 doi:10.20944/preprints201904.0170.v1

TITLE: Mapping research in assisted reproduction worldwide

RUNNING TITLE: Topic Modelling on ART research

AUTHORS: D. García[1], F. A. Massucci[2], A. Mosca[2], I. Ràfols[3], A. Rodríguez[1], R. Vassena[1,*]

ADDRESS:

[1] Clínica EUGIN, Barcelona 08029, Spain

[2] SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain

[3] Ingenio (CSIC-UPV), Universitat Politècnica de València, València, Spain & Centre for Science and Technology Studies (CWTS), University of Leiden, Netherlands

*Correspondence: Clínica EUGIN, Travessera de les Corts 322, 08029 Barcelona, Spain. Email: rvassena@eugin.es

 doi:10.20944/preprints201904.0170.v1

EXTENDED ABSTRACT:

Study question: What are the current trends of research in Human Assisted Reproduction around the world?

Summary answer: USA is the leading country, followed by the UK, China, France and Italy. The largest research area is “laboratory techniques”, although other areas such as “public health”, “quality, ethics and law” and “female factor” are gaining ground worldwide.

What is known already: Scientific research, especially in health and medical sciences, aims at addressing specific needs that society (and, especially, patients) perceives as pressing. One of the main challenges for policymakers and research funders alike is therefore to align research priorities to societal needs. We can thus think of research agendas in terms of a demand side (societal needs) and a supply side (research outputs). Research output in Human Assisted Reproduction has expanded in the past years, as indicated by the increasing number of scientific publications in indexed journals in this area. Nevertheless, no map of research related to assisted reproduction has been produced so far, hindering the identification of potential areas of improvement and need.

Study design, size, duration: 26,000+ scientific publications (articles, letters, and reviews) on Human Assisted Reproduction produced worldwide between 2005 and 2016 were analyzed. These publications were indexed in PubMed or obtained from reference list of indexed publications included in the analysis.

Participants/materials, setting, methods: The corpus of publications was obtained by combining the MeSH terms: “Reproductive techniques”, “Reproductive medicine”,

 doi:10.20944/preprints201904.0170.v1

"Reproductive health", "Fertility", "Infertility", and "Germ cells". Then it was analyzed by means of text mining algorithms (Topic Modeling (TM) based on Latent Dirichlet Allocation (LDA)), in order to obtain the main topics of interest. Finally, these categories were analyzed across world regions and time.

Main results and the role of chance: We identified 44 main topics, which were further grouped in 11 macro categories, form larger to smaller: "laboratory techniques", "male factor", "quality, ethics and law", "female factor", "public health and infectious diseases", "basic research and genetics", "pregnancy complications and risks", "general infertility and ART", "psychosocial aspects", "cancer", and "research methodology". The USA was the leading country in number of publications, followed by the UK, China, France and Italy. Interestingly, research contents in high income countries is fairly homogeneous across macro-categories, and it is dominated by "laboratory techniques" in Western and Southern Europe, and by "quality, ethics and law" in North America, Australia and New Zealand. In middle income countries we observe that research is mainly performed on "male factor", and noticeably less on "female factor". Finally, research on "public health and infectious diseases" predominates in low-income countries. Regarding temporal evolution of research, "laboratory techniques" is the most abundant topic on a yearly basis, and relatively constant over time. However, since production in most of the other categories is increasing, the relative contribution of this research category is actually decreasing. Publication is especially increasing in "public health and infectious diseases" (in all world regions, but especially in low income countries), "quality, ethics and law" (high income countries), and "female factor" (middle income countries).

Limitations, reasons for caution: Three main factors might limit the robustness of our work: the textual corpus analyzed is based on abstract and titles, the reproducibility of

 doi:10.20944/preprints201904.0170.v1

the stochastic algorithms applied, which may produce slightly differing results at each run, and the interpretation of the topics obtained.

Wider implications of the findings: This study should prove beneficial in the design of research strategies and policies that foster the alignment between supply (assisted reproduction research) and demand (society).

Study funding/competing interest(s): PTQ-14-06718 of the Spanish MINECO Torres Quevedo programme (FAM)

KEYWORDS: Topic Modelling; Latent Dirichlet Allocation; Text Mining; Assisted Reproduction; ART; IVF

 doi:10.20944/preprints201904.0170.v1

INTRODUCTION

One of the main challenges that research policymakers and funders are facing today is that of aligning research efforts with societal needs (Sarewitz 2007). This is particularly the case, in a context in which research is expected to help address major societal problems such as climate change and contribute to improve citizen wellbeing (Wallace 2015). Also, given that expenditure on research is stagnating in most western countries (OECD 2018), and the return on investment in health sciences is shrinking (Deloitte 2017), the need for accountability of research funding and research prioritization are increasingly becoming two pressing issues.

In order to align research with needs, two separate actions must take place; on the one hand, societal needs need to be identified, and patients' voices need to be included in a systematic and pervasive fashion in health related research funding decisions (Ràfols 2018). On the other hand, the landscape of the research that is currently been conducted must be described, so that an alignment of policies and funding can be achieved by comparing the supply (research conducted) with the demand (societal needs) (Cassi 2017; Ciarli 2019). For this reason, it is important to monitor the research landscape and to listen to current societal demands. These general considerations are especially relevant to the field of Human Assisted Reproduction research, where public funding is scarce, most research development follows local interests, and research priorities are usually not discussed at the national or international level.

Many techniques have been proposed to analyze, in a coarse grain fashion, the outputs of scientific production (Börner 2003; van Eck and Waltman 2007; Griffiths and Steyvers 2004). A fairly widespread approach is that of Topic Modeling (TM) (Hofmann 1999; Griffiths and Steyvers 2004; Steyvers and Griffiths 2007), a field of

 doi:10.20944/preprints201904.0170.v1

machine learning that, given a textual corpus, seeks to infer the set of topics that generated the observed word sequences.

In this work, we focus on the study of the current research supply in the field of Human Assisted Reproduction and the use of assisted reproductive technologies (ART). Specifically, we apply TM, implemented through Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) to infer the topics underlying the scientific publications produced by research on Human Assisted Reproduction in the time span 2005-2016.

