Comparison of the *h*-index scores among pathogens identified as emerging hazards in North America

Ruth Cox[1], K. Marie McIntyre[2], Javier Sanchez[1], Christian Setzkorn[2], Matthew Baylis[2], Crawford W. Revie[1]

Author affiliations

[1]Centre for Veterinary Epidemiological Research, Atlantic Veterinary College, University of Prince Edward Island, 550 University Avenue, Charlottetown, Canada

[2]Department of Epidemiology and Population Health, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK

Corresponding author:
Ruth Cox
Centre for Veterinary Epidemiological Research, Atlantic Veterinary College, University of Prince Edward Island, 550 University Avenue, Charlottetown, Canada, C1A 4P3.
Email: rucox@upei.ca
Telephone: +1 902 566 0815
Fax: +1 902 620 5053

Short title: The Hirsch index as a novel method to rank pathogens

Summary

Disease surveillance must assess the relative importance of pathogen hazards. Here we use the Hirsch index (*h*-index) as a novel method to identify and rank infectious pathogens that are likely to be a hazard to human health in the North American region. This bibliometric index was developed to quantify an individual's scientific research output and was recently used as a proxy measure for pathogen impact. Analysis of more than 3000 infectious organisms indicated that 651 were human pathogen species that had been recorded in the North American region. The *h*-index of these pathogens ranged from 0 to 584. The *h*-index of emerging pathogens was greater than non-emerging pathogens as was the *h*-index of frequently pathogenic pathogens when compared to non-pathogenic pathogens. As expected the *h*-index of pathogens varied over time between 1960 and 2011. We discuss how the *h*-index can contribute to pathogen prioritisation and as an indicator of pathogen emergence.

Keywords: *h*-index, bibliometric, pathogen hazard, pathogen prioritisation, pathogen emergence

Introduction

Effective disease surveillance and control rely on an ability to assess the relative importance of diseases and pathogens. Such prioritisation often involves the use of decision support tools to identify which diseases to target and where to focus resources and funding. These can be biased by the quality of evidence utilised, time taken for its collection and therefore the timeliness of results, or by the opinion of experts employed to make judgements on topics (McIntyre et al, 2011). In this work we use a quantitative method to identify and compare pathogens that are hazardous to human health in the North America region, which is quick and relatively simple to calculate, and we consider whether it might be used as a method to rank pathogens according to their impact. This novel method involves the use of the Hirsch index (*h*-index); a bibliometric index which was originally developed to quantify an individual's scientific research output (Hirsch, 2005), by accounting for the number of publications produced and the number of citations of those publications. "A scientist has an index *h* if *h* of his or her *Np* publications have at least *h* citations each and the other (Np-*h*) papers have ≤ *h* citations each" (Hirsch, 2005). It can be calculated using a range of bibliometric services such as those available from the Institute for Scientific Information's Web of Science (WOS) (Thomson Reuters, 2011), and is given as a standard metric in output generated by Google Scholar (Google, 2013). While the *h*-index was initially devised to assess the output of individual scholars, it has been extended to measure the productivity of research groups (Van Raan, 2006), and some services now provide *h*-index values associated with groups of keywords or phrases in a given bibliometric database (Thomson Reuters, 2011).

Recently the *h*-indices of a number of human pathogens (n=27) were shown to be significantly positively correlated with their impact as measured by disability adjusted life year (DALY) estimates (McIntyre et al, 2011). DALYs provide a combined measure of the years of healthy life lost as a result of poor health or disability in combination with an estimation of the potential years of life lost due to premature death. They were developed by the World Health Organization (Murray, 1994) and although they have only been estimated for a small number of diseases, they have become the most widely-used measure of the true burden of disease (Mathers et al, 2004). Thus, although the *h*-index is a measure that reflects global scientific interest in a pathogen (and inherently reflects research trends and funding), it is a reasonable proxy indicator of high impact human pathogens (McIntyre et al, 2011).

The aims of this work are to investigate:

1) whether the *h*-index might be used to identify the pathogens that are likely to have a high impact upon human health in the North American region.

2) how the *h*-index might be used to apply a relative ranking to a set of pathogens identified as emerging hazards. In this second aim we focus on examples of pathogens of interest to Canada because our research institution and funding agency are based in Canada.

3) how the *h*-index of a pathogen changes over time.

Materials and Methods

Identification of pathogen species

The 'ENHanCEd Infectious Diseases' (EID2) database (University of Liverpool, 2011) provided the raw data for this study. The purpose of this database is to provide a method of studying the main pathogens and hosts involved in disease transmission (McIntyre et al, 2013). It contains information about more than 740,000 organisms (such as vectors, hosts and pathogens) and their structure in the phylogenetic tree. This includes details regarding all pathogens that are known to infect humans and some known to infect domestic or companion animal species. Information about pathogens is assigned using data-mining of meta-data and semi-automated literature searches, for further details see McIntyre et al (2013).

Information was extracted about all organisms classified as ‘pathogen species’ that were known to infect humans. Analysis only included human pathogen species, since the database did not include all known North American animal host species. All data searches were undertaken in October 2011.

Information about the taxonomic division and pathogenic status of each pathogen was extracted from the database. Pathogenic status was defined as one of ‘frequently pathogenic’ (frequently causes morbidity and/or mortality in the general population), ‘non-pathogenic’ (does not frequently cause clinical signs within the general population, but may affect immune-compromised individuals) or ‘unknown’ (there was insufficient evidence to determine pathogenic effects).

