

Trends in Fetal Medicine: A 10-Year Bibliometric Analysis of Prenatal Diagnosis

Ferdinand Dhombres, Olivier Bodenreider

National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

The objective is to automatically identify trends in Fetal Medicine over the past 10 years through a bibliometric analysis of articles published in *Prenatal Diagnosis*, using text mining techniques. We processed 2,423 full-text articles published in *Prenatal Diagnosis* between 2006 and 2015. We extracted salient terms, calculated their frequencies over time, and established evolution profiles for terms, from which we derived falling, stable, and rising trends. We identified 618 terms with a falling trend, 2,142 stable terms, and 839 terms with a rising trend. Terms with increasing frequencies include those related to statistics and medical study design. The most recent of these terms reflect the new opportunities of next-generation sequencing. Many terms related to cytogenetics exhibit a falling trend. A bibliometric analysis based on text mining effectively supports identification of trends over time. This scalable approach is complementary to analyses based on metadata or expert opinion.

Keywords:

Bibliometrics, Prenatal Diagnosis

Introduction

The availability of new genomic analysis techniques is transforming research and practice in Medicine. This is especially true of Fetal Medicine with the emergence of non-invasive prenatal testing (NIPT) procedures enabled by sequencing of circulating cell-free DNA (cfDNA) from a simple maternal blood sample [14]. This evolution is expected to be reflected through manuscripts published in *Prenatal Diagnosis*, the official journal of International Society for Prenatal Diagnosis. In fact, such advances in Fetal Medicine are regularly screened by members of the editorial board and summarized in a yearly editorial “In case you missed it” [2,6,7]. For example, cfDNA was discussed in the editorial presenting trends of the year 2015 [6].

Trend analysis often relies on manual review and expert opinion. For example, significant trends have been identified in medical literature, including increase in frequency and complexity of statistical reporting [1] and increase in computerized tomography and magnetic resonance imaging in Radiology research [12]. Bibliometric techniques have also proved useful for identifying trends in scientific disciplines [15,21], and could be used for capturing an unbiased evolution of major themes in Fetal Medicine over a longer period of time. In context of trend analysis, bibliometric techniques of choice are not citation metrics [4,20,26] (e.g., impact factor and h-index), rather those techniques used for analyzing metadata associated with scientific articles [9,11,13,16] (e.g., indexing terms) and the text of these articles [10]. Surprisingly, use of text mining techniques on full-text articles has not been reported for trend analysis purposes.

The objective of this investigation is to automatically identify trends in Fetal Medicine over the past 10 years through a bibliometric analysis of articles published in *Prenatal Diagnosis*, using text-mining techniques.

Methods

We conducted a bibliometric analysis of 2,423 full-text articles published in *Prenatal Diagnosis* over a 10-year period, from January 1, 2006 to December 31, 2015. Our approach can be summarized as follows. We extracted salient terms from the articles; calculated their frequencies over time; and established evolution profiles for most frequent terms, from which we derived falling, stable, and rising trends.

Extracting salient Fetal Medicine terms

We processed the full-text articles to extract all sequences of consecutive words (“N-grams”) of 5 words or less, most likely corresponding to medical terms. Let us consider the sentence “Currently, commercial applications of cell-free fetal DNA testing include RhD blood group typing” [19]. Examples of N-grams extracted from this sentence include “fetal” and “DNA” (N=1); “fetal DNA” and “testing include” (N=2); “cell-free fetal DNA” (N=3); “RhD blood group typing” (N=4); and “testing include RhD blood group” (N=5). Not all N-grams are expected to correspond to medical terms, let alone to salient Fetal Medicine terms. We used Apache Solr (<http://lucene.apache.org/solr/>) to extract N-grams.