MATERIALS AND METHODS

Corpus identification

Our work aims at detecting the main research topics in Human Assisted Reproduction tackled by scientific publications produced worldwide in recent years. The first step consists in the identification of the appropriate textual corpus to analyze. To properly limit the perimeter of our analysis, we resorted to the bibliographic database PubMed of the US National Library of Medicine and exploited the indexation of Medical Subject Headings (MeSH terms) to identify relevant publications. We searched scientific publications on PubMed via a query that combined via a logical “OR” the following MeSH terms: 1) Reproductive techniques, 2) Reproductive medicine, 3) Reproductive health, 4) Fertility, 5) Infertility, 6) Germ cells. To capture the core of the relevant literature, we also included in our query all publications which, albeit not indexed by the above MeSH terms, appeared in the following journals: Human Reproduction, Human Reproduction Update, Molecular Human Reproduction, Fertility and Sterility, Reproductive Biomedicine Online.

 doi:10.20944/preprints201904.0170.v1

On the other hand, we did not want to include publications studying assisted reproductive technologies on animals for farming purposes and reproductive health issues related to contraception. For this reason, we removed from our query, via a NOT logical cause the MeSH terms Contraception and Contraceptive behavior. We further removed the following journals: Theriogenology, Journal of Animal Sciences, Biology of Reproduction, Animal Reproduction Science. The combined query retrieved publications associated with human reproduction but not contraception or research on animals for farming or veterinary purposes. Because PubMed suffers a delay in indexing publications with MeSH terms, we limited the search to the year 2016. To have fair time coverage, we analyzed all scientific publications produced between 2005 and 2016.

We identified a set of 108,377 publications produced between 2005 and 2016, all tagged with their respective PubMed identifier and all their associated MeSH terms (for each publication, the associated MeSH featured those encompassed in our initial query and a further characterizing set).

We considered in our analysis only documents of type *Article*, *Letter* and *Review*. To filter out those types of records we retrieved all identified publications in the bibliometric database Scopus, via their PubMed identifier and retained their metadata, which included affiliations, keywords and references.

We next enriched the corpus by including potential relevant publications that were not indexed in PubMed, by adopting the following strategy: first, we included all references of type *Article*, *Letter* and *Review* cited by more than 50 records of the initial corpus and published after 2003. Then, we first extracted groups of 20 common keywords across the original corpus via Matrix Factorization (Lin 2007) and then we searched in Scopus

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

for records published after 2003, which featured combinations of 4 of those keywords plus the term “assisted reproduction”. This further effort allowed us to collect a total of 106,101 publications potentially dealing with the research themes of interest for our work. Notably, the collected documents could still feature assisted reproduction research not focused on Humans. Therefore, to filter unwanted results (e.g. cattle reproduction for farming purposes), we proceeded to clean the corpus by computing statistics of records for journals and excluding those journals that were not expected to be publishing research within our scope of analysis; and by extracting clusters of common MeSH terms associated to the collected publications via Matrix Factorization and removing records associated to groups of MeSH outside the perimeter of our analysis.

Finally, by considering only those records for which an abstract was available, we constructed a final corpus of 22,133 textual documents which were then studied by means of text mining algorithms.

Topic Modelling (TM)

Before applying any text mining algorithm, all texts were preprocessed (Bird, Klein, and Loper 2009) by harmonizing spelling (all words were converted into British spelling), identifying as many synonyms as possible and converting all variants onto the same canonical form (for instance, by converting all instances of “sperm index” and its variations found the texts into “semen index”) and by applying standard techniques of textual stemming (i.e. removing plurals, putting verbs into infinitive tense etc) and textual chunking (i.e. identifying noun phrases such as “human assisted reproduction”, which are recognized as a single word).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Then we applied TM via LDA (Blei, Ng, and Jordan 2003) to the texts of the abstracts of all the 22,133 identified publications. In a nutshell, LDA assumes that the observed distribution of words in a textual corpus is determined by a statistical model that fixes both a word-topic and a document-topic association. Therefore, within LDA, one tries to infer which are the appropriate (unknown) documents/topics and word/ topics associations. This inference is performed by maximizing the likelihood, i.e. the probability that the observed corpus is indeed generated by the theoretical model. This maximization may be performed in several different ways: here, we adopt Gibbs Sampling (Bishop 2006), carried out via the Mallet software (Mccallum 2002)

The results of the LDA algorithm consist in a list of topics, and in the weighted relations between each document in the corpus and each topic (weights are normalized to 1). In practice, every topic is a list of characterizing words, while each document may be connected with non-negligible weight to more than one single topic.

In LDA, the number of topics is left as a free parameter: there are several possible ways to fix this number, but no real established standard (Arun et al. 2010). In this work, we selected the number of topics by first evaluating the maximum log-likelihood as a function of the number of topics and then we chose the number of topics for which the log-likelihood reached a plateau. This procedure allowed us to select 44 topics.

Topics' graphical representation

All the topics are further represented in map, arranged by the software ldavis (Sievert and Shirley 2014) in a 2-dimensional plane via multidimensional scaling (Borg and Groenen 2006). In this map, each circle represents a topic, and closer topics share a larger vocabulary (i.e. they are more similar). The area of the circle is proportional to the frequency of the topic in the corpus, while the color denotes which macro-category

 doi:10.20944/preprints201904.0170.v1

the topic belongs to. We grouped the 44 topics into 11 macro via a qualitative inspection of the conceptual content of each topic. This grouping is consistent with the spatial arrangement provided in the map: topics belonging to the same macro-category are located in neighboring regions.

Since we know the affiliation of all authors of the collected publications, the topics of each document can be transitively associated to the respective affiliations (and thus the relative countries). This linkage allows one to derive topic and topic groups distributions at the geographical level and to quickly evaluate the research portfolio on ART of different countries, or of aggregation of them. Here, we perform two analyses of this kind, looking at distributions of categories by geographical regions (by using the regions defined by the United Nations Statistics Division (Undata 2017), and by income levels respectively (using the income groupings defined by the World Bank (WorldBankData2017).

RESULTS

Human Assisted Reproduction research output per country and per year

The number of publications finally considered in our study, sorted per country of production, is reported in Table I and shown on a map in Figure 1. Not surprisingly, most of the records retrieved are produced in the United States, followed by the United Kingdom and China. France and Italy complete the top 5 of the most active countries in research on assisted reproduction in the time period analyzed.

In Figure 2, we show the trend of the number of documents retrieved per year between 2005 and 2016, and the variation with respect to the production in 2005. We observe a growth of scientific production of about 120% between 2005 and 2015. As a reference, we also show the number of documents returned by PubMed by simply searching for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

the MeSH term “IVF” and the combination of MeSH terms “IVF” and “Humans”. As one can see, we are actually covering a much wider perimeter of analysis (the number of documents per year of our dataset in Figure 2a is much bigger than those baselines), but we are analyzing a production whose growth rate is very similar to that of research on IVF only, thus suggesting we are not enlarging our perimeter in an inconsistent manner.