Zoonotic potential and emerging status of the pathogens were taken from (Taylor et al, 2001; Woolhouse and Gowtage-Sequeria, 2005) where available. Definitions were described in those publications as follows. Zoonotic potential was classified as either ‘non-zoonotic’ or ‘zoonotic’. Zoonotic pathogens were those that are naturally transmitted between vertebrate animals and humans. Pathogens previously but no longer transmitted from animals, such as HIV, were not regarded as zoonotic. Pathogens were classified as ‘non-emerging’ or ‘emerging’. Emerging pathogens were those that have appeared in a human population for the first time, or have occurred previously but are increasing in incidence or expanding into areas where they had not previously been reported over the last 20 years. Once the *h*-index of pathogens had been calculated (see below), the indices of pathogenic versus non-pathogenic, zoonotic versus non-zoonotic and emerging versus non-emerging pathogens were compared using the non-parametric Mann Whitney U test.

Calculation of the h-index of pathogens

The *h*-index scores were obtained for all pathogen species named in the database using WOS (Thomson Reuters, 2011). The following specific search protocol was followed in order to identify all scientific papers relating to each pathogen, and thus to calculate the *h*-index of the pathogen.

For each pathogen, literature searches were undertaken using search phrases specified using quotation marks (“”), the ‘topic’ search field and with lemmatization turned off. Search phrases were compiled which included the scientific name and any alternative names, synonyms or alternative spellings according to the National Center for Biotechnology Information (NCBI) taxonomy website (National Center for Biotechnology Information, 2011); searches for organisms contained ‘exclusion terms’ when necessary. Searches for viruses were more complex because of the frequent existence of synonyms and acronyms. Synonyms and acronyms were obtained from the NCBI taxonomy website (National Center for Biotechnology Information, 2011) or the International Committee on Taxonomy of Viruses (ICTV) website (International Committee on Taxonomy of Viruses, 2011) and were included as additional search terms. Since some acronyms were used for more than one virus, or occurred in a non-viral context, searches also included the term ‘virus’ if they had ‘virus’ within their pathogen name or if they were within the ‘virus’ division of the NCBI Taxonomy database and excluded any other entities (viral or non-viral) which shared the acronym. The Boolean operators ‘AND’, ‘OR’ and ‘NOT’ were used to link multiple search phrases. For example the query for *Sin nombre virus* contained the following search terms: ('sin nombre virus' OR 'sin nombre hantavirus' OR ('snv' AND 'virus')) AND NOT ('spleen' OR 'sindbis'). All searches were restricted to the years from 1900 to 2011, inclusive. Search terms are available on request from the authors.

Identification of pathogens that occur in the North American region

Having calculated the *h*-index for all pathogens in the EID2 database, we firstly, identified which pathogens are able to establish in the North American region, and secondly, ranked them according to their *h*-index. Thus we use previous occurrence in the North American region as an indicator of the pathogens that are more likely to emerge again in the same area, either because they are endemic and have the potential to re-emerge or because, in the past, they have had the opportunity to establish in the region. Clearly this is a simple indicator, however it provides a method of identifying pathogens that are able to occur in a specific geographical region. Our ranking of pathogens 'of interest' to Canada (see below) takes into account pathogens that are exotic to Canada.

Two methods were used to identify which of the pathogens had been recorded in at least one of the following North American countries: Canada, United States, Mexico or Greenland. These countries (which we defined as the North American region and now refer to as 'North America') were selected to comprise the North American land mass, while excluding the countries of Central America for simplicity.

The first method involved searching for the pathogens within the NCBI Nucleotide database (National Center for Biotechnology Information, 2011a). This database is a collection of genome sequences from sequencing projects around the world. The metadata for nucleotide sequences in some cases contains information about the location of pathogen isolation. In order to identify location, searches established where the pathogen and at least one of the geographical ‘Medical Subject Headings’(MeSH) terms for Canada, United States, Mexico or Greenland co-occurred. MeSH terms act as a controlled thesaurus and are used for indexing articles by the US Library of Medicine (US National Library of Medicine, 2012a). If one sequence from a pathogen had been recorded in North America within the Nucleotide database, then this was used as confirmation of pathogen presence. A second method was also used to identify pathogen location because the NCBI nucleotide database did not include location information about all pathogens in our study. This second method used the PubMed database, a database that contains more than 21 million citations of biomedical and life sciences literature (US National Library of Medicine, 2012b). The database was searched for all publications where the pathogen search terms (described above) and at least one of the geographical MeSH terms for Canada, United States, Mexico or Greenland co-occurred. The search terms had to be recorded in the title or abstract of the publication. There was a degree of inaccuracy associated with this method, since co-occurrence of a pathogen and a North American search term does not necessarily indicate that the pathogen has occurred in that region. Co-occurrence could also arise in publications that describe pathogen absence, animal models or simulation models for example. In order to account for this inaccuracy, only searches for pathogens which generated at least five references in the same country were used as confirmation of pathogen presence in North America. The threshold of five was chosen following sensitivity testing of the results from searches conducted for 21 randomly selected pathogens. In brief, this involved stratifying the scientific publications according to the pathogen and the continent to which they were linked via a MeSH term for a country. The association was checked to substantiate that the pathogen was found in hosts (including humans) within a MeSH term country. This indicated that on average 95% of the associations in single papers were accurate. Therefore setting the threshold at five papers would result in a positive predictive value PPV (i.e. proportion of predicted interactions for which papers provide supporting evidence) exceeding 99.9%. A high enough threshold to avoid false positives was balanced with the need to avoid causing any major bias against 'newer' pathogens that have fewer publications. For detailed description see McIntyre et al, (2013).

