting the Short Read Archive of more than 400GB of raw sequence data derived from massively parallel sequencing technologies; augmenting dbGAP with more than 20 GWAS studies and developing a new access system using the high-efficiency FASP protocol; and continuing the discovery initiative to alleviate difficulties in finding relevant Information in diverse resources and develop interface improvements such as Releated Reviews and Patient Drug Information. Outreach: Promoting Public Awareness and Access Consumer health websites and the NIH MedlinePlus Magazine transmit the latest useful research findings in lay language. In FY09, NLM increased distribution of the magazine to 600,000, launched a Spanish language edition, Salud!, and introduced online versions the magazines. Special population websites address specific minority health concerns in various racial and ethnic groups. NLM outreach programs enhance awareness of its Information Services. Special attention is given to minority groups and other underserved populations, including African American, Hispanic, and Native American communities, as well as health professionals serving minority populations and practicing in rural and inner city communities. In FY 2009, dozens of community-based projects were funded across the country. Health Services Research NICHSR improves access to health Services research through Information systems such as: HSRProj, a database covering over 5000 ongoing or recently completed health Services research projects; HSRR (Health Services and Sciences Research Resources), a database of research datasets, instruments and software relevant to health Services research, behavioral and social sciences, and public health; and HSTAT (Health Services/Technology Assessment Text), a full-text database including evidence reports, guidelines technology assessments, consensus statements, and treatment protocols. Advanced Information Systems, Data Standards and Research Tools In FY 2009, LHC and NCBI continued to conduct research in Biomedical informatics and

	computational biology Information systems, tested the effectiveness of medical informatics interventions, and developed new scientific computing tools. To cite a few examples, intramural researchers developed tools that support standards-based personal health records; applied **natural language processing** methods to extract Information from Biomedical literature; improved automatic detection of gene and protein names in scientific text; and provided software tools that enabled rapid expansion of the PubChem database. NLM made substantial contributions toward standardized reporting of genetic variations and clinical interpretation of genetic test results by augmenting RefSeqGene and dbSNP; expanding the LOINC ; and launching the Newborn Screening Coding and Terminology Guide to enable more effective use of newborn screening test results. Health Data Standards: As the central coordinating body for clinical terminology standards within DHHS, NLM supports nationwide implementation of an interoperable health Information technology infrastructure. NLM develops or licenses key clinical terminologies and problem lists designated as standards for U.S. health Information exchange. The Unified Medical Language System Metathesaurus, with more than 7.7 million concept names from more than 100 vocabularies, is a distribution mechanism for standard code sets and vocabularies used in health Data Systems. NLM also produces RxNorm, a standard clinical drug vocabulary; supports the LOINC nomenclature for laboratory tests and patient observations; and promotes International adoption of the SNOMED CT clinical terminology. In FY09, NLM released the first version of the CORE Problem List Subset of SNOMED CT and launched the Newborn Screening Codes and Terminology Guide, which provides a standard framework for reporting the results of newborn screening tests.

Note: AI keywords are highlighted with bold case.
Source: NIH RePORTER.

Annex.4. Examples of documents classified for estimating false positive rate

Table A.4. An Example in the Category A

Project title	RUMI: A patient portal for retrieving understandable medical information
Abstract	DESCRIPTION (provided by applicant): Despite improvements in its diagnosis and treatment, lung cancer remains the leading cause of cancer-related deaths worldwide. The high degree of mortality associated with this cancer, and the spectrum of different treatment options (which themselves involve significant morbidity), creates a difficult situation in which a patient is faced with critical, if not life-changing decisons. In an effort to make informed decisions, many turn to the Internet to find information on their disease. However, the quality of information is variable; the information can be difficult for the layperson to comprehend; and patients can have considerable problems understanding how the information they find is specifically applicable to their individual circumstances. The objective of this proposal is the development of a framework, named RUMI (Retrieving Understandable Medical Information), which challenges how cancer patients receive information today by making the process of care explicit to the patient, providing access to his/her medical record data in the direct context of a clinical guideline so they can see how decisions are made. The first step in our effort is the development of a comprehensive lung cancer knowledge and process model (LCKPM) that captures a clinical guideline and workflow. The LCKPM provides a foundation for connecting the patient's medical record to decision points and actions that occur over time, detailing the criteria in making a selection and the supporting rationale. Different layers in the LCKPM organize information, which provide links: to public information resources helping explicate unfamiliar medical jargon and concepts; to questions (and answers) that are frequently fielded by healthcare providers managing lung cancer patients; and to more meaningful clinical episodes as experienced by the patient. Based on the LCKPM, RUMI automatically maps a patient's medical record: free-text documents are analyzed via **information extraction** (IE) and **natural language processing** (NLP) methods, identifying key concepts and variables (e.g., as used in decisions). In parallel, we examine public web resources that provide consumer-level explanations and discussions of lung cancer; these sites are deconstructed into curated knowledge segments that can be used in more directed presentations to a given individual. The knowledge segments are used to explain medical concepts within a patient's reports; and the terminology within the clinical guideline process flow. Collectively, these developments implement an individually-tailored web portal that visualizes the patient's personal experience over the disease trajectory using a simplified event-driven timeline. The portal also provides customized information regarding clinical trials that the individual is eligible for, and pertinent past trials results. In addition to technical evaluation, the RUMI framework will be evaluated through the UCLA Lung Cancer Program's outpatient clinics in a controlled study to assess end impact. The result of this project will be a set of approaches to employ patients' own medical records and public information resources to inform and empower lung cancer patients as participants in their own healthcare and medical decision-making processes.