Human Assisted Reproduction research production by topic

Using TM, we extracted 44 topics linked with research on Human Assisted Reproduction between 2005 and 2016. The list of all topics and the 10 terms included in each is provided in Table II. In this table topics are sorted in decreasing size: the first topic is the most frequent, while the last one is the smallest in the corpus.

The set of words of a topic allows to grasp what the topic is about and to understand whether it is related with clinical, basic or social research. For instance, the first topic deals with general themes regarding ART, the second topic with ovarian stimulation and the third with pregnancy risks and complications. Notably, we were able to extract topics specifically related with either the female side of assisted reproduction research (e.g. topic 5, 28 and 29) or the male one (e.g. topic 6, 12 and 25). This result allows us to quantify the amount of research efforts devoted to either male or female related research and to further drill down into the specific themes. Lastly, we found a number of terms related to cancer (topics 19 and 21), which seem to be entirely devoted to study fertility preservation and reproductive complications of surgical procedures in cancer treatments.

To get a fairly general view of research themes without getting lost into the details of each of the 44 topics, we mapped each of these topics into 11 macro-categories (topic groups), related with broader areas of research. These groups are shown in the second

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

column of Table II, and each deal with basic research (e.g. topic group *basic/genetics*), clinical (e.g. topic groups *pregnancy complications and risks* and *laboratory techniques*) and psychosocial (e.g. topic groups *psychosocial aspects* and *public health and infectious diseases*) aspects of research on assisted reproduction.

Regions of topics

In Figure 3 we show the map of all the topics represented (an interactive version of this figure is available online at http://sirislab.com/lab/EUGIN/ARTtopicLDA.html). We can identify the upper part of the graph as the region of *Laboratory techniques*, the lower left part mostly as the region focused on *Basic research* and the lower right as the one concerned with the *Psychosocial aspects* of assisted reproduction research. *Female factor* research is almost homogeneously distributed along the x axis. Interestingly, a few topics dealing with *Cancer* and *Pregnancy complications* lie in between the *Female factor* and the *Psychosocial aspects and policies* region. In a similar fashion, *Male factor* research seems to be an intermediary between *Female factor* and *Basic research*.

World Human Assisted Reproduction research production per category

The effort of grouping topics into macro-categories allows us to gain an effective coarse-grained view of worldwide scientific production on Human Assisted Reproduction research. In particular, we are able to assess which size and growth or decline trends of the categories. In Figure 4, we show the percentage of publications linked with each of the macro-categories. As it can be observed, most of the publications produced in the 2005-2016 period focus on the different techniques applied in assisted reproduction (i.e. IVF, ICSI, genetic tests as PGD or PGS, etc.), grouped in the macro-category *Laboratory techniques*. Notably, *Male factor* is the second largest macro-category in percentage of publications. Interestingly, the third category is *Quality*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

ethics and law. This result indicates that ART is progressively including research on policies and on ethical implications. Conversely, the category *Psychological aspects* is the third smallest, suggesting there could be room for growth for research dealing with psychological aspects of assisted reproduction.

Human Assisted Reproduction research production per geographical region

Research in the western world regions (Western Europe, Northern America and Australia & New Zealand) is fairly homogeneously distributed across macro-categories and relatively similar among different regions, with some minor differences (Figure 5). For instance, we can detect a relative skew in research on *Laboratory techniques* in Western Europe and on *Quality, ethics and law* in Northern America, respectively. On the other side of the spectrum, Sub-Saharan countries devote most of their research efforts to *Public health and infectious diseases* issues. It is also apparent how most of the countries outside the western world are more skewed in their Human Assisted Reproduction research towards *Male factor* rather than *Female factor*.

Similar findings are found when grouping countries by income (Figure 6). High income countries have a fairly balanced research production, with a relative abundance of research on *Laboratory techniques* and *Quality, ethics and law* (the latter, especially driven by research in the United States). The relative importance of research on *Public health and infectious diseases* seems to grow when the level of income decreases. Also, research on *Male factor* has a larger weight in middle income countries with respect to high income countries. Similarly, and not surprisingly, the amount of research on *Laboratory techniques* seems to increase with income.

 doi:10.20944/preprints201904.0170.v1

Temporal trends in Human Assisted Reproduction research

The analysis of temporal evolution of research output for each specific category can help to understand whether there are hot (fashionable and growing) or cold themes in Human Assisted Reproduction research. We show in Figure 7a the temporal trends of each macro-category of research and in Figure 7b the evolution of the percentual weight of each of these categories. Overall, the patterns are quite stable. *Laboratory techniques* is relatively constant and the most abundant on a yearly basis (Figure 7a). However, since production in most of the other categories is increasing, the relative contribution of this research category is actually decreasing a little bit (Figure 7b). Research on *Quality, ethics and law* shows a slight increase, in line with the general trends, so that its weight remains roughly constant. Research on *Public health and infectious diseases* shows a rapid increase and gains a conspicuous percentage of the total production already in 2011. Research on the *Female factor* shows a similar trend, slowly becoming a major research theme.

To gain further insights, the analysis can be again broken down at the level of country income groups: the results are shown in Figure 8. One can see that, for high income countries, research on *Quality, ethics and law* has caught up research on *Laboratory techniques*; research on the *Male factor* is slowly decreasing, while research on *Public health and infectious diseases* is increasing. Middle income countries have been slowly increasing their research on *Laboratory techniques*, catching up with the pattern shown by high income countries. Lower middle income countries experienced a peak of research on the *Male factor* in 2011 and then a decrease and share, with low income countries, a very marked growth of research on *Public health and infectious diseases*.

 doi:10.20944/preprints201904.0170.v1

DISCUSSION

In this article we identify for the first time the current hot topics on Human Assisted Reproduction worldwide and the temporal trends for the period 2005-2016. We provide an innovative picture of the current research in ART. This could help to explore the areas where further research is needed, in combination to other studies focused on patients' needs. The societal demand side of Human Assisted Reproduction research is not analyzed in the present article. The analysis of societal needs requires the involvement of societal stakeholders, which can be supported by extensive fieldwork, carefully designed surveys and the compiling of statistical data (Petit-Zeman, Firkins, and Scadding 2010; Ràfols 2018; Ciarli 2019). Human Assisted Reproduction includes infertility treatments for male, female or unknown factors, and also treatments for people who resort to ART due to a lack of the male and/or the female gamete. It also includes gametes vitrification for medical and social reasons. Consequently, we preferred to investigate around Human Assisted Reproduction and not limit our scope to infertility or IVF, although they are the central issues of our work.

