Title

Opening up data from the massive citizen science project “Contagion! The BBC4 Pandemic”

Lead Applicant

Dr Petra Klepac

Details of proposal

Background Human interactions are crucial in shaping the spread of directly-transmitted infectious diseases. As a result, mathematical models commonly use data on human movement and social contact patterns to predict the dynamics of epidemics and potential control measures. Such data is typically obtained through self-reported surveys, or via anonymised, routinely collected mobile phone data. Both of these approaches currently have limitations in terms of scale, data availability and/or covariates relevant to infectious disease transmission. To mark the centenary of the 1918 Spanish Flu pandemic, the BBC commissioned a feature-length documentary on influenza epidemiology. A central part of this documentary was a custom-made phone app called BBC Pandemic, designed in collaboration with the team named in this proposal, which collected movement and social contact data from volunteers across the UK that could be incorporated into mathematical models of infectious diseases. This citizen science experiment records volunteers’ hourly locations to the nearest square kilometre over a 24-hour period, and self-reported contact-data over that 24-hour period. The preliminary data-set analysed for the documentary had data from 30,000 people. The app will continue to gather data until December 31, 2018 and the total number of people in the dataset already exceeds 50,000, creating the largest dataset of its kind. (i) Aims Increase the visibility and accessibility of this unique open-access data set, support its adoption by the research community and promote the principles of open science. Widen access to the data to a general audience and develop an understanding of the use of models to inform public health policy. This will be achieved by developing a website that in addition to general information about the project has four main parts: Spatial patterns of mobility - mobility patterns in different areas, and in different age groups can be extracted from the hourly locations (example kernels from preliminary dataset are shown in Figure 1). Social mixing patterns - from the self-reported contact data we can extract the age-dependent mixing patterns. Figure 2 shows the mixing patterns from the preliminary dataset for users who work in healthcare compared to those who are employed but not in healthcare. Interactive simulations of the model - users can vary input parameters by using sliders and selecting different aspects of data and see how that affects the outcomes of the models, and how that influences control and vaccination strategies. Dr Kucharski has been involved in creating an interactive educational website on a smaller scale (https://rozeggo.shinyapps.io/deploy/). Downloads - possibility to download data and code for the modelling work for this project. These include: (i) movement kernels, (ii) contact matrices, (iii) shapefiles we created for the spatial model, (iv) code. In addition to the website, the BBC Pandemic app will be made openly available alongside a high-impact open access publication summarising the main patterns in the dataset and their implications for modelling of infectious diseases. Target audience The project audience will include researchers working in social behaviour, infectious disease modelling, public health, economics, and other fields. The project will also target members of the public interested in social interactions and movements, as well as people who want to learn about infectious disease modelling and its impact on informing public health strategies. We also want the target audience to include groups such as schools, for which we will offer suitable project activities that can be published in Plus magazine, a magazine that brings the beauty and applications of mathematics to the general audience and has already written about our work (https://plus.maths.org/content/very-useful-pandemic). Finally, volunteers who participated in the study and used the BBC Pandemic app, or viewers of the BBC documentary, will be able to see the impact of their contributions. Activities Alongside the launch of the website, the team will promote the data and project via media coverage, talks at academic conferences, science festivals, and links with existing public engagement programmes run by the team members. We also hold a workshop to encourage further discussion and collaboration between interested researchers. (ii) Influencing open research practices Access to mobile phone GPS datasets, which are often proprietary, is currently extremely limited. This creates a barrier to entry for junior researchers, prevents reproducibility and limits potential for follow-on analysis. By making a large and unique movement dataset open and explorable, our project will overcome all these issues. In the process, we also aim for the project to become a leading example of how citizen science can contribute to high-profile research outputs. This study is the first of its kind, but we hope other studies will follow in other countries in future, and this example will also lead the way in showing best practice for sharing data as widely as possible. (iii) Monitoring, evaluation, and success indicators We will evaluate the project success via several quantitative metrics, including: number of dataset downloads; paper citations; social media activity; and google analytics to see how people interact with the website and where are they most engaged. We will also use qualitative indicators of success, such as enquiries and interest from other researchers, and follow on projects and collaborations either directly or indirectly inspired by our work.