Comparison of the h-index calculated from WOS with PubMed

The *h*-index was calculated from the WOS, which differs in its bibliographic content to other bibliographic databases. In order to compare the output from WOS with the PubMed database, non-parametric Spearman rank was used to correlate the WOS *h*-index of pathogens that occurred in North America with the number of publications for that pathogen in PubMed.

Ranking 'pathogens of concern' in Canada.

Additional descriptive analysis focused on 'pathogens of interest' in Canada. These were identified from three different sources. The first source was a recent publication which highlighted pathogens that are likely to emerge in North America in response to climate change (Greer et al, 2008). The second source was researchers from the Zoonotics Division of the Public Health Agency of Canada (PHAC) (N. Ogden *pers comm.*). PHAC provided funding support to the project and the researchers provided a list of pathogens that are of interest due to their potential to become emerging hazards in Canada. Details about the characteristics of these pathogens, including whether they have occurred in North America, were collated and they were ranked according to their *h*-index. The third source was the Ontario burden of infectious disease study (Kwong et al, 2010), a study that describes the mortality and morbidity of infectious diseases in Ontario. It lists three measures of disease burden for infectious diseases that have occurred in Ontario. The measures are YLL: Years of Life Lost due to premature mortality, YERF: Year-Equivalents of Reduced Function as a result of disease or condition, and HALY: Health Adjusted Life Years, which are calculated by adding the YLL and YERF for each pathogen. Spearman rank correlation was used to compare each of these measures with the *h*-index of the pathogen. Kwong et al, (2010) calculated the burden for a total of 69 diseases, however we only included those for which the pathogen was specified and could therefore be matched with an *h*-index. Thus, general terms describing a disease or condition such as 'Septicaemia' were excluded from this analysis.

Calculation of change in h-index over time

Time-bounded *h*-index scores were obtained for a selected set of pathogens using the same phrase searches as described above. However, here the cumulative *h*-index was calculated every year from 1960 to 2011 inclusive to assess how the index changed over time. The pathogens chosen were *Chikungunya virus, Hendra virus, Monkeypox virus, Nipah virus, Rift Valley Fever virus, Trypanasoma cruzi* (the cause of Chagas disease) and *West Nile Virus.* These pathogens were chosen as example pathogens that have been classified as either 'emerging' or 'non-emerging'; with the intent to compare the change in *h*-index of both types of pathogens. Furthermore, they are examples of pathogens that were deemed to be of particular interest to the

PHAC (N. Ogden *pers comm*.) due to their potential for emergence in Canada. In addition we calculated the cumulative *h*-index for *Plasmodium falciparum*, because it is a pathogen of worldwide concern and because preliminary calculations showed that it has one of the highest *h*-indices.

In order to assess the rate of change in *h*-index for these pathogens, two negative binomial models were evaluated. The first model assessed the cumulative *h*-index as the outcome where the time since first publication was used as an offset. Since the rate of change will be largely influenced by the number of publications for pathogens of major importance (e.g. *P. falciparum*), the second model assessed the rate of change of *h*-index by year. The outcome for this model was calculated by subtracting the cumulative *h*-index for a particular year from the previous year, except for the first year of the series. Similarly, the number of years since first publication was used as an offset. Both models included a categorical variable indicating the pathogen, a variable indicating the calendar year when *h*-index was computed and the interaction between these two variables. These models were assessed using the Deviance and X2 goodness of fit tests (Dohoo et al, 2009). The predicted rates from these models were calculated and plotted against time for each pathogen.

Results

Pathogens likely to have a high impact in North America

A total of 3627 pathogen species were recorded in EID2 and of these 1827 were classified as human pathogens species. Of these, 651 were human pathogens that have been recorded in North America. These consisted of 474 pathogen species that have occurred in North America according to the Nucleotide database, and an additional 177 pathogens that were identified when the pathogen search terms occurred in at least five publications in conjunction with the North American search terms entered into the PubMed database. A total of 258 occurred in both the Nucleotide database and the PubMed searches.

The *h*-index of the human pathogen species ranged from 0 to 584 and was highly over dispersed (Figure 1). Only a limited number of pathogens had an *h*-index over 100, with most pathogens scoring a relatively low value (median=37). Although, the *h*-index was calculated from WOS, which differs to some extent in its bibliographic content from PubMed, the *h*-index was significantly correlated with the number of publications recorded in the PubMed database (Spearman rank correlation r_s=0.736, p<0.001, n=651), (Figure 2).

The largest proportion (42.2%) of pathogen species were bacteria, followed by fungi (21.2%) and viruses/prions (16.0%) (Table 1). The 10 pathogens with the highest *h*-index included one yeast (*Saccharomyces cerevisiae*), five viruses and four bacteria species (Table 2).