Infertility is a disease (Zegers-Hochschild et al. 2009) that affects couples and individuals globally. However, in low income countries access to IVF (as well as other ART) is limited, even non-existent (Inhorn and Patrizio 2015). This may explain the low amount of research on ART performed in these countries; according to our results, research in these countries is smaller and more focused on *Public Health* and *Infectious Diseases*.

Globally, the four main areas of research are *Laboratory techniques*, *Male factor*, *Quality of ART, ethics and law*, and *Female factor*, which are chiefly developed by

 doi:10.20944/preprints201904.0170.v1

middle-to-high income countries. Research on infertility related to *Laboratory techniques* is more prevalent in high income countries, a feature that could be related to the higher complexity of research on these topics and the larger associated costs. Relative research efforts on *Male factor* compared to *Female factor* are not related to income group, but we observe clear differences among regions. *Male factor* predominates in all regions of Asia, Latin America, Northern Africa and Southern Europe, while *Female factor* heads research only in Eastern Europe. In Western Europe, Northern America, Australia and Sub-Saharan Africa the ratio between male and female factor is nearly 1. This suggests that the decision to carry out ART research on a certain topic might, in part, shaped by social and cultural aspects. One aspect to be taken into account since it could have biased these findings is that the definition of infertility as of male or female origin is not always consistent across the countries, and the prevalence of male infertility in the general population (whether in a relationship or not) has not been established yet (Barratt et al. 2017).

Regarding temporal trends, the total scientific production worldwide is increasing yearly, especially in the areas of *Public Health and infectious diseases* and *Female factor*. As a consequence, the dominant area of *Laboratory techniques* is losing relative prevalence at international level.

A limitation of this study is that it focuses only in the number of publications. However, it is important to note that increased quality in studies also leads to improved quality of care (Duffy et al. 2017). However, assessing how to improve quality of care is no easy task – for example, several treatments/techniques seem to increase pregnancy outcomes only slightly, and but even these small effects need large (and expensive) trials to achieve statistical significance (Evers 2013).

 doi:10.20944/preprints201904.0170.v1

In conclusion, this work may help open the debate on patient needs and research focus in the realm of Human Assisted Reproduction. Since this is the first and so far the only report in the field, it will permit understanding of current ART research, and discussing future research directions.

AUTHORS ROLES

FAM, AM, IR: involved in study design, data analysis, and manuscript preparation. DG, AR, RV: involved in study design, implementation and supervision, expert knowledge and manuscript preparation.

ACKNOWLEDGMENTS

We acknowledge the efforts performed by researchers in Human Assisted Reproduction for their contribution to this work.

FUNDING

FAM acknowledges financial support from the grant PTQ-14-06718 of the Spanish MINECO Torres Quevedo programme.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

 doi:10.20944/preprints201904.0170.v1

REFERENCES

Arun R, Suresh V, Veni Madhavan CE, and Narasimha Murthy MN. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Zaki, MJ, Yu, JX, Ravindran, B, and Pudi, V (eds) Advances in Knowledge Discovery and Data Mining. 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 391--402.

Barratt CLR, Bjorndahl L, De Jonge CJ, Lamb DJ, Osorio Martini F, McLachlan R, Oates RD, van der Poel S, St John B, Sigman M, *et al.* The diagnosis of male infertility: an analysis of the evidence to support the development of global WHO guidance-challenges and future research opportunities. *Hum Reprod Update* 2017: **23**; 660-680.

Bird S, Klein E, and Loper E. Natural language processing with Python: analyzing text with the natural language toolkit 2009. " O'Reilly Media, Inc.".

Bishop CM. Pattern recognition and machine learning. *springer* 2006.

Blei DM, Ng AY, and Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research* 2003: **3**; 993--1022.

Borg I and Groenen P. Modern Multidimensional Scaling: Theory and Applications. *Journal of Educational Measurement* 2006: **40**; 277-280.

Börner K, Chen, C., & Boyack, K. W. Visualizing knowledge domains. Annual review of information science and technology, 37(1), 179-255. 2003.

Cassi L, Lahatte, A., Rafols, I., Sautier, P., & De Turckheim, E. . Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics* 2017: **11**; 1095-1113.

Ciarli TaR, I. The Relation between Research Priorities and Societal Demands: The Case of Rice. Research Policy. 48 (4), 949-967. 2019.

Deloitte. A new future for R and D? Measuring the return from pharmaceutical innovation 2017. 2017. Deloitte.

Duffy J, Bhattacharya S, Herman M, Mol B, Vail A, Wilkinson J, Farquhar C, Cochrane G, and Fertility G. Reducing research waste in benign gynaecology and fertility research. *BJOG* 2017: **124**; 366-369.

Evers JL. The wobbly evidence base of reproductive medicine. *Reprod Biomed Online* 2013: **27**; 742-746.

Griffiths TL and Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences* 2004: **101**; 5228-5235.

Hofmann T. Probabilistic latent semantic analysis. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. 1999. pp 289--296.

Inhorn MC and Patrizio P. Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Hum Reprod Update* 2015: **21**; 411-426.

Lin C-J. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation* 2007: **19**; 2756--2779.

Mccallum AK. MALLET: A Machine Learning for Language Toolkit. 2002.

OECD. Gross domestic spending on R\&D (indicator). 2018.

Petit-Zeman S, Firkins L, and Scadding JW. The James Lind Alliance: tackling research mismatches. *Lancet* 2010: **376**; 667-669.

Ràfols I, and Yegros, A. Is research responding to health needs? Social Observatory of La Caixa Foundation (refereed online publication), March 2018. 2018.

Sarewitz D, & Pielke Jr, R. A. The neglected heart of science policy: reconciling supply of and demand for science. environmental science & policy, 10(1), 5-16. 2007.

Sievert C and Shirley K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014. pp 63--70.

Steyvers M and Griffiths T. Probabilistic topic models. *Handbook of latent semantic analysis* 2007: **427**; 424--440.

Undata. United Nations Statistics Division https://unstats.un.org/unsd/methodology/m49/. 2017.

van Eck NJ and Waltman L. VOS: A New Method for Visualizing Similarities Between Objects. In Decker, R and Lenz, H-J (eds) Advances in Data Analysis: Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8--10, 2006. 2007. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 299--306.

Wallace ML, & Rafols, I. Research portfolio analysis in science policy: moving from financial returns to societal benefits. Minerva, 53(2), 89-115. 2015.

WorldBankData2017. World Bank Country and Lending Groups 2018.