Decision

Not shortlisted

Comment on decision from Wellcome

This was felt to be a high quality proposal to make a valuble data resource available and useable. However, the level of innovation as well as the potential impact of this proposal to transform health research through openness was considered limited.

Title **Data2Paper**
Lead Applicant **Mr Neil Jefferies**
Details of proposal 'Data2Paper' is a 'one-click' process to streamline the data paper publication workflow. Metadata about a dataset (together with a link to the dataset itself) are transferred from a data repository, via a cloud-based helper app which allows the addition of methodological detail, to a relevant publisher as a data paper submission. The app leverages DataCite DOI, ORCID and Scholix functionality to avoid data re-entry, saving time and avoiding errors. If accepted and published, subsequent details of the paper can be fed back to the repository to enrich the original record via ORCID and Scholix functionality. Data papers increase the opportunities for citation, improve the reproducibility of science through the dissemination of methodological information and also permit the release of negative results, thereby reducing repeat failures and increasing the efficiency of research resource utilization. Making the publication process as friction-free as possible incentivizes researchers to deposit their data in repositories and through data papers encourage and enable its sharing, verification and re-use. In addition, data papers themselves are first class research outputs that operationalise the FAIR Data Principles. With funding from Wellcome, our plan is to Pilot the the app with selected communties of repositories, researchers and journals/preprint respositories with a view to tracking at least 20 papers through the Data2paper workflow and then, subsequently, surveying researchers and publishers about the experiences: Bio-sciences: Bio-Studies portal, F1000 Research, Wellcome Earth Sciences: AGU/EGU, Earth Sciences Information Platform Data Journals: Data-in-Brief, GigaScience, Scientific Data Pre-print Repository: Zenodo Analyse and publish the feedback and correspondingly adapt the Data2Paper approach in terms of: Researcher efficiency: Time saved, improved accuracy of data entry, improved awareness of state of work-in-progress Researcher user experience: Areas for improvement, missing features, showstoppers, ideas for future development Research motivation: Does improved efficiency remove a disincentive to publish data papers, does the journal information on Open Access and APC's help inform open decisions, can this be improved? Use this information to help build long term operational, service, governance and business models for Data2paper - we are currently exploring a membership model, similar to that of ORCID. The project's scale is global as it works across national and disciplinary boundaries by joining multiple repositories and publication outlets via existing scholarly communication channels. The current Project Team and Steering Group are predominantly European but have strong links with the US institutions, networks and activities e.g. Research Data Alliance, Force11, Stanford University, Elsevier US, Metadata2020, Fedora Repository Platform and the COAR Next Generation Repositories Working Group. The anticipated impact on the research stakeholders is as follows: All proponents of FAIR Data will see growth in FAIR Data research objects. Researchers - will be able to achieve a citable output with the minimum of trouble. Also be able to demonstrate compliance with funder and institutional data deposition mandates. Funders - this service encourages better research data management. Researchers are more likely to engage with the repositories if they are likely to derive a citable research object at the cost of a few minutes' work. There would be additional metrics available, as well as better information about re-use. It should also encourage better data citation practices than are currently in evidence. Publishers - can secure a pipeline of (better quality) data paper submissions directly to journal submission systems. Higher Education Institutions - additional opportunity to demonstrate research impact and derive metrics. Repositories - improves their range of services, an opportunity to engage researchers not only to comply but also to engage with data management and deposition. ORCID - this represents an opportunity to enhance ORCID's value proposition by increasing its directly useful

function for both researchers and HEIs. Data2paper has an Advisory Panel with representatives of all the key stakeholder communities who monitor progress and provide guidance and direction when required. The panel comprises: Carole Goble, (University of Manchester), Wouter Haak, (Elsevier), Hollydawn Murray, (F1000 Research), Tim Smith, (Zenodo/Invenio/CERN), David Carr, (Wellcome), David Kaye (Jisc), Josh Brown (ORCID). Key factors for success: Integration with target repositories and journals completed At least 20 paper completed Data2paper workflow Demonstrated improvements in research efficiency and motivation sufficient to justify the service
Decision **Not shortlisted**
Comment on decision from Wellcome This proposal was to evaluate the effectiveness of the Data2Paper app. The level of ambition proposed was limited, and so the potential impact of this proposal to transform health research through openness was unclear.

