Future MICROBIOLOGY

COMMENTARY

Why is scientific research on 'data-poor' microorganisms being ignored?

"Should we really be complacent in allowing these specimens and their valuable, albeit 'data poor', information to be lost for future generations?"

John P Hays*

First draft submitted: 7 April 2017; Accepted for publication: 12 April 2017; Published online: 26 May 2017

Many academic laboratories across the world may still hold stocks of bacteria, viruses, fungi and parasites that are (many) decades old. These precious isolates are most probably frozen and lying forgotten at the back of a freezer, or freeze dried in glass vials and sitting gathering dust in the corner of cupboards and store rooms within microbiological laboratories across the world. Significantly, these microorganisms will most likely be preserved along with only a limited amount of detailed epidemiological data, which means that these 'data-poor' isolates (lacking all but basic epidemiological data) are very likely to be discarded due to their perceived lack of scientific value, possibly to make room for the storage of more recent isolates, to reduce storage costs, or for the decommissioning or rehousing of microbiological laboratories. Importantly, even if these data-poor microorganisms are eventually investigated by scientists, the research results obtained will almost invariably be rejected for publication by scientific journals due to a perceived lack of scientific significance or interest. Should we really be complacent in allowing these specimens and their valuable, albeit 'data poor', information to be lost for future generations? After all, these isolates represent an untapped resource of biological and genetic information for future generations, which once discarded cannot be retrieved. Why are scientists and scientific journal editors not discussing this issue and providing guidance, collaboration, research and publishing opportunities for data-poor microorganisms? Here the author explains his viewpoint and provides access to a simple website [1] for the exchange of basic information on data-poor isolates and for the promotion of this subject.

KEYWORDS
• data-poor microorganisms • scientific journals • whole-genome sequencing • WGS

Definitions

In the context of this perspective article, 'data-poor' microorganisms means microorganisms collected a minimum of 5 years ago, where some or all of the accompanying epidemiological data is missing. That said, these data-poor microorganisms should still retain some essential epidemiological information, country of origin and decade of isolation as an absolute minimum. In this context, genus and species names do not have to be known as these can be readily determined after culture and identification (e.g., via the use of matrix-assisted laser desorption/ionization TOF mass spectrometry). Of course, it is also possible to include data-poor microorganisms in

*Department of Medical Microbiology & Infectious Diseases, Erasmus University Medical Centre (Erasmus MC), Rotterdam, The Netherlands; j.hays@erasmusmc.nl

Future Medicine part of fsg

"Currently, it is the final part of the structure of these scientific data packets that dominates scientific thinking, that is, data have to be placed into a 'short-term' context (regional, global or chronological) in order to have a publishable or usable scientific value."

this definition which possess more refined epidemiological data, such as region of origin, year of isolation, genus and species etc. However, this article is specifically concerned with those isolates where 'epidemiologically rich' research involving detailed and accurate geographical and chronological epidemiological data is not available and where there is a high risk of the data-poor isolates being discarded due to a perceived lack of scientific value. Admittedly, this definition is somewhat arbitrary and deliberately excludes data-poor microorganisms that have been collected less than 5 years ago, largely because it is presumed that advances in computer-based data storage technology means that data-poor microorganisms less than 5 years old are becoming much less likely to exist. Further, techniques such as whole-genome sequencing (WGS) are beginning to be routinely used to characterize the epidemiological relationships between hospital outbreaks of infection [2] or the epidemiology of foodborne-disease outbreaks [3], which means that there is currently more emphasis on collecting and correctly storing microbiologically-linked epidemiological data. Of course, data-poor microorganisms collected less than 5 years ago may still represent scientifically interesting isolates if they possess rare genotypic or phenotypic characteristics.