Note: AI keywords are highlighted with bold case.
Source: NIH RePORTER.

Table A.5. An Example in the Category B

Project title	Molecular Roles of Cdk5 in Neuronal Functions and Pain Signaling
Abstract	Cdk5 and orofacial mechanical pain: Chronic orofacial pain is a significant public health concern. Patients with orofacial pain conditions often experience mechanical and thermal allodynia or hyperalgesia. Nevertheless, there are few animal models for such conditions. In order to study orofacial hypo/hyperalgesia and determine whether Cdk5 activity is involved, we have utilized special devices to quantify the responses of mice to painful mechanical orofacial stimulation. Using these devices, we are exploring the link between Cdk5 and orofacial pain as manifested by mechanical hypoalgesia or hyperalgesia. Is Cdk5 involved in orofacial mechanical pain sensation? To answer this fundamental question, we are using a modified orofacial stimulation test (OST) on wild-type, p35-/-, and Tgp35 mice to study the effect of different mechanical stimuli on their behavior. This new behavioral testing method uses a conflict paradigm that allows animals to make a choice between receiving a reward (30% sucrose) or escaping aversive stimuli, so the animals have control over the amount of nociceptive stimulation and can modify their own behavior. Additionally, this technique provides investigator-independent testing using automatically recorded behavior of the observed animals; they incur less stress, and their behavior can be measured repeatedly in a non-biased fashion. Three different levels of painful conditions were achieved by interfering with their access to a reward (30% sucrose) using plates with different numbers of Nitinol wires (pain level 1: 6+6 wires, level 2: 9+9 wires, and level 3: 13+13 wires). Our current findings reveal aversive behavior to mechanical stimulation with orofacial mechanical hypersensitivity in Tgp35 mice (which have increased Cdk5 activity), as evidenced by shortening of the total licking time and number of attempts the mice make to access the reward. The number of reward licking/facial contact events decreased substantially in these mice with increased mechanical pain intensity. In contrast, mice lacking p35 (with decreased Cdk5 activity) displayed mechanical hypoalgesia. To the best of our knowledge, we are the first to report using the orofacial mechanical stimulation test in mice to demonstrate that Cdk5 plays an important role in orofacial mechanonociception. Phosphoproteomic analysis for novel Cdk5 substrates: Because of the important role of Cdk5 in pain signaling, neurotransmitter release, and neurodegeneration, we examined whether it was practical to conduct phosphoproteomics screening, initially to compare phosphoproteins in Cdk5+/+ and Cdk5-/- mice. The p35-/- and p39-/- mice still have residual Cdk5 activity, so we used Cdk5-/- mice; because these mice show perinatal lethality, we used whole-brain protein extracts obtained from E18.5 Cdk5+/+ and Cdk5-/- mouse embryos. Phosphoproteomic analysis was performed in collaboration with Christian Gonzalez of the University of Chile, Santiago, Chile. Phosphoproteins were isolated from our whole-brain protein extracts and digested with trypsin, and the resulting peptides were isotopically labeled for their relative and absolute quantification. We were able to identify changes in the levels of 40 phosphoproteins containing one or more Cdk5 phosphorylation site(s). We classified these 40 phosphoproteins according to their functions using **gene ontology** software. With the **gene ontology** analysis, we were, for example, able to classify 11 of the phosphoproteins as involved in neuronal morphogenesis while 13 were grouped in the category of signaling pathways. In summary, having shown that Cdk5 modulates orofacial mechanical pain, our current research is focused on further confirming these findings with additional Cdk5 mouse models. Furthermore, we will vigorously pursue molecular investigations into identifying novel Cdk5 substrates involved in pain signaling.

Note: AI keywords are highlighted with bold case.
Source: NIH RePORTER.