Title

NiftyNet as an open and reproducible artificial intelligence research platform for real-time surgical and interventional support

Lead Applicant

Prof Tom Vercauteren

Details of proposal

Vision and aims Improving existing procedures or enabling new surgical and interventional applications through computer assistance typically requires real-time processing, analysis or even understanding of visual information captured intra-operatively by a variety of medical devices. Recently, a number of research groups have demonstrated the potential of deep learning to generate a quantum leap in our capacity to extract information from interventional imaging in real time. Yet, few have taken their research beyond the engineering proof of concept and into the clinic. The general field of artificial intelligence (AI) is booming and industrial-quality open-source software (e.g. TensorFlow https://tensorflow.org/) is now available. Tailoring such generic machine learning frameworks, validating the resulting trained algorithms, ensuring interoperability with clinical devices, and embedding AI modules in real-time application-specific pipelines and user interfaces nonetheless requires substantial efforts that few research group take up. This proposal targets computational researchers working in the field of computer-assisted intervention. Our audience can further be subdivided in: 1) those working primarily on the development of machine learning for visual understanding; and 2) those working primarily on the integration and translation of new applications. We will provide tools for both of these audiences to better focus on their main research question while benefiting from established validated platforms to streamline their developments while maximising interoperability with clinical devices, research reproducibility and reusability.   Activities To achieve our vision, our proposal will build upon and integrate two key software libraries: 1) NiftyNet (http://niftynet.io), for the core AI aspects; and 2) GIFT-Grab (https://github.com/gift-surg/GIFT-Grab) for the underpinning real-time data acquisition, processing and display pipelines. Both libraries require core developments and overarching integration work to generate robust real-time computer-assisted intervention pipelines. Below are a list of tasks we will be delivering. Each task entails software development work done under best software development practices concurrently with high-quality documentation as well as infrastructure and community support. Task 1. The majority of successful open-source software follow a "release early, release often" strategy to create a tight feedback loop between the core development team and the community. To initiate short release cycles from the onset, we will initially aim for a "low-hanging fruit" task. We will make use of the existing NiftyNet infrastructure, implement (when needed) and train established models for prototypical tasks which are often presented as open competitions/challenges (see e.g. https://endovis.grand-challenge.org/) in our community. We will share pre-trained models for these competitions and promote the use of NiftyNet as baseline within these competitions.   Task 2. In the meantime, we will develop a minimum viable product integrating GIFT-Grab and NiftyNet to demonstrate real-time inference on simple video sources. We envisage that this minimum viable product may not run in real time at full image definition and full framerate but will serve as a means of gathering feedback and keeping a high development pace. We will also setup the infrastructure to share validated and documented hardware configurations to support the developed pipelines. Task 3. To go beyond initial computational bottlenecks, we will extend GIFT-Grab to maximise its use of the graphical processing unit (GPU), for example by leveraging GPUDirect (https://developer.nvidia.com/gpudirectforvideo) for frame grabbing, by making use of inference optimiser such as TensorRT (https://developer.nvidia.com/tensorrt) and directly feeding into OpenGL display foregoing the need for CPU-GPU data transfers (a.k.a Graphics Interoperability). Task 4. To further trigger the interest of the community we will demonstrate the relevance of our approach to support innovative computational imaging devices and multi-stream inputs. We will collaborate with the development team of SUPRA (https://github.com/IFL-

CAMP/supra), an open GPU-enabled framework for low-level ultrasound management. We will embed SUPRA as a supported input sources for GIFT-Grab and thus provision for real-time AI that can take advantage of both low-level radio-frequency ultrasound data and post-processed ultrasound image streams. We will again share validated hardware configurations extending to any adjunct required for example for calibration purposes. Influencing open research practices NiftyNet is the first open-source deep learning platform dedicated to medical imaging and is led by our group. It extends TensorFlow for scalability and is starting to gain momentum in the diagnostic/pre-operative image analysis community. NiftyNet is currently the only consolidated effort to share AI models and gather a Model Zoo within the medical imaging community. This proposal will anchor this practice and extend it to the companion field of computer-assisted intervention. Evaluation criteria and success indicators Evaluating the success of an open-source software is notoriously complex. Beyond the somewhat simplistic quantitative measures such as the number of releases, the amount and frequency of code changes, the number of active issue reports and code merge requests, we will focus on indicators that we believe are important for this proposal, namely the number of new and active external code contributors and the number of research papers citing and using our software. A key success milestone will also be met if a first working system is demonstrated in July 2019 during the second international conference on Medical Imaging and Deep Learning (MIDL) which we are co-organising in London. Success indicators will further include our capacity to attract additional supporting development resources such as through donated hardware or sponsored student Google Summer of Code (GSOC, https://summerofcode.withgoogle.com) projects.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an ambitious proposal to develop an open source software solution. However, the potential impact of this proposal to transform health research through openness was unclear, and the evaluation plan would have benefited from more detail as well as more focus on user feedback.