Intuitively, common English words (i.e., non-medical words) or expressions and general medical terms are unlikely to be terms of interest. In contrast, terms frequently occurring in *Prenatal Diagnosis* are more likely to be salient terms. Therefore, as shown in Figure 1 we filtered out all N-grams entirely composed of common English words, e.g., “commercial applications” (filter #1); selected N-grams present in more than 10 articles in at least one year (filter #2); selected N-grams present in UMLS Metathesaurus [3], a large medical dictionary (filter #3); but excluded N-grams corresponding to general medical terms (isolated adjectives and terms categorized as “Concepts & Ideas” in UMLS Semantic Network), e.g. “mmol” and “arterial” (filter #4). Finally, one author (FD) manually reviewed terms excluded by these filters and rescued salient Fetal Medicine terms that were not covered by the medical dictionary (e.g., “cell-free fetal DNA”).

Calculating term frequencies

For each medical term, we recorded the number of articles in which it appears, for each year of the decade 2006–2015, and for the whole decade. Additionally, we determined the cumulative proportion of occurrences for each term in each year.

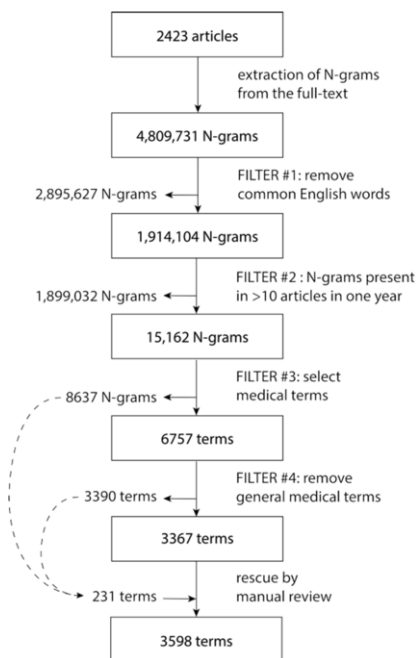


Figure 1 - Term extraction strategy applied to the 2,423 articles from Prenatal Diagnosis (2006-2015). Counts represent numbers of distinct terms (N-grams).

Establishing evolution profiles

Intuition here is that terms used mostly at the beginning of the 10-year period under investigation are becoming less popular (denoting a falling trend). In contrast, terms used mostly at the end of the decade have become more popular recently (denoting a rising trend). In practice, we used cumulative frequency over time to determine when terms were used most. We divided the decade into 3 periods, namely 2006- 2009, 2010-2011, and 2012-2015. For a given term, if 50% or more of all occurrences were observed during 2006-2009, its overall frequency is decreasing (falling trend). In contrast, if 50% or more of all occurrences were observed during 2012- 2015, its overall frequency is increasing (rising trend). Otherwise, the term was deemed “stable”. We extracted 30 most frequent terms in each trend group for visualization and further analysis. Additionally, we extracted 200 most recent terms among those exhibiting a rising trend, as they are likely to denote “hot terms”. Finally, we surveyed frequency evolution for a selection of terms, including those identified by editors of *Prenatal Diagnosis* as reflecting advances in Fetal Medicine for 2015[6].

To produce the evolution profiles, we used the R Foundation Computing environment [17] along with packages for text and data management [24,25] and for data visualization [22,23]. Excluding manual review of terms, it took about four hours to process the documents and compute evolution profiles.

Results

Extracting salient Fetal Medicine terms

From the 2,423 articles, we identified 3,598 salient medical terms. On average, the terms occurred in 101.9 articles over the decade. Our manual review rescued 231 (2.7%) of the 8,637 terms that had been inappropriately filtered out, including

“prenatal ultrasound”, “maternal plasma”, “fetal nuchal translucency” and “cell-free DNA”. These terms were present in 178.5 articles on average, ranging from 33 (for “fetoscopic laser photocoagulation”) to 883 articles (for “fetal medicine”).

Establishing evolution profiles

Distribution of terms according to year in which their cumulative frequency reaches 50% of their total document frequency is presented in Figure 2. We identified 618 terms with decreasing frequencies over time (falling trend), 2,142 stable terms, and 839 terms with increasing frequencies (rising trend). Not surprisingly, while stable terms occur in a large number of articles, terms with decreasing or increasing frequencies occur in fewer articles.