Information about emergence status (emerging or non-emerging) and zoonotic potential (zoonotic or non-zoonotic) was obtained from two publications. These publications assigned an emergence status to 462 (71%) and a zoonotic status to 464 (71%) of the 651 pathogens included in our analysis. Of the 462 pathogens that had an emergence status, 26.2% were classified as emerging. Pathogen species with the highest *h*-index recorded in WOS that were classified as emerging included *Escherichia coli*, *Human immunodeficiency virus 1* and *2* and *Hepatitis C virus* (Table 2). Emerging pathogens had a significantly higher *h*-index than non-emerging pathogens (Mann Whitney U, p<0.001) (Table 3). A total of 464 of the pathogens had been assigned a zoonotic potential status and 67.9% of these were zoonotic (Table 3). The *h*-index values of zoonotic and non-zoonotic pathogens were not significantly different (Mann Whitney U, p = 0.718). Pathogens that were frequently pathogenic had significantly higher *h*-index scores than pathogens that were non-pathogenic (Mann Whitney U, p<0.001) (Table 3). There were 13 pathogens of 'unknown' pathogenicity, which were excluded from this analysis.

Using the h-index to apply a relative ranking to pathogens of interest

Additional analysis focused on pathogens that had been identified as potential emerging hazards within Canada in the literature or by PHAC. These pathogens were both endemic and exotic to Canada. Of the pathogens of interest to PHAC, two (*Plasmodium falciparum*) and Verotoxic *E. coli*) cause notifiable diseases in Canada (Public Health Agency of Canada, 2010) and three (*Nipah virus*, *Hendra virus* and *Rift Valley Fever virus*) had not previously been recorded in North America (and therefore did not feature in our list of North American pathogens).

All of the pathogens of interest from both sources were classed as frequently pathogenic. Of the pathogens that were highlighted by Greer *et al.* (2008), those with the highest *h*-index were *E. coli*, *P. falciparum* and *Streptococcus pneumoniae* (Table 4). All had previously been recorded in North America. Additional pathogens of concern to PHAC with the highest *h*-indices were *Trypanosoma cruzi* (the cause of Chagas Disease), *Nipah Virus* and *Hendra Virus* (Table 5). Only *T. cruzi* has been recorded in North America and had a much higher *h*-index (130) than any of the others deemed to be of interest by PHAC. Overall, the median *h*-index of the pathogens listed in Tables 4 and 5 is 82, which is considerably greater than the median value of 37 for all

North American pathogens analysed. The *h*-index of 31 of these 33 pathogens were ranked in the top 50% of the North American pathogens (figure 1). The only exceptions were the food and water borne pathogens *Cryptosporidium hominis* and *Shigella boydii.*

The *h*-index of pathogens was positively correlated with the HALY measure of pathogen impact in Ontario (Spearman rank correlation r_s=0.627, p<0.001, n=41), (Figure 3). The *h*-index was also positively correlated with the two measures that make up the HALY score, namely the YLL (r_s=0.676, p<0.001, n=41) and the YERF (r_s=0.448, p<0.003, n=41). Of the 20 pathogens with the highest HALY score in Ontario, a total of 8 also feature in the top 20 pathogens with the highest *h*-index, while 15 feature in the top 50 and 16 have an *h*-index of greater than 100. The strength of this correlation is likely influenced by a few very high impact pathogens and we highlight that there are also a few pathogens that have a relatively high *h*-index score, although a relatively low HALY measure, e.g. malaria.

Change in h-index over time

The *h*-index of pathogens varied considerably over time. Figure 4 shows the time series for seven pathogens; *P. falciparum* was excluded from this figure because it has a high *h*-index that tends to obscure the series of the other pathogens. The *h*-index of *Rift Valley fever virus* and *Monkeypox virus* increased gradually from 1960 onwards. This was also the case for *T. cruzi,* although the *h*-index value for this pathogen was much greater than for any of the other six. The *h*-index of *Chikungunya virus* increased gradually from 1960 onwards, showing a steep increase in 2005 until 2007. *West Nile virus* showed a steady increase in *h*-index from 1960, until around 1998 at which point it was associated with a sharp increase which only recently appeared to have reached a plateau. The *h*-index scores for *Hendra virus* and *Nipah virus* were zero until the mid-1990s, but then increased relatively rapidly until around 2005 when both appear to have plateaued to some extent.

When the *h*-index was adjusted for the number of years since the first record of the pathogen (or in the case of 'older' pathogens the record in 1960, when our dataset began), the pathogens with the highest *h*-index were *P. falciparum* and *T. cruzi* (Figure 5a) throughout this time. In the 2000s the *h*-index of the pathogens *Hendra virus* and *Nipah virus* increased more rapidly than the other pathogens tested. When the yearly rate of change of the *h*-index was measured (adjusted by discounting the *h*-index from previous years), the *h*-index of *P. falciparum* and of *West Nile virus* increased at a higher rate than any of the other pathogens (Figure 5b). In comparison the rate of change of the *h*-index for *T. cruzi* gradually decreased from 1960. Finally, both *Hendra virus* and *Nipah virus* showed a rapid increase in the 1990s, although in more recent years (since the early 2000s) the rate of change of the *h*-index of these viruses has decreased.

Discussion

Pathogens likely to have a high impact in North America

The *h*-index of a pathogen can be viewed as an indicator of the relative scientific interest in that pathogen. Although it likely reflects trends in research interest, research funding and regional bias, the *h*-index of a limited number of pathogens has been correlated with their DALY measure which suggests that it might be used as a measure of impact (McIntyre et al, 2011). We focused on human pathogens that have been recorded in North America. We used previous occurrence in North America as an indicator of the pathogens that are more likely to emerge again in the same area, because geographic proximity is a characteristic that has been deemed a risk for emergence, for example in Canada (Cox et al, 2012). Clearly, this is a simple indicator and other non-endemic pathogens that have not previously been recorded in the region could still emerge.