Data loss

The lack of explicit epidemiological information for 'data-poor' microorganisms may frequently be traced back to the fact that at the time of isolation, the standard method for recording clinical information was pen and paper, and that after decades of storage, these paper files (but not the relevant microbiological isolates) were somehow 'lost' to their original owner. This process of 'losing' clinical information could be associated with the simple degradation of paper files over time, the disposal of the paper files after the relevant microbiologist has obtained alternative employment or the disposal of the paper files after the microbiologist has retired. Exacerbating this issue is the fact that the relevant paper files would be stored separately from the microbial isolates, meaning that the link between the paper data and the microbial isolates is lost over many years of storage. Additionally, psychologically speaking, microbiologists may generally attach less significance to disposing of 'pieces of paper' rather than actual microbial isolates themselves. Interestingly, the fact that clinical or diagnostic study records are currently stored electronically rather than on paper does not necessarily mean that this problem will disappear, as pressing the 'delete' key on data stored in a computer is actually much easier than the physical act of disposing of paper in a waste bin. However, it is now much easier to include all of the information obtained from a clinical or diagnostic study within a single document on a computer and to easily make back-ups of large data files without the chance of individual pieces of paper being lost or misplaced. Of course, using a computer in conjunction with single large data files also means that lots of data can be easily deleted or lost perhaps due to computer errors that lead to disk formatting or the loss of a data file's password over time, for example, after the retirement of the respective data manager. Additionally, large data files can be stored and conveniently transported on 'memory sticks' and one of the problems of using memory sticks is the fact that they are small and very easy to transport (and therefore very easy to lose), as well as the fact that (accidentally) deleting large amounts of information on a memory stick is much easier than (accidentally) deleting several volumes of paper records.

Contextualization & scientific dogma

So how important is the presence of epidemiologically rich data within the context of microbiology, scientific research and the advancement of knowledge? This depends on contextualization or the actual context in which the data is placed. Currently, it is very easy to say that data-poor microorganisms possess no real publishable value because of the lack of precise clinical or diagnostic information accompanying the isolates. There currently exists therefore a tendency to ignore research on data-poor microorganisms and to risk simply discarding uninvestigated microorganisms because the 'limited' information they possess does not fit into the currently accepted practice of generating and publishing scientific information within a 'single packet' of data. These data packets (usually scientific publications) tend to possess a strictly defined structure, which sequentially include: a definition of the problem (introduction); a definition of the techniques used to investigate the problem (materials and methods); a definition of the results (results); and the placement of the findings into a specific context, including possible pitfalls of the research

itself and suggestions for future research objectives (discussion and conclusion). Currently, it is the final part of the structure of these scientific data packets that dominates scientific thinking, that is, data have to be placed into a 'short-term' context (regional, global or chronological) in order to have a publishable or usable scientific value. This emphasis on contextualization is very similar to that which occurs when producing a movie or writing a book. An emphasis on context is used to interest the viewer or reader in order to provide entertainment, and ultimately to generate sales of the product. Without some form of contextualization, the movie or film would appear to be unfinished and generate fewer readers or sales. Within the context of scientific advances and the role of data-poor microorganisms, the question therefore remains as to whether current scientific research and publishing dogma is actually 'fit-for-purpose' and synergizes or antagonizes with our current scientific capabilities and future scientific needs? Is this dogma actually generating hurdles to possible future scientific breakthroughs and the advancement of knowledge by relying too much on packaging and short-term contextualization in order to sell scientific articles as short stories? After all, the more interesting the story, the higher the article impact and the increase in citations and author recognition! Therefore, even in an era of open access publishing and 'author pays' financial models, it is doubtful that journal editors will be persuaded to publish scientific research based on data-poor microorganisms.

Another important point is that the perceived value and job security of scientists are becoming more and more based on 'excellent' research that quickly generates short-term interest for members of the scientific community, journalists, politicians or the general public. So-called 'blue sky' and long-term research have become less compatible (and at the same time less financially attractive in the short term) than research that can quickly be patented to generate profit or high-impact publications. Research into data-poor microorganisms does not fit within this pattern of 'short-termism', especially in an era where research funding is under pressure and priorities have to be made with respect to global research agendas. Having said that, the author feels that there is still a case for research into data-poor microorganisms to be funded and to be published. Instead of simply taking a short-term view, we need to find ways to encourage microbiologists to share and study data-poor isolates, while at the same time encouraging scientific journals to publish the results obtained.