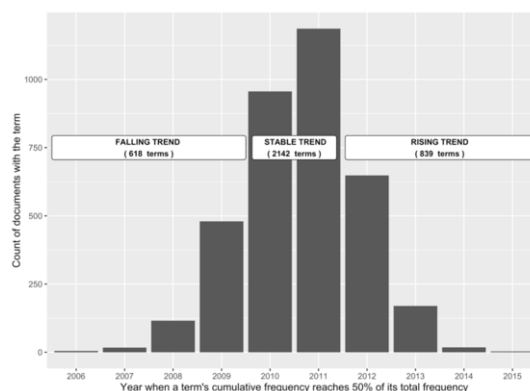


Figure 2 - Distribution of terms according to the year in which their cumulative frequency reaches 50% of their total document frequency.

Falling trend. Among the most frequent terms with decreasing frequencies, we found many terms related to Cytogenetics (e.g., “FISH”, “cytogenetic analysis”, “molecular cytogenetic”, “cytogenetic studies”). Of note, the term “case report” is the term whose frequency decreased most dramatically, dropping from 121 articles in 2006 to 58 articles in 2015. The top 30 terms exhibiting a falling trend are shown in Figure 3a. These terms reached 50% of their total document frequency before 2010.

Stable trend. Not surprisingly, many common terms in Fetal Medicine have relatively stable frequencies (Figure 3b). For example, the terms “pregnancy”, “fetus”, “ultrasound” were present in over 2,000 articles, and the terms “gestational age”, “karyotype”, “maternal age” and “amniocentesis” in over 1,000 articles. The terms “chorionic villus sampling” and “placenta”, present in over 500 articles are also stable over the decade. These terms reached 50% of their total document frequency in 2010 or 2011.

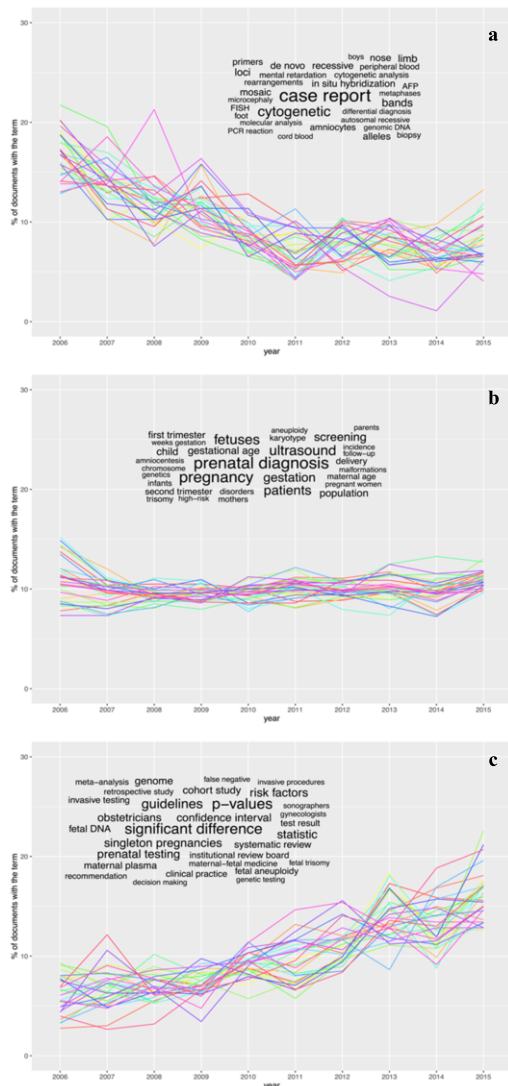


Figure 3 - Evolution of term frequency (coloured lines) over time for the top 30 terms exhibiting a falling trend (a), a stable trend (b) and a rising trend (c). (The font size in the term cloud is proportional to term frequency.)

Table 1 - List of the 20 most frequent of most recent terms exhibiting a rising trend (“hot terms”).