Zegers-Hochschild F, Adamson GD, de Mouzon J, Ishihara O, Mansour R, Nygren K, Sullivan E, van der Poel S, International Committee for Monitoring Assisted Reproductive T, and World Health O. The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) Revised Glossary on ART Terminology, 2009. *Hum Reprod* 2009: **24**; 2683-2687.

 doi:10.20944/preprints201904.0170.v1

FIGURE LEGENDS

Figure 1: World distribution of publications on Human Assisted Reproduction research. The US, UK, China, France and Italy dominate the ranking.

Figure 2: Comparison of temporal trends between the dataset analyzed in the present paper (Final Dataset) and documents yielded by PubMed by searching by MesH term "IVF" and the combination of the MeSH terms "IVF" and "Humans". Documents per year (**2a**) and growth rate (**2b**). We are collecting publications well beyond research on IVF, but in a research field which features a similar growth rate to that of research on IVF.

Figure 3: A map of the 44 topics extracted via LDA from the corpus analyzed in the present work. Coordinates are calculated via a dimensional reduction algorithm: each circle is a topic, closer circles denote similar topics. The size of the circles is proportional to the weight of the topic in the whole corpus, while colors denote the macro-categories to which each topic was manually mapped to.

Figure 4: Percentage of publications for each of the macro-categories defined in the present work. Most of the publications collected deal with assisted reproduction techniques.

Figure 5: Percentage of publications of each macro-category at geographical grouping level. Each group is defined by the United Nations, publications are associated to all affiliations.

Figure 6: Percentage of publications of each macro-category for countries grouped for income level. Each group is defined by the World Bank, publications are associated to all affiliations.

 doi:10.20944/preprints201904.0170.v1

Figure 7: Total (7a) and percent (7b) temporal trends of each research macro-category analyzed in the present paper. Some categories attain a constant evolution (e.g. techniques) and some categories grow. Since the total number of publications increases, categories with constant trends actually experience a decline of their percent contribution to the total production, while those categories growing like the total production keep a constant percent weight.

Figure 8: Temporal trends of each research macro-category analyzed in the present paper, broken down at country groups for income level. Each group is defined by the World Bank, publications are associated to all affiliations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Table I

Country	N. records
United States	6083
United Kingdom	2347
China	1630
France	1368
Italy	1285
Canada	994
Germany	929
Australia	907
Netherlands	881
Spain	746
Belgium	626
Japan	565
Turkey	562
Brazil	488
Israel	474
Sweden	468
India	461

 doi:10.20944/preprints201904.0170.v1

Denmark	444
Switzerland	384
Greece	357

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Table II

TOPIC	TOPIC GROUP	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10
1	GENERAL INFERTILITY & ART	infertility	assisted reproductive technology	treatment	couple	pregnancy	risk	technique	method	management	fertility
2	LABORATORY TECHNIQUES	day	hcg	ovarian stimulation	oocyte	cycle	controlled ovarian hyperstimulation	stimulation	gonadotropin releasing hormone agonist	follicle	oocyte retrieval
3	PREGNANCY COMPLICATIONS AND RISKS	assisted reproductive technology	child	pregnancy	risk	birth	twin	singleton	birth weight	increase risk	intracitoplasmatic sperm injection
4	LABORATORY TECHNIQUES	embryo transfer	day	embryo	cycle	transfer	live birth rate	live birth	pregnancy	clinical pregnancy rate	pregnancy rate
5	FEMALE FACTOR	age	infertility	pregnancy	risk	risk factor	time	body mass index	exposure	men	year
6	MALE FACTOR	men	semen quality	varicocele	semen concentration	semen analysis	seminal parameter	male infertility	infertile men	semen motility	semen count
7	PUBLIC HEALTH AND INFECTIOUS DISEASES	adolescent	sexual health	men	reproductive health	young people	sex	sexually transmitted infection	sexuality	hiv	intervention
8	QUALITY, ETHICS	embryo	law	assisted reproductive	child	decision	regulation	couple	debate	embryo	donation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

	AND LAW			technology						donation	
9	PSYCHOSOCIAL ASPECTS	infertility	couple	men	treatment	depression	infertile woman	anxiety	partner	infertile couple	quality
10	QUALITY, ETHICS AND LAW	care	guideline	quality	clinic	health care	doctor	nurse	recommendation	health professional	health care provider
11	QUALITY, ETHICS AND LAW	physician	respondent	gynecologist	survey	questionnaire	attitude	obstetrician gynecologists	abortion	recommendation	response rate
12	MALE FACTOR	dna fragmentation	spermatozoon	dna damage	semen sample	dfi	dna integrity	tunel	fragmentation	integrity	men
13	LABORATORY TECHNIQUES	oocyte	embryo	day	vitrification	blastocyst	culture medium	human oocyte	cryopreservation	fertilization	ivm
14	BASIC RESEARCH AND GENETICS	spermatozoon	protein	human semen	human spermatozoon	acrosome reaction	capacitation	expression	semen motility	progesterone	cell
15	BASIC RESEARCH AND GENETICS	male infertility	gene	deletion	polymorphism	frequency	mutation	azoospermia	infertile men	microdeletions	snp
16	METHODOLOGY RESEARCH	meta analysis	live birth	systematic review	treatment	live birth rate	controlled trial	quality evidence	clinical pregnancy	intervention	bias
17	PUBLIC HEALTH AND INFECTIOUS	reproductive health	health	policy	health service	health care	intervention	family planning	community	cost	government

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

	DISEASES										
18	PREGNANCY COMPLICATIONS AND RISKS	pregnancy	case report	ectopic pregnancy	gestation	risk	hyperstimulation syndrome	diagnosis	syndrome	embryo transfer	treatment
19	CANCER	cancer	fertility preservation	fertility	treatment	cancer patient	chemotherapy	cancer treatment	diagnosis	breast cancer	risk
20	PSYCHOSOCIAL ASPECTS	child	donor	parent	family	oocyte donation	mother	recipient	oocyte donor	semen donor	couple
21	CANCER	endometriosis	surgery	disease	treatment	recurrence	hysterectomy	laparoscopy	cervical cancer	diagnosis	ovarian cancer
22	QUALITY, ETHICS AND LAW	resident	training	medical student	performance	trainee	obstetrics	score	skill	obstetrics and gynecology	time
23	QUALITY, ETHICS AND LAW	obstetrics and gynecology	resident	obstetrics	training	survey	trainee	residency	respondent	specialty	physician
24	BASIC RESEARCH AND GENETICS	chromosome	frequency	aneuploidy	chromosomal abnormality	fluorescence in situ hybridization	carrier	translocation carrier	translocation	spermatozoon	karyotype
25	MALE FACTOR	men	azoospermia	semen extraction	obstructive azoospermia	fsh	spermatozoon	semen analysis	testis	non obstructive azoospermia	male infertility