Title **Development of the world's first comprehensive database and search tool for characterisation of bacterial surface polysaccharides**
Lead Applicant **Dr Johanna Kenyon**

Details of proposal

As we transition into a genomics era, whole-genome sequencing is becoming an accessible standard for analyses in research, clinical and public health laboratories. Some specialist tools and databases exist for microbiologists to identify clinically relevant features (i.e. antibiotic resistance genes) in bacterial genomes. However, a complete compendium of information for major bacterial virulence determinants does not exist, and access to reference strains reported in the literature is difficult, creating serious barriers for research. Some of the most critical bacterial virulence factors are cell surface polysaccharides. These include the capsule (CPS, K), lipooligosaccharide (LOS) with a carbohydrate outer-core (OC), and lipopolysaccharide (LPS) that is LOS with O-antigen (O) polysaccharide attached. Genes that direct their synthesis are generally encoded in loci that exhibit extensive variability between strains. However, polysaccharide nomenclature systems are unfortunately variable between species and are often non-descriptive to non-experts. Due to their importance in virulence and epidemiology, polysaccharides are targeted in strain typing, vaccines and phage therapeutics. Therefore, the ability to extract interpretable, actionable information from bacterial genomes via tools that do not require expert knowledge is critical. In response to this need, members of our research team developed 'Kaptive-Web' (kaptive.holtlab.net; github.com/kelwyres/Kaptive-Web); a web-based tool for rapid typing of K and O loci in genomic sequences from the clinically significant pathogen, Klebsiella pneumoniae (see Additional Information 1). While extremely useful for epidemiological typing, this tool does not provide access to reference strains or data on polysaccharide structures/biosynthesis that would enhance its value to researchers, clinicians and public health professionals. Use on other organisms is also limited to the command-line version, Kaptive, which requires some degree of computational expertise to download and host on a computer system. Thus, an opportunity exists to expand Kaptive to maximise access to all polysaccharide data arising from bioinformatics and laboratory research, and broaden application for use on other bacterial pathogens. Our vision is to create the most comprehensive database and search tool for bacterial surface polysaccharides to-date. This will combine datasets from multiple research disciplines that have never been presented together on the same platform, in a searchable format for non-experts. Initially, the capabilities of the tool would be expanded for use on Acinetobacter baumannii, the World Health Organisation's number one priority pathogen for therapeutics development. Polysaccharides produced by this species are currently being targeted for strain typing and novel vaccine/phage therapies highlighting an urgent need for improved data accessibility. Our team has shown that A. baumannii produces CPS and LOS, with more than 120 different CPS and 16 OC gene clusters identified and fully annotated with a transparent, easily interpretable nomenclature system. Our recent studies on A. baumannii CPS have uncovered biosynthesis pathways and carbohydrates not previously found in nature. We are also compiling the world's first bank of strains for representative polysaccharide types to eliminate barriers to future research (see Additional Information 2). However, a complete compendium of this data does not currently exist, and providing access to this information via an improved Kaptive-Web interface is the core focus of this proposal. The overarching aims are to: 1) Modify the existing Kaptive code and Kaptive-Web interface to incorporate additional features that maximise output; 2) Release all A. baumannii CPS and LOS datasets on the new Kaptive platform. Kaptive-Web was developed using web2py and utilises the command-line Kaptive script available at https://github.com/katholt/Kaptive. Kaptive-Web is currently hosted by the Australian NeCTAR cloud on a 16-core 64-GB RAM server that can run up to 15 simultaneous submissions. To expand on the functional capabilities of the tool, modifications must be made to both the command-line and web scripts using Python. This will occur in two stages. Stage 1: Integration of core functions including: i) the ability to optionally provide raw short-read sequence input rather than pre-assembled genome data; ii) the addition of output fields to facilitate access to the relevant published literature; iii) enhanced visual representation of results; iv) incorporating a search feature to enable identification of strains, loci, or other information. Stage 2: Addition of features that allow structural and biosynthesis data to be included with K and OC loci on the platform, broadening the depth of information available to users. This will first include uploading all K and OC loci sequences, fully annotated using transparent nomenclature. A relational database and user interface will then be created to interactively display the corresponding chemical structures and sugars, synthesis genes, encoded enzymes with known/predicted functions, and links to reference strains that we will source and deposit into the National Collection of Type Cultures UK. The resulting platform will, for the first time, enable users to upload raw sequence to rapidly receive a wealth of information about polysaccharide genetics, structures and biosynthesis, and access relevant strains. Such data are key to monitoring the influence of infection control efforts targeting CPS and/or LOS, and are essential for predicting antigen or phage binding variants to inform treatment decisions. We will evaluate the impact of this tool by assessing web page statistics and use by numerous teams investigating polysaccharides. Ultimately, Kaptive has the capacity to inspire action towards an integrative research initiative enabling the discovery, accessibility, and reusability of consolidated information.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal had good potential to impact health research. However, the level of innovation proposed was limited and the evaluation plan would have benefited from more detail.