Rank	“hot topic”	Rank	“hot topic”
1	systematic review	11	positive predictive value
2	DNA sequences	12	maternal plasma DNA
3	fetal cell-free DNA	13	non-invasive prenatal diagnosis
4	invasive prenatal testing	14	microarray analysis
5	cell-free DNA	15	clinical setting
6	web	16	non-invasive prenatal testing
7	plasma DNA	17	single nucleotide polymorphism
8	aneuploidy detection	18	prospective cohort study
9	Genomics	19	collaborative study
10	exclusion criteria	20	genetic counselors

Table 2 - Trend for terms in four categories of interest (the most recent terms exhibiting a rising trend are marked ***, the arrows represent rising (↗), stable (→) and falling (↘) trends; df: document frequency).

Category	Term	Df	Trend
invasive diagnostic procedures	amniocentesis	981	→
	chorionic villus sampling	573	→
	fetal blood sampling	128	→
	cordocentesis	116	→
	ultrasound guidance	94	→
next-generation genetics	fetal cell-free DNA ***	244	↗
	fetal cells	229	→
	cell-free DNA ***	210	↗
	microdeletion	177	↗
	comparative genomic hybridization	172	→
	CGH	163	→
	massively parallel sequencing ***	107	↗
	copy number variation	89	↗
	genome sequencing ***	81	↗
	chromosomal microarray ***	79	↗
	CNVs ***	68	↗
	copy number variants ***	65	↗
	whole genome sequencing ***	65	↗
imaging procedures	shotgun sequencing ***	57	↘
	direct sequencing	48	↘
	exome ***	29	↗
	ultrasound	2032	→
	ultrasound examination	683	→
	ultrasound scan	501	→
	ultrasound screening	411	→
	MRI	346	→
	first trimester ultrasound	305	→
	magnetic resonance imaging	299	→
	fetal echocardiography	206	→
	Doppler ultrasound	194	→
	imaging procedures	170	→
fetal therapy procedures	fetal MRI	162	→
	three dimensional ultrasound	160	→
	second trimester ultrasound	102	↘
	X-ray	100	↗
	umbilical artery Doppler	70	→
	transvaginal ultrasound	55	→
	computed tomography	28	→
	heart ultrasound	138	→
	fetal therapy	91	→
	fetal surgery	80	↗
	laser surgery	76	→
	fetal intervention	75	→
	amnioreduction	61	→
	fetoscopy	40	↗
	fetoscopic laser surgery ***	37	↗
	fetoscopic laser coagulation	33	→
	in utero treatment	29	→
	in utero therapy	981	→

Rising trend. Terms with increasing frequencies include those related to statistics (e.g., “statistical analysis”, “p-value”, “significant difference”), medical study design and methods (e.g., “cohort study”, “systematic review”, “meta-analysis”, “ethics committee”, “institutional review board”), and clinical practice documents (e.g., “guidelines”, “recommendations”) (Figure 3c). These terms reached 50% of their total document frequency after 2011. Of particular interest, “hot terms” (i.e., the most recent terms exhibiting a rising trend) generally reflect the new opportunities of next-generation sequencing (“cell-free DNA”, “non-invasive prenatal testing”, “microarray analysis”). The list of the 20 most frequent hot terms is provided in Table 1; these terms exhibit a document frequency ranging from 377 to 123 over the decade.

Trend for specific terms. As expected, many terms identified by editors of Prenatal Diagnosis as reflecting advances in Fetal

Medicine for 2015 [6] were also captured by our approach among the recent terms exhibiting a rising trend (e.g., “fetoscopic laser surgery”, “monochorionic diamniotic twin pregnancies”, “placental insufficiency”, “placental function”). Further analysis of these terms is presented in the discussion section. We also surveyed specific terms for invasive diagnostic procedures, imaging procedures, fetal therapy procedures and next-generation genetics. As shown in Table 2, terms in two of these categories, namely invasive diagnostic procedures and imaging procedures, are generally stable. In contrast fetal therapy procedures and next-generation genetics tend to exhibit a rising trend, some of these terms having appeared very recently (“hot terms”).