The species with the highest *h*-index values included yeast (*Saccharomyces cerevisiae*), which can cause opportunistic infection, food-borne pathogens (*E. coli*), person to person transmitted viruses (*Hepatitis* B and C virus, *Human Immunodeficiency virus*, human herpesvirus), bacteria that cause multiple clinical infections (*Staphylococcus aureus*) and person to person transmitted bacteria (*Helicobacter pylori*). While some of these pathogens have a high impact on the human population, others are likely to have generated a high *h*-index for other reasons. For example, the vast majority of publications about *S. cerevisiae* are related to its industrial use in brewing and baking, rather than opportunistic infection. Similarly, the high *h*-index of *S. aureus* is likely to be associated with non-zoonotic infections in multiple species, rather than simply human illness.

There are two implications of these findings. Firstly that there may be a need to refine our search terms, as we increase our understanding of the biases of the *h*-index. Secondly, that the *h*-index may be most useful for ranking selected pathogens of concern. We suggest, therefore that it might be most reliably used as one complementary component of a pathogen prioritisation risk assessment particularly since such studies often rely on qualitative data or expert opinion (Cox et al, 2012, Krause and Working Group on Prioritization at Robert Koch Institute, 2008). Indeed the first publication on the *h*-index notes (when assessing the *h*-index of the individual researcher) that, ‘a single number can never give more than a rough approximation to an individual's multifaceted profile, and many other factors should be considered in combination in evaluating an individual' (Hirsch, 2005).

Using the h-index to apply a relative ranking to pathogens of interest

An additional part of our work focused on pathogens that have been identified as 'pathogens of interest' in Canada. All of the pathogens that were identified by PHAC or by Greer et al, (2008) had a relatively high ranking *h*-index. Those with the highest *h*-index were *E. coli, P. falciparum* and *S. pneumonia*. These pathogens are likely to have a high *h*-index either because they tend to be virulent and/or because they spread relatively easily in the human population, either via vectors, food and water or from person to person. It is important to note that our analysis only included species level pathogens and that we did not differentiate between strains of pathogens. This may be a useful distinction in future analyses. *E coli*, for example, are a large and diverse group of bacteria, which includes both virulent and non-virulent strains as well as zoonotic and non-zoonotic strains. In our analysis, *E. coli* has been classified as zoonotic, because at least some strains are zoonotic. This group is likely to score a high *h*-index not only due to the impact of virulent strains such as zoonotic Verotoxic *E. coli* O157, but also due to the prevalence of non-zoonotic illness such as urinary tract infections and neonatal meningitis.

Within our list of pathogens 'of interest' there are some that score a relatively high *h*-index but that do not cause especially severe disease. Examples, with an *h*-index greater than 100, include *Salmonella enterica*, *Respiratory synctial virus* and influenza virus. These tend to cause mild symptoms in the general population, (although they can be severe in individuals who are immunocompromised). They are likely to have generated scientific interest (and therefore a high *h*-index) due to their morbidity and their ease of transmission. *Salmonella enterica* for example, which has an *h*-index of 107, causes a diarrheal infection and occurs worldwide. In Canada there are an estimated 6,000 to 12,000 cases per year (Health Canada, 2012), although it is likely that cases are under-reported and that the actual number of infections is much higher.

A positive correlation between the *h*-index and the HALY score indicated that the *h*-index is a proxy for this measure of pathogen impact in Ontario and it could therefore be used to rank pathogens that are known to occur in a specific region. While we found a positive relationship between the two measurements, we also show that the ranking needs to be interpreted in the correct context. For example, *P. falciparum* scores a high *h*-index, but a relatively low HALY in Ontario. This shows that it has a high impact on a global scale, but that its impact within the cooler climate of Ontario is relatively low because it does not commonly occur.

Overall, our analysis of pathogens ‘of interest’ from three different sources, supports the idea that the *h*-index could be a practical method to compare potential pathogen hazards. There are two particular ways that it could be best used. Firstly as a method to separate high and low priority pathogens and therefore act as a rapid screening method for pathogens that require further risk analysis. Secondly, to rank pathogens that are ‘of interest’ in a specific region. For example to rank pathogens that are exotic to a region, but are of concern due to their global impact, or to rank pathogens that are endemic in a region and that occur frequently enough to have become ‘of interest’.

Change in h-index over time

Analysis of time series data demonstrated that the *h*-index of a pathogen changes over time, even after accounting for the increasing trend in total number of publications. We hypothesise that the rate of change of the *h*-index might be used as a crude indicator of a pathogen's emergence and/or the spread of infection. *Hendra virus,* for example, was discovered in horses in Australia in 1994 and its *h*-index began increasing from 0 in 1995. Similarly the *h*-index of *Nipah virus* increased from the time that it was discovered in a pig population in Malaysia in 1999. The *h*-index of both of these recently emerging pathogens has increased rapidly since their identification compared to the other pathogens studied here. It is also notable that the *h*-index of *West Nile virus*, which increased steadily from 1960 showed a relatively rapid increase from 1999 onwards and we hypothesise that this increase coincides with the emergence of the disease in the Eastern United States in 1999 (Soverow et al, 2009). Finally, the increase in the *h*-index of Chikungunya virus from 2005 to 2007 coincides with its outbreak in the Indian Ocean territories in 2005 (Schuffenecker et al, 2006).