Finally, when we talk about the development and understanding of many different aspects of science and culture, we often refer to the past, for example, the development of vaccination, antibiotics, flight, fashion, music, the Internet etc., as occurring within various decades (e.g., within the 1950s, 60s, 70s and 80s) or even within various centuries (e.g., the 1700s, 1800s etc.). Although important scientific and cultural events in themselves, the fact that these events are now 'aged' (more than 10 years old), means that not being able to define an exact month or year when the events occurred does not necessarily significantly impact on the context or value of the development of the knowledge being described. Alternatively, the publication of 'epidemiologically poor' data lacking accurate chronological context (e.g., day, month etc.) does not significantly impact on the relationship between the past and current contextualization of the events being described, that is, the meaning of these developments to the current context of science, fashion, music etc., is not lost, but placed into a relative (non-short-term) context with respect to the greater range of time that has passed. Additionally, in a global context, the significance of regional data tends to become less significant when national or international data are especially relevant. For example, considering a microbiological example, involving global microbiological 'outbreak events', scientists are happy to publish their results based on national rather than regional findings. One example of this is the reporting of the NDM-1 resistance gene, where regional considerations were not considered important within the context of global events [4,5]. In these cases, the generation and publication of low-fidelity findings obtained from data-poor microorganisms may outweigh the need for epidemiologically rich, regional-based data.

Added value

So if data-poor isolates are to be dusted off and included in relevant microbiological research studies, what type of experiments should be performed and how should the data obtained be incorporated into modern day knowledge regarding the origin and spread of (pathogenic) microorganisms at the regional, national and global level? One of the most powerful

"Importantly, although it is tempting to think of the added value of data-poor microorganisms as being restricted to research and publications involving comparative genome sequencing, there actually exist many more possibilities for research involving data-poor microorganisms."

techniques currently in use in microbiology is WGS and this seems to be the most logical and useful place for utilizing data obtained from data-poor microbial specimens (see also whole-genome multi-locus sequence typing above). WGS facilitates the identification and mapping of core, accessory, virulence and antibiotic resistance genes, as well as facilitating research into the colonization, global spread and pathogenicity of bacterial clones. The use of WGS on data-poor microbial isolates could reveal unknown or unappreciated (but potentially pathogenic) genes that have ceased to circulate or are currently circulating 'under-the-radar' in microbial clones. In this respect, 'under-the-radar' means that the genes are present in microbial populations, but that they have not yet been discovered due to the limited number of WGS studies that have currently been performed on circulating global isolates. Also very important, the genetic data obtained from data-poor microorganisms can be incorporated into current WGS studies to generate more complete views of genetic shifts in microbial clones without having to link the isolates to a specific point of time. A detailed genetic map obtained from data-poor isolates and current microbial clones could indicate a loss or gain of genetic information between a data-poor isolate and a currently circulating strain, providing microbiologists with a targeted focus for future research. For example, WGS of data-poor isolates of *Escherichia coli* obtained approximately 30 years ago in Rotterdam, The Netherlands could be compared with a modern day *Escherichia coli* isolates obtained from Rotterdam in December 2017, this as a basis for investigating the loss or gain of antibiotic resistance genes within the Rotterdam area over the previous 30 years. The age difference between the data-poor microorganisms investigated neutralizes the perceived need for epidemiologically rich data in the contextualization of the results.