Discussion

Trends in Fetal Medicine

Through a bibliometric analysis of articles published in *Prenatal Diagnosis*, using text mining techniques, we were able to identify trends in Fetal Medicine over the past 10 years.

Trends for diagnostic techniques. As expected, terms related to noninvasive prenatal testing exhibit a rising trend. More generally, terms denoting new genetic methods (e.g., “next generation sequencing”, “whole genome sequencing”, “single nucleotide polymorphism” or “microarray analysis”) are on the rise. In contrast, terms related to Cytogenetics (e.g., “molecular cytogenetic” or “FISH”) were highly used at the beginning of the decade, but are now less popular, reflecting a paradigm shift in Fetal Medicine. Interestingly, terms denoting invasive sampling techniques (“amniocentesis”, “choriocentesis”) remain stable in Fetal Medicine discourse, with a high number of occurrences across the decade, possibly because they continue to be mentioned as a reference when discussing newer techniques.

Trends in study design. In addition to trends for diagnostic techniques, our analysis identified trends in study design, namely an evolution toward structured studies reflected by a falling trend for “case report”, as well as a rising trend for “retrospective study” and for “meta-analysis”. The most recent terms (“hot terms”) include “prospective cohort study” and “systematic review”. Case reports are still given consideration for publication as research letters in *Prenatal Diagnosis*. However, a partnership with the journal *Clinical Case Reports* since 2013 may be the reason why fewer case reports end up being published in *Prenatal Diagnosis* nowadays. The rising trend for statistical methods, tests and variables is consistent with the observed evolution of study design towards structured epidemiological and clinical studies reported in the general medical literature [1].

Text mining vs. expert opinion

While our analysis is generally consistent with the trends identified by the editors of *Prenatal Diagnosis* as reflecting advances in Fetal Medicine over the past few years [2,6,7], some terms related to fetal surgery do not appear in our lists of terms exhibiting a rising trend, simply because their frequency is below that of top terms in this group. For example, although it exhibits a rising trend, the term “fetoscopic laser surgery” occurs only in 40 articles during the decade. Similarly, the terms “fetal therapy”, “in utero treatment”, “fetoscopy”, “fetal surgery”, “diaphragmatic hernia”, “spina bifida” or “twin-twin transfusion syndrome” are stable but occur in less than 210 articles. Evaluation of placental function was also deemed as a major advance in 2015 [6], and our analysis also finds a rising trend (but

limited frequencies) for “placental function”, “placental dysfunction” and “placental insufficiency”. Interestingly, although clearly identified in our analysis, trends in study design discussed above were not reported in editorials of the journal (probably because they do not reflect advances in diagnostic techniques per se). Moreover, stable and falling trends are not reported in editorials, but they are identified by our bibliometric analysis.

Text mining vs. metadata analysis

The medical literature referenced in PubMed/MEDLINE is indexed with Medical Subject Heading (MeSH) thesaurus. Therefore, an analysis of the indexing terms (MeSH descriptors) assigned to *Prenatal Diagnosis* articles could also help identify trends in Fetal Medicine. However, MeSH descriptors have limited granularity and there is often a delay between publication and indexing.

Limited granularity. MeSH has a limited number of descriptors for indexing *Prenatal Diagnosis* articles. In addition to the descriptor “*Prenatal Diagnosis*”, there are 7 more specific descriptors, namely “*Amniocentesis*”, “*Chorionic Villi Sampling*”, “*Fetoscopy*”, “*Maternal Serum Screening Tests*”, “*Ultrasonography, Prenatal*”, “*Cervical Length Measurement*”, and “*Nuchal Translucency Measurement*”. Arguably, this granularity is insufficient for specific bibliometric analyses and cannot match granularity resulting from text mining techniques.