There is likely to be a bias in the *h*-index towards 'old' pathogens compared to newly emerging pathogens, for which papers have not yet had time to accumulate citations. Indeed, it has been suggested that the *h*-index can only provide a realistic assessment of the achievement of academics (and therefore in our work – the impact of a pathogen) who have been publishing for at least ten years (Harzing, 2008). One way to compare between pathogens with different lengths of 'academic publishing' is to divide the *h*-index by the number of years since the first publication, a measure referred to as the 'm-quotient' (Hirsch, 2005). Our analysis, which controlled for the number of years of publication, revealed how the rate of change of *P. falciparum* and *West Nile virus* was higher than the other pathogens tested. The high rate of increase in the *h*-index of the malaria pathogen reflects the impact of the disease for which there were approximately 219 million cases worldwide in 2010 (World Health Organisation, 2012). Although the impact in terms of mortality rates has fallen by 26% since 2000, the increasing *h*-index also accounts for the fact that malaria is a risk to over half of the world's population, and that international disbursements and government funding for malaria control rose steeply during this time (World Health Organisation, 2012). We suggest that the *h*-index of *West Nile virus* has increased at a high rate because this reflects the impact of the pathogen as it has spread across the USA since it emerged in 1999 and because its emergence has been attributed to climate warming (Soverow et al, 2009). In contrast, the rate of change in *h*-indices of other pathogens such as *T. cruzi* and *Chikingunya virus* have decreased yearly. These pathogens have both been described as 'neglected tropical diseases', which tend to be endemic in low income, developing regions and typically have a high morbidity, but low mortality (Hotez, 2011). The rate of change in the *h*-indices of newly emerging pathogens (*Nipah virus* and *Hendra virus*) showed a different pattern to that of the 'older' pathogens, with a rapid increase following their emergence which then slowed in more recent years. This trajectory is likely to reflect the increasing scientific interest in a newly emerging pathogen, which then levels off as knowledge is established.

Identifying patterns of disease emergence using bibliometric measures or electronic resources has proven a valuable tool to augment disease monitoring and surveillance. For example, patterns of disease reporting in ProMED (the Internet-based 'Program for Monitoring Emerging Diseases' (International Society for Infectious Diseases, 2013)) have been used as an early-warning of disease emergence (Cowen et al, 2006), while records of Internet queries have been used to track the spread of influenza infections (Google Flu Trends, 2012; Ginsberg et al, 2009) and Methicillin Resistant *S. aureus* (Dukic et al, 2011). Similarly, social networking tools such as Twitter, have proven to be real-time indicators of public health concerns, since the number of Twitter posts relating to 'swine flu' and/or 'H1N1' in 2009 correlated well with H1N1 incidence data (Chew and Eysenbach, 2010). Twitter has also been used to measure the uptake of research findings, with the number of tweets generated within the first 3 days of an article's publication being a good predictor of highly cited articles. A proposed 'twimpact factor' has therefore been suggested as a timely metric to gauge research impact and influence (Eysenbach, 2011). A comparison of 'twimpact factor' with *h*-index may provide some predictive value in the case of disease monitoring.

In comparison to the *h*-index, the indicators described above are more instantaneous measures and it is unlikely that the change in *h*-index could be used for real-time surveillance purposes due to the time lag in the measure of the *h*-index and the relative impact. In addition, newly emerging pathogens are likely to be under-represented. However, the trajectory of the *h*-index may be relatively predictable if combined with other measures. Work has shown that it is possible to predict the future *h*-index of scientists as far as five to ten years into the future, on the basis of additional publicly available information, including years since publishing their first article, number of distinct journals published in and the number of articles in five prestigious journals (Acuna et al, 2012).

Assessment of a wider range of pathogens would be beneficial, with a particular focus on emerging pathogens. Specific incidences of emerging diseases and global emerging disease hotspots have been identified in the past (Jones et al, 2008). Comparison of the *h*-index of pathogens with their global emergence may reveal the typical time delay between disease emergence and changes in associated *h*-indices, as well as whether there is a level of increase in *h*-index that can be reliably interpreted as an early warning of future disease emergence.

Comparison of the h-index with other bibliometric sources

The *h*-index scores in the present study were generated from one bibliometric source and comparison was not made with other sources. Other bibliometric services, such as SCOPUS or Google Scholar, search different literature sources over differing temporal periods. Although alternative sources produce slightly different *h*-index values, these tend to be comparable across platforms (McIntyre et al, 2011). Our work demonstrated a clear correlation between the *h*-index calculated in WOS and the number of publications recorded in PubMed.

Overall the *h*-index combines an assessment of both the quantity of publications and the quantity of citations. A pathogen cannot have a high *h*-index without having a substantial number of papers published about it. However the number of papers is not enough – a reasonable number of these papers need to have been cited in order to increase the *h*-index value. The *h*-index thus corrects for pathogens that might have a limited number of highly cited papers, or many that have not been cited. It therefore tends to highlight pathogens that generate a continuous stream of publications with above average publication impact. While the *h*-index is the most commonly cited metric, alternative methods of assessing research output have been suggested (Harzing, 2007; Alonso et al, 2009) and might be considered in future assessments of pathogen impact. For example, the *g*-index could be used to give more weight to highly-cited articles (Egghe, 2006) and has been suggested as a useful complement to the *h*-index (Harzing, 2008). We also suggest an evaluation of the measure '*cf*', which takes into account the differences in number of citations received by all articles in a given year, so that scientific impact can be compared across years (Radicchi et al, 2008).