Title

Automating dataset and topic discovery to optimise and accelerate health research development

Lead Applicant

Ms Karen Tingay

Details of proposal

A. Vision To provide the research community with a freely-available, open-access platform which promotes discoverability and reuse of published research outputs. Aims: To improve the discoverability of datasets and topics using text mining of academic publications, To improve the usability of the bibInsight platform to non-technical users To validate the accuracy and efficiency of our previously-developed data mining methods for use in developing health and social-sciences research using domain experts To provide a tool to support data discovery initiatives, such as persistent identifiers and RECORD, while these are becoming mainstream. Activities: We plan to expand on existing work using text-mining and machine learning methods applied to academic literature databases. The purpose is to develop tools for discovering appropriate data sources and emerging topic trends from user-defined literature searches. Existing work on which this project will build includes a functional but rudimentary platform to identify both survey and clinical data sources from abstracts, and to discover emerging topics.

Platform Our platform, bibInsight, can ingest data from a variety of sources such as PubMed, OVID, Scopus and others, to be analysed as one data source. The current means of data processing and interrogation include disambiguation of affiliations via the Global Research Identifier Database (GRID), discovery of emerging topics assisted by Medical Subject Headings (MeSH) terms, and relevance-ranking assisted by topic modelling. All this, however, currently exists in a form usable only by researchers with a solid computer science background. Part of this project will focus on making this functionality widely available through a readily-usable online analysis platform that we will develop. Dataset discovery This is achieved using machine learning techniques by training a classifier to recognise the context within which potential datasets might be mentioned in a given abstract. This technique has already enabled one researcher to discover 30 new cohorts in two hours, while it has taken at least two researchers to discover 150 in a five year period. We would like to tailor these methods to health- and social science-specific contexts and automate them as much as possible without losing relevance.

Identification of emerging topics This is achieved by analysing trends in the use of the standardised MeSH terms, assisted by metadata about the terms themselves. This allows us to discriminate between increased usage of a term due to recent introduction, or due to greater research interest. Our aim is to expand these methods to predict emerging topic trends as early as possible. Target audiences: Researchers, research data controllers, and cohort/data owners. For researchers, fast and accurate access to information held within academic publications will improve reusability of methods and data, and accelerate research development. For research data controllers, given the lengthy process of acquiring datasets, proactively approaching owners of relevant datasets may speed up the process. Knowing what topics are emerging may help them to provide more accurate advice for researchers looking for an appropriate dataset. For cohort and data owners, the ability for researchers to find their data, and use it in a variety of applications, is vital in securing further funding, and essential to ensuring that the utility of their research is maximised. B. Influence on open research practices Our current focus is on helping researchers make sense of unstructured information held in publications. We anticipate that this work will impact on the following three aspects of research practice Increased visibility of research outputs: Impact on this area will be achieved via the use of existing infrastructure in a more efficient way. Our current techniques make it easier to extract accurate information from unstructured sources such as academic papers. However, these techniques are not yet publicly available. By developing this platform, we will provide a readily-available means for researchers to discover relevant datasets and topics even if these have not achieved a high profile.