Delay between publication and indexing. There is a delay between time of publication and indexing. For example, in May 2016, 53% of articles published by *Prenatal Diagnosis* in 2015 were still awaiting indexing. Moreover, MeSH thesaurus is updated on a yearly basis, with some exceptions for public health emergencies (e.g., the term “*Zika Virus Infection*” was added to MeSH ahead of normal maintenance cycle). There is usually a delay between emergence of a new phenomenon and its availability as a MeSH descriptor. For example the term “*Maternal Serum Screening Tests*” was introduced in MeSH in 2013, whereas the first articles on the subject were published over 30 years ago [5]. (Of note, a specific term for “cell-free DNA” is currently under consideration for introduction in MeSH.) Therefore, our approach based on text mining is better suited for identifying trends in a timely fashion.

Limitations and perspectives

For text mining purposes, we had to extract text of articles from PDF documents, which are optimized for human readability, rather than automatic text processing. For example, we had to eliminate text of headers and footers to avoid extracting the name of the publisher present on each article as a “frequent term”. Similarly, we had to ignore words containing digits, which resulted in absence of potentially important terms, such as “*b2-microglobulin*”, “*CRISPR/Cas9*”, and many gene names (e.g. “*CHD7*” or “*FGFR3*”). Availability of *Prenatal Diagnosis* corpus in computer-friendly formats, such as XML, would make text mining analyses simpler and more reliable.

As mentioned earlier, we had to manually review terms excluded by our medical term filter and rescue 2.7% of them for analysis, including “*fetal nuchal translucency*” and “*cell-free DNA*”. This is a consequence of limited coverage of Fetal Medicine terms in standard terminologies integrated in UMLS. Recent inclusion of Human Phenotype Ontology [18] into UMLS (version 2015AB) brought some important terms for postnatal phenotypes, but coverage of Fetal Medicine remains limited [8].

Conclusion

Through a bibliometric analysis of articles published in *Prenatal Diagnosis*, using text-mining techniques, we were able to identify trends in Fetal Medicine over the past 10 years. These trends are related to diagnostic techniques (Cytogenetics is progressively replaced by non-invasive techniques based on Genomics) and to study design (Fetal Medicine increasingly relies on scientific methods, including statistics and bioinformatics).

Our bibliometric analysis identified trends that are consistent with those identified by experts (about recent diagnostic techniques), but also identified other interesting trends (about study design), and provided an account for terms exhibiting falling trends and stable terms. In practice, bibliographic analysis and expert opinion are complementary approaches to identifying trends in Fetal Medicine.

PubMed/MEDLINE indexing based on MeSH offers limited granularity and a delay that is not compatible with identification of trends in a rapidly evolving domain, such as Fetal Medicine. We observed that coverage of Fetal Medicine, in MeSH, and standard terminologies integrated in UMLS is limited. List of terms identified through our text mining analysis could be basis for developing a terminology for Fetal Medicine. The list of terms and their evolution profiles are available upto request to the authors.

In summary, a bibliometric analysis based on text mining effectively supports identification of trends over time. This scalable approach is complementary to analyses based on metadata or expert opinion.

Acknowledgements

This work was supported in part by Intramural Research Program of NIH, National Library of Medicine, French Gynecology and Obstetrics Association (*Collège National des Gynécologues et Obstétriciens Français*), and Philippe Foundation.