Table A.6. An Example in the Category C

Project title	The Future of Genomics Medicine in Patient Care: Contributions from CHOP
Abstract	DESCRIPTION (provided by applicant): The Center for Applied Genomics (CAG) at The Children's Hospital of Philadelphia (CHOP) has established a pediatric biorepository with over 60,000 children consented for access to **electronic health records** (EHRs) with regular updates, re-contact, and high-density genome-wide association (GWA) array data. For this project, we have proposed five specific aims reflecting the workgroup mechanism of the eMERGE consortium. These will build upon substantive Preliminary Data derived during eMERGE II, where a series of GWA, sequencing, pharmacogenomics (PGx), EHR integration, and return-of-results (RoR) projects helped establish a platform for translational and eMERGE III efforts. In the first of these aims, we propose to continue to expand the eMERGE phenotype library and we propose four lead phenotypes: obesity, epilepsy, intellectual disability, and autism, in which we have a strong record in discovery, integration, and translation. Under specific aim 2, we propose to leverage CAG's position as a world-leader in genomics research to characterize rare variants in 2,000 CAG patients, where we have already catalogued several hundred rare variants in the accompanying Appendix. Leveraging this resource and expertise at CAG and eMERGE, we propose to return actionable findings to a minimum of ~160 parents (Specific Aim 3). This effort will build upon our existing RoR platform established during eMERGE II that returned results to parents of 160 CHOP children with autism, and to several hundred individuals with PGx risk profiles. Our fourth aim is to evaluate the health impact, cost-effectiveness and ELSI implications of RoR, and we aim to longitudinally track all families to whom results are returned at four time-points, leveraging existing resources and surveys developed with pediatric eMERGE partners. Finally, we propose a massive expansion of our EHR integration, established under eMERGE II, which also provide and integrate education resources for patients and medical professionals across the eMERGE network. Ultimately, we anticipate that the immediate outcome of these efforts will be improved healthcare for patients at CHOP and expanding to the entire eMERGE network. Further, with our eMERGE partners, we aim to establish a blueprint for integrating genomics and EHR data on such a scale that it will have real potential to fundamentally change medical practice in the US.

Note: Project terms of this document contain an AI-keyword “data mining”.
Source: NIH RePORTER.

Table A.7. An Example in the Category D

Project title	Establishment of an H3Africa Biorepository at Contract Laboratory Services
Abstract	DESCRIPTION (provided by applicant): This H3Africa biorepository project covers the extension and improvement of the biorepository capacity at the Contract Laboratory Services (CLS), which is subsidiary of the University of the Witwatersrand. Established in 2000, CLS is a focused organization providing diagnostic laboratory services, including sample storage facilities for clinical trials and research studies. It already manages a large repository (>700,000 samples) of human samples on behalf of a number of collaborating clinical trial networks including NIH funded CIPRA project Safeguard the household (2002-2008), Central HIV Aids Vaccine Immunology (CHAVI), the International HIV/AIDS Vaccine Initiative (IAVI), the Adult and Pediatric AIDS Clinical Trials Group (ACTG/PACTG/IMPAACT), Partners in Prevention (PIP and PreP), Microbicide Trial Network (MTN), and acts as the central repository for the South African Cancer Epidemiological Research Group. Due to our experience in clinical trials and laboratory support work in Africa, we value capacity building in this environment to address and support health needs as critical. In order to be a key stakeholder in the H3Africa mission, the project funding will be used to expand and optimize CLS' sample storage infrastructure in order to accommodate samples from H3Africa in a dedicated state of the art facility. The project will also assist with the implementation of several important CLS activities in order to better serve the needs of the H3Africa initiative, including the upgrading of its laboratory and sample management software, the use of radio frequency identification tags to reduce human error and sample degradation, and the further training of its staff in quality assurance, and repository management tools and standards. In the first phase of the project (UH2), CLS will undertake a more detailed assessment and feasibility analysis of trends in biorepository technology and the needs of H3Africa, particularly in respect of sample tracking, analysis, storage, retrieval and distribution. It will also use this period to build further links with all the H3Africa /genomic research centers in order to align CLS biorepository capacity building with the needs of our collaborators and prepare detailed specifications for the second phase (UH3) activities. In the second phase (UH3), CLS will activate a dedicated H3Africa biorepository, collecting, storing and distributing samples from and to all over Africa as specified in the scoping studies of UH2. It will also implement several state-of-the-art repository practices in order to ensure the highest care and management of H3Africa samples. In all stages specific care will be taken to maintain the high standards of sample protection, as required by CLS' clinical trial partners. We note that CLS already has the following achievements in this regard: ISO 15189 accreditation for all its laboratories (and excellent inspection reports) through the South African National Accreditation System (SANAS) since 1999 accreditation for the CLS laboratory through the British Quality Group DAIDS approved through PPD and FHI monitoring IATA and LDMS certification of its repository. CLS has a proven track history of managing and growing its biorepository over the last 10 years in a highly cost-efficient manner; this project will further support expansion and continuous improvement of these services.

Note: Project terms of this document contain an AI-keyword “bioinformatics” and “robotics”.
Source: NIH RePORTER.