Preventive Medicine 111 (2018) 466–472

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Review Article

Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis☆

Andrea Ramirez Varela[a,*], Michael Pratt[b], Jenine Harris[c], Jesse Lecy[d], Deborah Salvo[e], Ross C. Brownson[c,f], Pedro C. Hallal[a]

[a] Post-Graduate Program in Epidemiology, Federal University of Pelotas, Brazil
[b] University of California San Diego, School of Medicine and Institute for Public Health, United States
[c] Prevention Research Center in St. Louis, Brown School, Washington University School of Medicine, Washington University in St. Louis, United States
[d] Syracuse University, United States
[e] The University of Texas Health Science Center at Houston (UTHealth), School of Public Health in Austin, United States
[f] Department of Surgery and Alvin J. Siteman Cancer Center, Washington University School of Medicine, Washington University in St. Louis, United States

ARTICLE INFO

Keywords:
Physical activity
Network analysis
History
Research
Review

ABSTRACT

Little has been published about the historical development of scientific evidence in the physical activity (PA) and public health research field. The study aimed to examine the evolution of knowledge in this field.

A structured literature review using formal citation network analysis methods was conducted in June-2016. Using a list of influential PA publications identified by domain experts, a snowball sampling technique was used to build a compact citation network of 141 publications that represents the backbone of the field. Articles were coded by study type and research team characteristics, then analyzed by visualizing the citation network and identifying research clusters to trace the evolution of the field.

The field started in the 1950s, with a health sciences focus and strong North American and European leadership. Health outcome studies appeared most frequently in the network and policy and interventions least. Critical articles on objective measurement and public policy have influenced the progress from an emphasis on health outcomes research at early stages in the field to the more recent emerging built environment and global monitoring foci. There is only modest cross-citation across types of study. To our knowledge, this paper is the first to systematically describe the development of research on PA and public health. The key publications include fundamental ideas that remain citable over time, but notable research and dissemination gaps exist and should be addressed. Increasing collaboration and communication between study areas, encouraging female researchers, and increasing studies on interventions, evaluation of interventions and policy are recommended.

1. Introduction

Physical inactivity is an important risk factor for chronic diseases such as diabetes, coronary heart disease, some cancers, depression and dementia (Bauman, 2004; Ding et al., 2016; Ekelund et al., 2016; Kohl et al., 2012; Lee et al., 2012; Sallis et al., 2016), and costs 67.5 billion dollars globally annually in health care expenditures and lost productivity (Ding et al., 2016). Since the first epidemiologic studies published in the 1950s there has been enormous growth in the number of papers, researchers, study types, and disciplines engaged in research on physical activity. However, little has been published about the historical development of scientific evidence in the field of physical activity and public health. Available publications are in the format of commentaries, review articles, and historic narratives (Blair and Powell, 2014; Paffenbarger et al., 2001; Park, 1995); but all lack a quantitative research approach. Citation analysis is a powerful tool that allows for a visual and objective representation of the past, present, and potential future directions of a research field (Lecy and Beatty, 2012). This information is important to identify knowledge gaps and communication barriers among research and practice communities, and may be helpful in moving the field forward. The aim of this study was to use citation analysis to provide insight into the evolution of knowledge in the field of physical activity and public health.

☆ No financial disclosures or conflicts of interest were declared by the authors of this paper.

* Corresponding author at: Post-Graduate Program in Epidemiology, Federal University of Pelotas, Rua Marechal Deodoro 1160 - Centro, Pelotas, RS 96020-220, Brazil.
E-mail address: aravamd@gmail.com (A.R. Varela).

https://doi.org/10.1016/j.ypmed.2017.10.020
Received 18 May 2017; Received in revised form 3 October 2017; Accepted 17 October 2017

2. Methods

A structured literature review was conducted from February 2015 to June 2016 using citation network analysis (Lecy and Beatty, 2012). A stepwise protocol (before, during, after) for citation data collection was conducted, through seven steps:

2.1. Prior to citation data collection

2.1.1. Identification of most cited publications

In June 2015, the most cited documents in the field since 1950 were searched in ISI Web of knowledge and Google Scholar (Lecy and Beatty, 2012), using the following search criteria: "physical activity AND public health", allowing keywords to be found anywhere in the text. All languages, countries, and study types were included. Documents included published articles, commentaries, books, and others (reports, dissertations). Resulting documents were ranked by number of citations. A final list of the forty most cited publications was derived from ISI Web of Knowledge and Google Scholar, by combining both lists, removing duplicates (n = 11), and leaving the 40 unique most cited articles. Citation counts data was not normalized by publication date.