26	MALE FACTOR	oxidative stress	spermatozoon	oxygen specie	semen motility	human spermatozoon	concentration	lipid peroxidation	apoptosis	semen function	motility
27	LABORATORY TECHNIQUES	intracitoplasmatic sperm injection	semen injection	oocyte	fertilization	fertilization rate	spermatozoon	couple	embryo quality	imsi	fertilization failure
28	FEMALE FACTOR	follicular fluid	serum	concentration	endometriosis	granulosa cell	antibody	infertile woman	asa	antisperm antibody	elisa
29	FEMALE FACTOR	hsg	sensitivity	specificity	hysteroscopy	procedure	laparoscopy	tubal patency	hysterosalpingography	fallopian tube	uterine cavity
30	FEMALE FACTOR	expression	endometrium	implantation failure	implantation	cell	infertile woman	unexplained infertility	endometriosis	endometrial receptivity	endometrial biopsy
31	LABORATORY TECHNIQUES	intrauterine insemination	couple	unexplained infertility	treatment	pregnancy	clomiphene citrate	pregnancy rate	cycle	infertility	ovulation
32	LABORATORY TECHNIQUES	preimplantation genetic diagnosis	embryo	preimplantation genetic screening	aneuploidy	couple	day	biopsy	transfer	chromosome	advanced maternal age
33	MALE FACTOR	expression	testis	germ cell	spermatogenesis	sertoli cell	cryptorchidism	cell	mrna	spermatogonia	leydig cell
34	PUBLIC HEALTH AND INFECTIOUS DISEASES	hiv	c. trachomatis	infection	c. trachomatis infection	infertility	risk	pcr	detection	ureaplasma urealyticum	transmission

35	FEMALE FACTOR	infertility	adenomyosis	uterus	fertility	myomectomy	fibroid	treatment	diagnosis	hysteroscopy	endometrial polyp
36	BASIC RESEARCH AND GENETICS	dna methylation	gene	expression	methylation	rna	telomere length	methylation pattern	igf	gene expression	dna
37	FEMALE FACTOR	amh	ovarian reserve	antral follicle count	age	fsh	follicle count	anti mullerian hormone	dor	amh level	inhibin b
38	PREGNANCY COMPLICATIONS AND RISKS	tocolysis	delivery	pregnancy	preterm labor	preterm birth	gestational age	pregnant woman	caesarean section	gestation	birth
39	PUBLIC HEALTH AND INFECTIOUS DISEASES	method	mtdna	spermatozoon	dna	mitochondrial dna	detection	semen cell	cell	propofol	victim
40	BASIC RESEARCH AND GENETICS	stem cell	cell	human	oocyte	cumulus cell	cell type	mitochondrion	scnt	somatic cell	granulosa cell
41	MALE FACTOR	device	method	exposure	laboratory	casa	worker	variation	vcl	measurement	lead
42	MALE FACTOR	treatment	acupuncture	treatment group	andrology	sci	significant difference	cord injury	medication	pregnancy rate	observation group
43	PUBLIC HEALTH AND INFECTIOUS DISEASES	hpv	vaccine	vaccination	human papilloma virus	hpv vaccine	hpv infection	hpv vaccination	hybrid	cervical cancer	growth

 doi:10.20944/preprints201904.0170.v1

44	PUBLIC HEALTH AND INFECTIOUS DISEASES	celiac disease	hepatitis b virus	hepatitis b virus infection	medicine	gynecology and obstetrics	hepatitis b virus dna	activity	site	th century	film

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Figure I

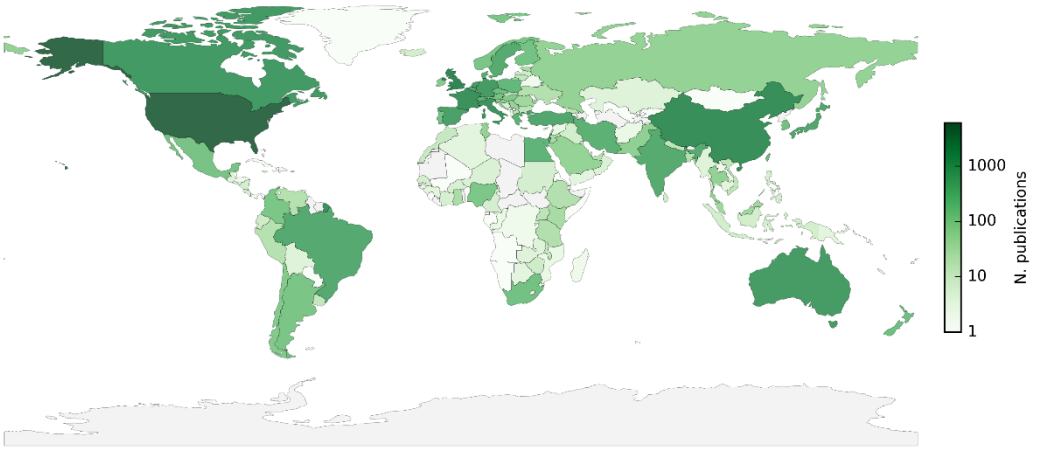
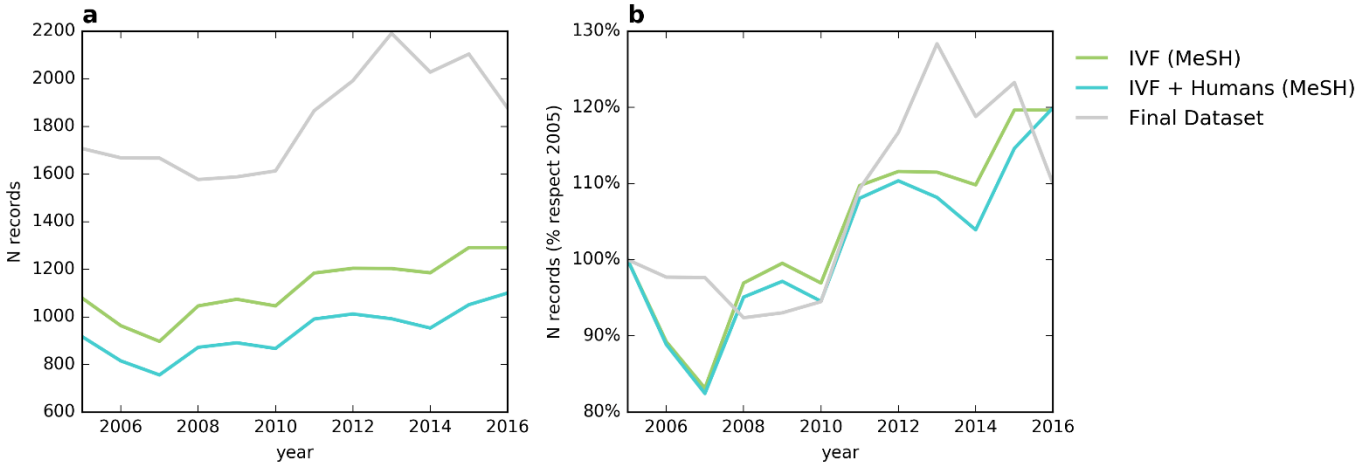