In conclusion, the *h*-index is a quantitative measure that can be used to estimate the potential impact of a pathogen and that can be calculated quickly and easily. It can be used to identify and to rank individual pathogens or types of pathogens (e.g. zoonotic, emerging and pathogenic) and to measure changes over time. It could provide a rapid method of screening for pathogens that are likely to be important and therefore it would be particularly useful if incorporated into a prioritisation tool to complement a set of more qualitative criteria.

Acknowledgments

The authors would like to acknowledge the discussion and advice offered by Alex Demarsh, Phillip Hepworth and Nicholas Ogden. Funding for this work was provided by the Public Health Agency of Canada (http://www.publichealth.gc.ca), grant number: 4500217733, awarded to CWR and JS and was also part of the ENHanCE project (http://www.liv.ac.uk/enhance/) funded under the ERA EnvHealth scheme by NERC (grant number NE/G002827/1, http://www.nerc.ac.uk), awarded to MB.

Author contributions

Conceived and designed the study: RC, KMM, JS, CS, MB, CWR

Provided the *h-indices* for analysis: KMM, CS, MB

Performed the data collection: RC, KMM, CS

Analyzed the data: RC, KMM, JS, CS, CWR

Wrote and commented on the manuscript: RC, KMM, JS, CS, MB, CWR

Figure captions

Figure 1 A scatter-plot showing the *h*-index value (y axis) of 651 pathogen species that are infectious to humans and have been recorded in the North American region against the rank position of each of those pathogens (x axis).
Points that have been coloured black indicate pathogens that were identified as potential emerging hazards and therefore of interest in Canada.

Figure 2 Correlation of *h*-index with number of publications reported in the PubMed database for 651 human pathogen species.
(The x-axis has been truncated at 300 to better demonstrate the association of the *h*-index and the total number of publications. There were only four pathogens with an *h*-index greater than 300. These were *Saccharomyces cerevisiae*, *Escherichia coli*, *Human Immunodeficiency virus 1* and *2*).

Figure 3 Correlation of *h*-index with Health Adjusted Life Year measurement of 36 infectious diseases that occurred in Ontario in 2010.
Data labels show the pathogen or disease named in the study by (Kwong et al, 2010).

Figure 4 The *h*-index score by year from 1960 to 2009 for seven selected pathogens.

Figure 5 The modelled rate of change of the *h*-index from 1960 to 2009 for eight pathogens.
A. The *h*-index has been adjusted according to the number of years since the first record.
B. The yearly rate of change of the *h*-index. This model has been adjusted according to the number of years since the first record and it also discounts the *h*-index from previous years.

Table captions

Table 1 Taxonomic classification of 651 human pathogen species that have been recorded in the North American region.

Table 2 Pathogen species with the highest *h*-index recorded in Web of Science from those human pathogen species recorded in the North American region. All were classed as frequently pathogenic.

Table 3 Summary of *h*-index values for human pathogen species that have been recorded in the North American region, grouped according to emerging, zoonotic and pathogenic status
A total of 651 human pathogen species were recorded in the North American region, however not all had been assigned a status for each characteristic.

Table 4 The *h*-index of 33 pathogens that have been identified as an emergence risk in Canada by Greer et al (2008).

Pathogen names in grey indicate a pathogen that can cause the associated disease, but that is not commonly the main cause of the disease in the North American region.

Table 5 The *h*-index of six pathogens that have been identified as pathogens of emergence concern in Canada by the Public Health Agency of Canada (N. Ogden *pers comm*.).

References

Acuna, D. E., S. Allesina, and K. P. Kording, 2012: Future impact: Predicting scientific success. Nature 489, 201-202.

Alonso, S., F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera, 2009: h-index: A review focused on its variants, computation and standardization for different scientific fields. J. Informetr. 3, 273-289. Chew, C., and G. Eysenbach, 2010: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 5, e14118.

Cowen, P., T. Garland, M. E. Hugh-Jones, A. Shimshony, S. Handysides, D. Kaye, L. C. Madoff, M. P. Pollack, and J. Woodall, 2006: Evaluation of ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. J. Am. Vet. Med. Assoc. 229, 1090-1099.

Cox, R., C. W. Revie, and J. S. Sanchez, 2012: The use of expert opinion to assess the risk of emergence or re-emergence of infectious diseases in Canada associated with climate change. PLoS ONE 7: e41590.

Dohoo, I. R., W. Martin, and H. Stryhn, 2009: Veterinary Epidemiological Research, Veterinary Epidemiological Research, 2nd edn , pp. 865-866. VER Inc.

Dukic, V., M. David, and D. S. Lauderdale, 2011: Internet Queries and Methicillin-Resistant *Staphylococcus aureus* Surveillance. Emerg. Infect. Diseases 17, 1068-1070.

Egghe, L., 2006: Theory and practice of the g-index. Scientometrics 69, 131-152.

Eysenbach, G., 2011: Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. J. Med. Internet. Res. 13, e123.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, 2009: Detecting influenza epidemics using search engine query data. Nature 457, 1012-1014.