2.1.2. Identification of most cited authors

To validate the preliminary list generated in Step 1, we undertook a systematic process for expert identification. A separate search for the most cited authors in the field was conducted in Scopus, Google Scholar and ISI Web of Knowledge using the same search criteria from step 1. Lists from each source were combined, duplicates were excluded, and authors were ranked based on their H index. Country of affiliation was included in the list. Authors were divided into five categories based on their expertise: 1) Physical activity levels, trends and measurement – the science of physical activity surveillance and measurement; 2) Determinants and correlates of physical activity – understanding why some people are active and others are not; 3) Health outcomes of physical activity – studies on the health outcomes of physical activity with physical activity as the main exposure variable; 4) Interventions in the field of physical activity that aim to increase physical activity as the primary objective; and 5) Policy and practice in the field of physical activity and public health. This classification system was originally developed for the 2012 Lancet Physical Activity Series (A. E. Bauman et al., 2012; Hallal et al., 2012; Kohl et al., 2012).

A list of the three most cited authors per category was created. To ensure adequate global research representation for each category, if all three authors were from the USA, the top two only were included and a third was identified by selecting the highest cited author from a non-US institution. The final list included academics from USA, Australia, Brazil, Japan, and Norway.

2.1.3. Expert validation of the list of most cited publications

Between April and June 2015 a letter of request including the list of 40 most cited physical activity articles from step 1 was sent by email to the 15 experts identified in step 2. Experts chose the ten articles they considered most influential for the field's development, and ranked by importance (ranking: 1–10). They were encouraged to suggest articles that were not on the list. The final list of most influential articles consisted of 15 articles with 4 or more expert votes each. The response rate from the initial author list was 80% (12/15). Three authors did not reply and therefore the next author in the list was invited to participate until 15 responses were achieved.

2.2. Citation data collection

2.2.1. Data collection with Citation Network Analyzer tool

The Citation Network Analyzer-CNA tool developed by Lecy et al. (Lecy and Moreda, 2011) was chosen for data collection because of its functions of citation link identification, citation patterns tracking, and selection of highly influential publications based on the PageRank indicator, which fit the study objectives.

A citation network was collected in August 2015 using the most influential papers (n = 15, step 3) as "seeds" for a snowball sample. The sampling technique builds the network by identifying articles citing the seeds, articles citing those, etc. Since a snowball sample grows exponentially, we utilized a constrained snowball, which collects only a percentage of articles at each level, retaining the highly-cited articles and discarding the rest, resulting in a compact sample that represents the backbone of a literature since it contains the most-cited articles and linkages between them and is not biased by researcher preferences (Lecy and Beatty, 2012).

Two parameters were considered for the constrained snowball sampling: a) number of levels of data collection from the seeds; and b) percentage of articles to be sampled by level. Seed articles constitute the baseline level of data collection. Two levels of data were collected: articles that cite the seeds are in level 1 and articles that cite level 1 are in level 2. For this study, it was estimated that an initial five level selection strategy would produce over 10,000 publications, too many for practical analysis and effective interpretation. Thus, data were collected in two levels from the previously selected fifteen seeds, with a sampling of the top 2% most cited articles at each level (Lecy and Beatty, 2012). This produced a citation network with 5217 articles and 9132 citation links.

This sample was further refined by filtering by the group of those at the 75th percentile and above for total citations, i.e., only articles with at least 674 Google Scholar citations. This subsample contained 1131 articles, including 80% of the original seed articles identified by field experts (step 3). Since this is the most highly-cited set of articles in the network it represents the arterial flows of research through the field (Lecy and Beatty, 2012). Appendix graph A includes a representation of the complete citation network explaining the need for filtering in order to conduct the main path analysis.

2.3. After citation data collection

2.3.1. Main path identification

The sample was further refined through main path analysis, a method to identify the set of articles that mathematically represents the optimal path for information to flow through the network between the seed articles and the last level of collected data. Links with the highest transversal weights were retained. Transversal weight is the proportion of all paths between the first/source document (not citing any others in the network) and the last/sink document (not cited by any others in the network) that contain a particular link or article. It represents the extent to which an article or link is needed for keeping the network connected (De Nooy et al., 2011; Harris et al., 2009). Using the search path count strategy to extract the main path (De Nooy et al., 2011), and based on a transversal weight cut point of 0.03 to ensure the inclusion of at least 80% of the seed articles, we obtained a network of 141 articles. This set contains the nodes with strongest citation linkages as you move forward in time from the seeds, representing the strongest path by which knowledge in the field has been generated and disseminated, i.e., it is the backbone of the literature (Harris et al., 2009).

2.3.2. Network data extraction

Abstracts from the main path articles (n = 141) were coded according to the previously described five categories, plus first author gender and country of residence based on affiliation. Main path abstract coding was conducted independently by authors AR and DS, who agreed with a weighted kappa of 0.77 and percent agreement of 82.8%. Discrepancies were resolved until reaching 100% agreement.

2.3.3. Statistical and graphic analyses

Traditional and network descriptive statistics and exponential random graph modeling (ERGM) were conducted to examine network