Importantly, although it is tempting to think of the added value of data-poor microorganisms as being restricted to research and publications involving comparative genome sequencing, there actually exist many more possibilities for research involving data-poor microorganisms. These include (to name only a few) – pathogenicity studies: investigating the pathogenicity of historically significant microorganisms [6], discovery of variation in (unknown) mobile genetic elements [7], discovery of variation in (unknown) virulence factors and plasmids [8], identification of new antibiotic/antiviral/antifungal/antiparasitic mechanisms and antibiotics [9], investigating antibiotic 'tolerance' [10], transformation, transduction, conjugation efficiencies [11]; diagnosis and treatment: development of new diagnostic tests and targets [12], discovery of previously unknown antibiotics [13], investigating microorganisms and cancer therapy [14], extended standardization of genotyping schemes [15]; gene regulation: investigating global gene regulation, for example, SOS and stringent responses [16], linking transcription factors to global physiology [17], reconstruction and prediction of transcription regulatory networks [18]; evolution: evolutionary ecological transitions [19]; and the immune system: differences in inflammatory responses [20].

The results obtained from this kind of research are actually data-independent in that all of the epidemiological data necessary to generate scientific value and publishable results are contained within the actual data-poor microorganism itself. Providing a context for the data provides an extra level of interest to the results, but is not by itself mandatory for generating added value.

"Scientists and scientific journal editors need to think about how collaborative agreements, financial support and research expertise can be best integrated into an international approach designed to make best use of these valuable 'once-lost-gone-forever' resources."

The solution

If we accept that data-poor microorganisms currently represent a large, but untapped, resource of underused yet potentially invaluable scientific data, then we should take the necessary steps to inventorize, analyze and even biobank these data-poor isolates (or at least biobank the data obtained from these data-poor microorganisms) for future generations [21]. This process will invariably require the integration of scientific collaboration, funding and technological (e.g., bioinformatics) expertise, which could eventually be obtained via multinational or national funding calls. However, other options do exist, including the generation and pooling of data-poor microorganisms/data between individual institutions, such that consortia of institutions can pool their resources, while at the same time sharing the burden of the overall costs of the research involved. Alternatively, data-poor microorganisms could be potentially interesting for pharmaceutical companies to investigate, and license agreements or material transfer agreements could be arranged on an individual institution basis.

Whichever (if any) of these collaborative solutions become available, it will be necessary to establish some form of data exchange that will publicize the availability of data-poor microorganisms and the willingness of (academic) laboratories to collaborate in sharing isolates, obtaining funding, processing data and publishing the results obtained. In this respect, the author has established a basic online website [1], in order to encourage contact details to be exchanged between institutions or individual scientists that are interested in publicizing their data-poor microorganisms or expertise.

Conclusion

Data-poor microorganisms are a precious resource of genetic and phenotypic microbial information that are currently being underutilized and are threatened with extinction. Scientists and scientific journal editors need to think about how collaborative agreements, financial support and research expertise can be best integrated into an international approach designed to make best use of these valuable 'once-lost-gone-forever' resources. Scientists and scientific journals should re-think the current emphasis on the short-term contextualization of research data and accept imprecisely defined data-poor contexts in exchange for the acquisition of long-term insights into microbial genomics, transcriptomics, metabolomics, virulence and microbial evolution. Publishing scientific data and the dissemination of information and hypotheses is the true driver of the progression of (microbiological) science. By denying the acceptance of data-poor science, we are literally throwing away a potential resource of new knowledge, insight and novel scientific development.

Financial & competing interests disclosure

This publication was supported by a Horizon2020 European Union grantunder grant agreement GA-633780 ("DIAGORAS" – www.diagoras.eu). The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1 Data Poor Microorganisms. www.datapoor.info

2 Ruppé E, Olearo F, Pires D *et al.* Clonal or not clonal? Investigating hospital outbreaks of KPC-producing *Klebsiella pneumoniae* with whole-genome sequencing. *Clin. Microbiol. Infect.* doi:10.1016/j.cmi.2017.01.015 (2017) (Epub ahead of print).