Figure II



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Figure III

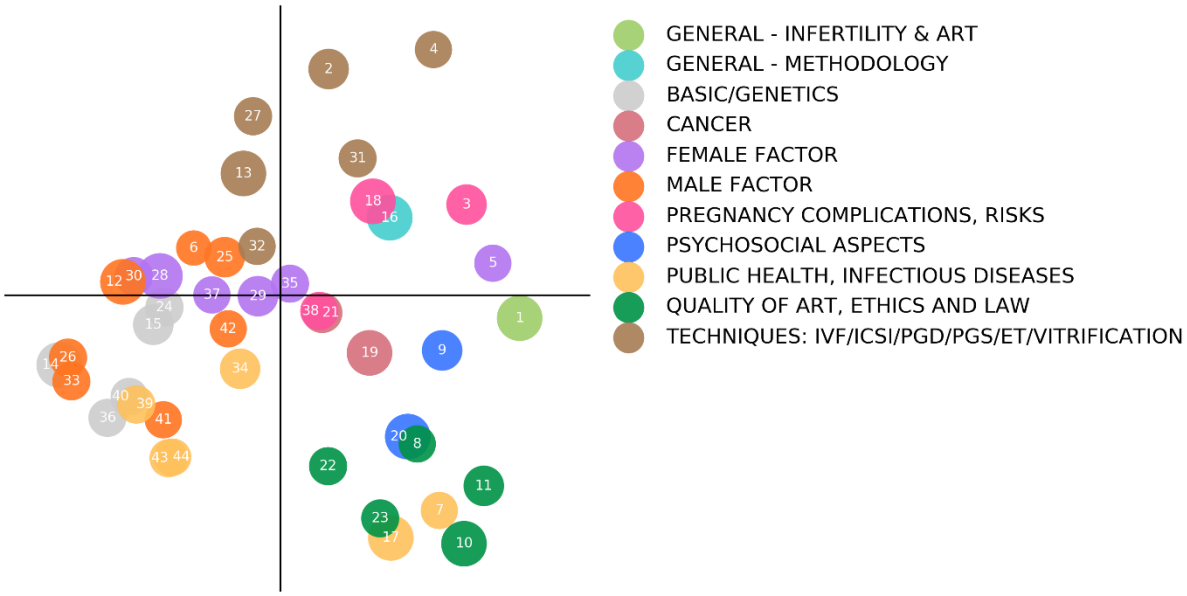
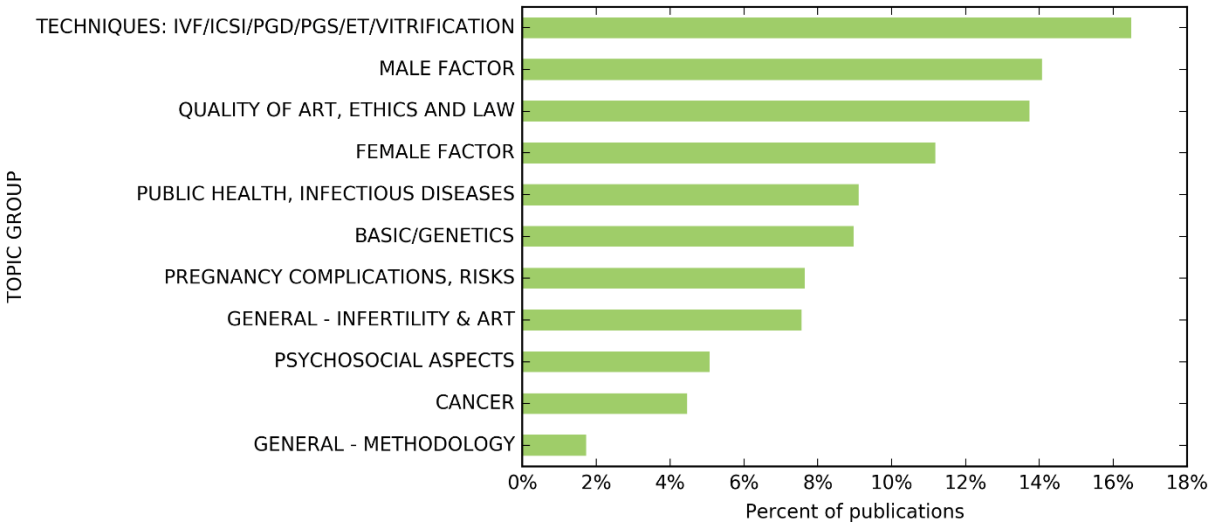