References

- [1] L.D. Arnold, M. Braganza, R. Salih, and G.A. Colditz, Statistical trends in the Journal of the American Medical Association and implications for training across the continuum of medical education, *PLoS One* **8** (2013), e77301.
- [2] D.W. Bianchi, T. Van Mieghem, L.G. Shaffer et al, In case you missed it: the Prenatal Diagnosis section editors bring you the most significant advances of 2013, *Prenat Diagn* **34** (2014), 1-5.
- [3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res* **32** (2004), D267-270.
- [4] R. Bolli, The 10 Most Read Articles Published in Circulation Research in 2015, *Circ Res* (2016).
- [5] H.S. Cuckle, N.J. Wald, and R.H. Lindenbaum, Maternal serum alpha-fetoprotein measurement: a screening test for Down syndrome, *Lancet* **1** (1984), 926-929.
- [6] J. Deprest, A. Ghidini, T. Van Mieghem et al, In case you missed it: the Prenatal Diagnosis editors bring you the most significant advances of 2015, *Prenat Diagn* **36** (2016), 3-9.
- [7] B.H. Faas, A. Ghidini, T. Van Mieghem et al, In case you missed it: the Prenatal Diagnosis editors bring you the most significant advances of 2014, *Prenat Diagn* **35** (2015), 29-34.
- [8] M. Girdea, S. Dumitriu, M. Fiume et al, PhenoTips: patient phenotyping software for clinical and research use, *Hum Mutat* **34** (2013), 1057-1065.
- [9] Y. Hong, Q. Yao, Y. Yang et al, Knowledge structure and theme trends analysis on general practitioner research: A Co-word perspective, *BMC Fam Pract* **17** (2016), 10.
- [10] L.J. Jensen, J. Saric, and P. Bork, Literature mining for the biologist: from information retrieval to biological discovery, *Nat Rev Genet* **7** (2006), 119-129.
- [11] F. Li, M. Li, P. Guan et al, Mapping publication trends and identifying hot spots of research on Internet health information seeking behavior: a quantitative and co-word biclustering analysis, *J Med Internet Res* **17** (2015), e81.
- [12] K.J. Lim, D.Y. Yoon, E.J. Yun et al, Characteristics and trends of radiology research: a survey of original articles published in AJR and Radiology between 2001 and 2010, *Radiology* **264** (2012), 796-802.
- [13] R. McLean, K. Mendis, B. Harris et al, Retrospective bibliometric review of rural health research: Australia's contribution and other trends, *Rural Remote Health* **7** (2007), 767.
- [14] M.E. Norton, B. Jacobsson, G.K. Swamy et al, Cell-free DNA analysis for noninvasive examination of trisomy, *N Engl J Med* **372** (2015), 1589-1597.
- [15] S. Perez, V. Laperriere, M. Borderon et al, Evolution of research in health geographics through the International Journal of Health Geographics (2002-2015), *Int J Health Geogr* **15** (2016), 3.
- [16] A. Pinter, Changing Authorship Patterns and Publishing Habits in the European Journal of Pediatric Surgery: A 10-Year Analysis, *Eur J Pediatr Surg* **25** (2015), 353-358.
- [17] R Core Team, R: A Language and Environment for Statistical Computing (version 3.2.2), *R Foundation for Statistical Computing, Vienna, Austria*, (2015).
- [18] P.N. Robinson, S. Kohler, S. Bauer et al, The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, *Am J Hum Genet* **83** (2008), 610-615.
- [19] L.C. Sayres, M. Allyse, M.E. Norton et al, Cell-free fetal DNA testing: a pilot study of obstetric healthcare provider attitudes toward clinical implementation, *Prenat Diagn* **31** (2011), 1070-1076.
- [20] M. van Wesel, Evaluation by citation: trends in publication Behavior, Evaluation Criteria, and the strive for high impact publications, *Sci Eng Ethics* **22** (2016), 199-225.
- [21] G.T. Venable, B.A. Shepherd, C.M. Loftis et al, Bradford's law: identification of the core journals for neurosurgery and its subspecialties, *J Neurosurg* **124** (2016), 569-579.
- [22] J. Weiner, tagcloud: Tag Clouds. R package version 0.6, (2015).
- [23] H. Wickham, ggplot2: Elegant Graphics for Data Analysis., *Springer-Verlag New York* (2009).
- [24] H. Wickham, stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0, (2015).
- [25] H. Wickham and R. François, dplyr: A Grammar of Data Manipulation. R package version 0.4.3, (2015).
- [26] A. Zietman, The Red Journal's Top Downloads of 2014, *Int J Radiat Oncol Biol Phys* **93** (2015), 4-6.

Address for correspondence

Olivier Bodenreider, MD, PhD

8600 Rockville Pike, 38A/09S904, Bethesda, MD 20894, USA

Phone: (301) 827-4982, E-mail: olivier.bodenreider@nih.gov