3 Angelo KM, Conrad AR, Saupe A *et al.* Multistate outbreak of *Listeria monocytogenes* infections linked to whole apples used in commercially produced, prepackaged caramel apples: United States, 2014–2015. *Epidemiol. Infect.* 145(5), 848–856 (2017).

4 Chen Y, Zhou Z, Jiang Y, Yu Y. Emergence of NDM-1-producing *Acinetobacter baumannii* in China. *J. Antimicrob. Chemother.* 66(6), 1255–1259 (2011).

5 Poirel L, Lagrutta E, Taylor P, Pham J, Nordmann P. Emergence of metallo-beta-lactamase NDM-1-producing multidrug-resistant *Escherichia coli* in Australia. *Antimicrob. Agents Chemother.* 54(11), 4914–4916 (2010).

6 Wagner DM, Klunk J, Harbeck M *et al.* *Yersinia pestis* and the plague of Justinian 541–543 AD: a genomic analysis. *Lancet Infect. Dis.* 14(4), 319–326 (2014).

7 Toleman MA, Bennett PM, Walsh TR. ISCR elements: novel gene-capturing systems of the 21st century? *Microbiol. Mol. Biol. Rev.* 70(2), 296–316 (2006).

8 Berger P, Knodler M, Forstner KU *et al.* The primary transcriptome of the *Escherichia coli* O104:H4 pAA plasmid and novel insights into its virulence gene expression and regulation. *Sci. Rep.* 6, 35307 (2016).

9 Dos Santos DF, Istvan P, Quirino BF, Kruger RH. Functional metagenomics as a tool for identification of new antibiotic resistance genes from natural environments. *Microb. Ecol.* 73(2), 479–491 (2017).

10 Levin-Reisman I, Ronin I, Gefen O, Braniss I, Shoresh N, Balaban NQ. Antibiotic tolerance facilitates the evolution of resistance. *Science* 355(6327), 826–830 (2017).

11 Yildirim S, Thompson MG, Jacobs AC, Zurawski DV, Kirkup BC. Evaluation of parameters for high efficiency transformation of *Acinetobacter baumannii*. *Sci. Rep.* 6, 22110 (2016).

12 Krahling V, Becker D, Rohde C *et al.* Development of an antibody capture ELISA using inactivated Ebola Zaire Makona virus. *Med. Microbiol. Immunol.* 205(2), 173–183 (2016).

13 Davis E, Sloan T, Aurelius K *et al.* Antibiotic discovery throughout the Small World Initiative: a molecular strategy to identify biosynthetic gene clusters involved in antagonistic activity. *Microbiologyopen* doi:10.1002/mbo3.435 (2017) (Epub ahead of print).

14 Felgner S, Kocijancic D, Frahm M, Weiss S. Bacteria in cancer therapy: renaissance of an old concept. *Int. J. Microbiol.* 8451728 (2016).

15 Chenal-Francisque V, Lopez J, Cantinelli T *et al.* Worldwide distribution of major clones of *Listeria monocytogenes*. *Emerg. Infect. Dis.* 17(6), 1110–1112 (2011).

16 Handel N, Hoeksema M, Freijo Mata M, Brul S, ter Kuile BH. Effects of stress, reactive oxygen species, and the SOS response on *de novo* acquisition of antibiotic resistance in

Escherichia coli. *Antimicrob. Agents Chemother.* 60(3), 1319–1327 (2015).

17 Berthoumieux S, de Jong H, Baptist G *et al.* Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.* 9, 634 (2013).

18 van Hijum SA, Medema MH, Kuipers OP. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol. Mol. Biol. Rev.* 73(3), 481–509 (2009).

19 Moran NA, Wernegreen JJ. Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.* 15(8), 321–326 (2000).

20 Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog.* 7(3), e1001307 (2011).

21 Kang B, Park J, Cho S *et al.* Current status, challenges, policies, and bioethics of biobanks. *Genomics Inform.* 11(4), 211–217 (2013).