Figure IV



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Figure V

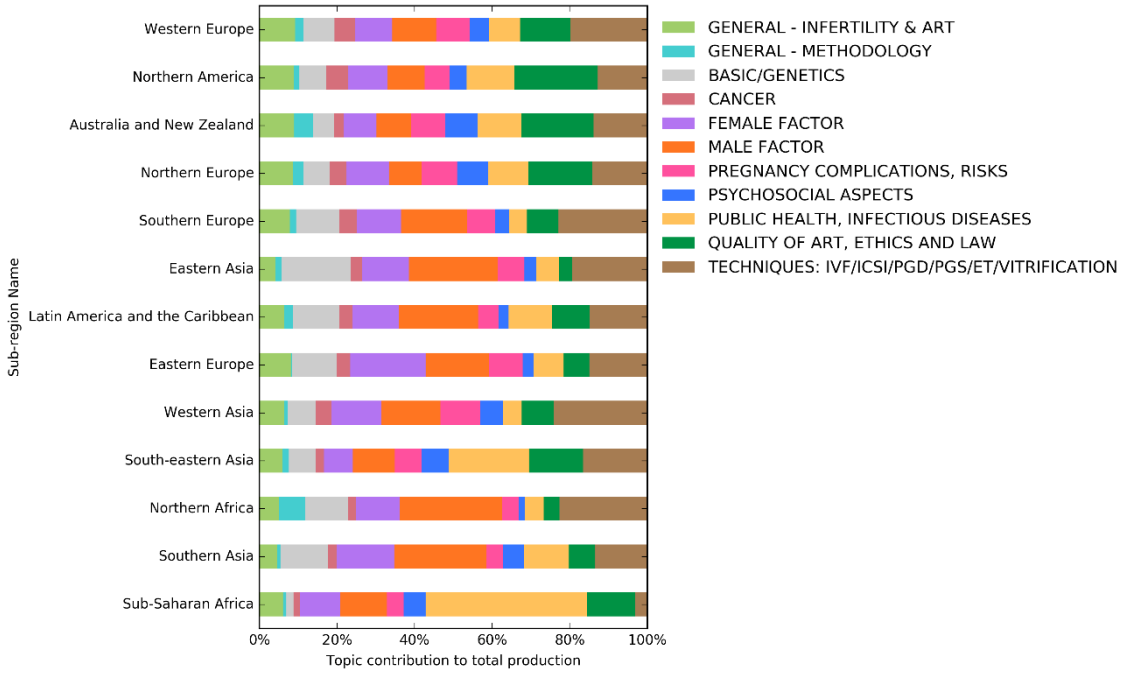
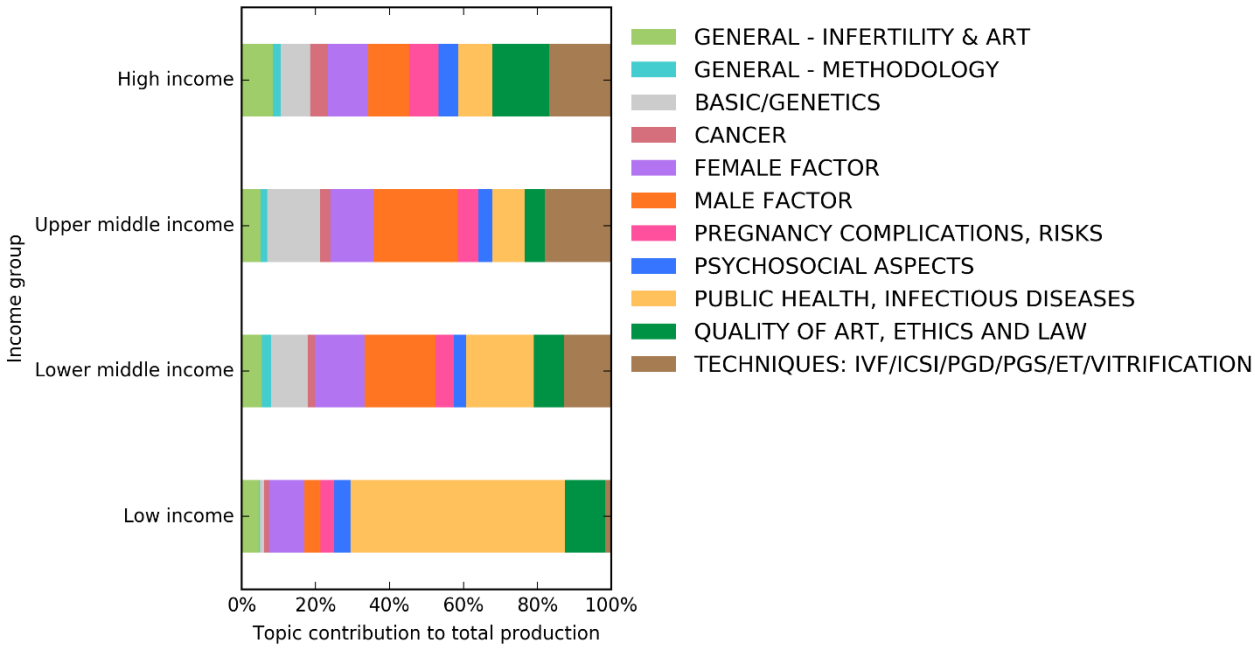


Figure VI



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2019 doi:10.20944/preprints201904.0170.v1

Figure VII

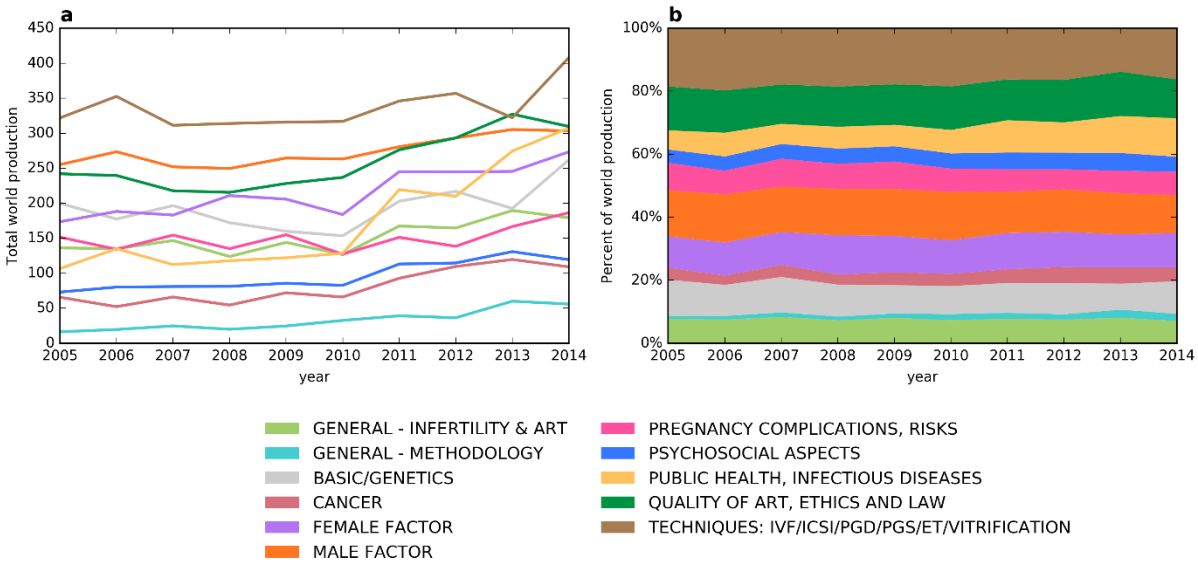


Figure VIII