Google, 2013: Google Scholar. Available at: http://scholar.google.ca/ (Accessed May 2012).

Google Flu Trends, 2012: Google flu trends around the world. Available at: http://www.google.org/flutrends/ (Accessed May 2012).

Greer, A., V. Ng, and D. Fisman, 2008: Climate change and infectious diseases in North America: the road ahead. CMAJ 178, 715-722.

Harzing, A., 2008: Reflections on the *h*-index. Available at: http://www.harzing.com/pop_hindex.htm (Accessed May 2012).

Harzing, A., 2007: Publish or perish. Available at: http://www.harzing.com/pop.htm. (Accessed May 2012)

Health Canada, 2012: Salmonella prevention. Available at: http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/food-aliment/salmonella-eng.php. (Accessed Dec 2012).

Hirsch, J. E., 2005: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. USA. 102, 16569-16572.

Hotez, P. J., 2011: The neglected tropical diseases and the neglected infections of poverty: Overview of their common features, global disease burden and distribution, new control tools, and prospects for disease elimination , National Academies Press, Washington DC.

International Committee on Taxonomy of Viruses, 2011: Taxonomy of Viruses. Available at: http://www.ictvdb.org/Ictv/index.htm (Accessed October 2011).

International Society for Infectious Diseases, 2013: ProMED. Available at: http://www.promedmail.org/.

Jones, K. E., N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak, 2008: Global trends in emerging infectious diseases. Nature 451, 990-993.

Krause, G., and Working Group on Prioritization at Robert Koch Institute, 2008: How can infectious diseases be prioritized in public health? A standardized prioritization scheme for discussion. EMBO Rep. 9 Suppl. 1, S22-7.

Kwong, J.C., N. S. Crowcroft, M. A. Campitelli, S. Ratnasingham, N. Daneman, S. L. Deeks, D. G. Manuel, and Ontario Burden of Infectious Disease Study Advisory Group 2010: Ontario Burden of Infectious Disease Study (ONBOIDS). An OAHPP/ICES Report. Ontario Agency for Health Protection and Promotion, Institute for Clinical Evaluative Sciences. Available at: http://www.publichealthontario.ca/

Mathers, C. D., D. Ma Fat, J. T. Boerma and World Health Organization, 2004: The global burden of disease: 2004 update. Available at: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/.

McIntyre, K. M., I. Hawkes, A. Waret-Szkuta, S. Morand, and M. Baylis, 2011: The H-index as a quantitative indicator of the relative impact of human diseases. PLoS ONE 6: e19558.

McIntyre, K. M., C. Setzkorn, M. Wardeh, P. J. Hepworth, A. D. Radford, and M. Baylis, 2013: Using open-access comprehensive database for the study of Mammalian and avian livestosk and pet infections. Prev. Vet. Med. http://dx.doi.org/10.1016/j.prevetmed.2013.07.002

Murray, C.J.L, 1994: Quantifying the burden of disease: the technical basis for disability adjusted life years. Bulletin of the World Health Organisation 72, 429-445.

National Center for Biotechnology Information, 2011a: Nucleotide database. Available at: http://www.ncbi.nlm.nih.gov/nuccore (Accessed October 2011).

National Center for Biotechnology Information, 2011b: Taxonomy browser. Available at: http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html (Accessed October 2011).

Public Health Agency of Canada, 2010: National notifiable diseases: notifiable diseases on line. Available at: http://dsol-smed.phac-aspc.gc.ca/dsol-smed/ndis/list-eng.php (Accessed May 2010).

Radicchi, F., S. Fortunato, and C. Castellano, 2008: Universality of citation distributions: toward an objective measure of scientific impact. PNAS 105, 17268-17272.

Schuffenecker, I., I. Iteman, A. Michault, S. Murri, L. Frangeul, M. C. Vaney, R. Lavenir, N. Pardigon, J. M. Reynes, F. Pettinelli, L. Biscornet, L. Diancourt, S. Michel, S. Duquerroy, G. Guigon, M. P. Frenkiel, A. C. Bréhin, N. Cubito, P. Després, F. Kunst, F. A. Rey, H. Zeller, S. Brisse, 2006: Microevolution of Chikingunya viruses causing the Indian Ocean outbreak. PLoS. Med. 3(7): e263.

Soverow, J. E., G. A. Wellenius, D. N. Fisman, and M. A. Mittleman, 2009: Infectious disease in a warming world: how weather influenced West Nile virus in the United States (2001-2005). Environ. Health. Perspect. 117, 1049-1052.

Taylor, L. H., S. M. Latham, and M. E. Woolhouse, 2001: Risk factors for human disease emergence. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 356, 983-989.

Thomson Reuters, 2011: Overview - Web of Science. Available at: http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/(Accessed October 2011).

University of Liverpool, 2011: ENHanCEd Infectious Diseases database (EID2). Available at: www.zoonosis.ac.uk/eid2 (Accessed October 2011).

US National Library of Medicine, 2012a: MeSH. Available at: http://www.ncbi.nlm.nih.gov/mesh (Accessed October 2011).

US National Library of Medicine, 2012b: PubMed. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed (Accessed October 2011).

Van Raan, A., 2006: Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics 67, 491-502.

Woolhouse, M. E., and S. Gowtage-Sequeria, 2005: Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11, 1842-1847.

World Health Organisation, 2012: World Malaria Report 2012, World Health Organisation, Geneva.