Accelerated research. Impact in this area will be achieved via a reduction of the effort required at the early stages of a research project, measured in person-months. Reduced cost to the research institute. Impact in this area is achieved via the use of publicly available datasets which can be interrogated in a unified way using specialised freely-available analysis techniques. This is in sharp contrast to existing tools which come with high licensing costs for research teams and institutions and, by extension, research funders. C. Evaluation We will use domain experts to evaluate both topic and data discovery analyses as applied to health and social science contexts. This will be achieved through a straight comparison of the relevance of results obtained respectively via the automatic route and manual effort as assessed by domain experts. Domain experts will also be used to ensure that our methods and evaluation criteria are accurate for the subject areas, e.g. alternative spellings, abbreviations, context-specific disambiguation (i.e. support referring to physical or emotional) Efficiency of the tools will be compared against current practices involving human effort, most commonly measured in person-months.

Researchers will be engaged in both the development and evaluation of the methods through workshops that aim not only to inform researchers about the tools, but also to inform the developers of these tools about specific researcher requirements.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to extend and increase the accessibility of the bibInsight platform, which would increase discoverability of datasets. However, the level of innovation proposed was considered limited.

Title

Opening Psychology to the Individual

Lead Applicant

Dr Chris Noone

Details of proposal

Our Vision The Opening Psychology to the Individual project will create a novel and valuable online resource for Health Psychologists hosted by the Center for Open Science (COS). The resource will facilitate the use of N-of-1 methods to test theories of health behaviour at the individual-level at which they were designed to operate, rather than the group-level at which they are currently tested. N-of-1 studies involve repeated measures within the individual over time as the unit of measurement to allow hypotheses focusing on individual behaviour to be tested. Calls for greater use of methods which focus on individual-level behaviour have been made for more than 10 years (Davidson et al., 2016; Molenaar, 2004). Notably, N-of-1 methods are recommended by both the Medical Research Council and the American Medical Association as a vital tool in the development and testing of complex interventions (Craig et al., 2008). Increasing the use of N-of-1 methods in Health Psychology is fundamental to the establishment of an accurate evidence base for health behaviour theories. There are both mathematical and empirical reasons for this. The mathematical issue relates to the assumption of ergodicity. For pooling data across individuals to give an accurate depiction of individual-level behaviour, the data must be ergodic - meaning that all statistical characteristics must be equivalent at both within-individual and between-individual levels (Molenaar & Campbell, 2009). Kievit and colleagues (2013) cite an accessible example of how difficult it is to satisfy this assumption in psychology – if measures of IQ were ergodic, then all individuals would have IQ lower than 100 for half the time since IQ measured at the group-level has a median of 100. The empirical basis for the need to test theories of individual behaviour using individual-level data comes from a recent study which, for the first time, quantified the lack of generalisability from group-level data to the individual (Fisher, Medaglia, & Jeronimus, 2018). This study showed that, across 6 samples where intensive repeated-measures data were collected, the variance of estimates of the relationships between psychological constructs and behaviours were 2 to 4 times bigger for within-individual estimates than for between-individual estimates – a clear violation of ergodicity. In plain language, the correlations found in the group data cannot be applied to explain any particular individuals' behaviour. This finding undermines the use of group-level data to test theories of health behaviour which aim to predict individual behaviour. Despite the recognised value of N-of-1 methods in Health Psychology, these methods are rarely implemented in practice. Open Science provides a framework to up-skill researchers in new methods and promote their uptake by providing access to transparent protocols, open workflows, data sharing and open-source statistical packages. Our Aims This project has five aims: Develop an online resource through the COS including user-friendly protocols and repository for N-of-1 data and tools to facilitate the design, implementation, analysis and synthesis of N-of-1 studies for testing theories of health behaviour (see the appendix) Develop an exemplar protocol for a collaborative project focusing on theoretically-derived predictors of medication adherence Describe the resource in a journal article Demonstrate the resource through workshops at two international health psychology conferences Evaluate the usability, feasibility and acceptability of the resource in a qualitative study with health psychology researchers Influence of the Opening Psychology to the Individual project on open research in Health Psychology Though there have been welcome developments in the promotion of open science in Health Psychology, including the new journal Health Psychology Bulletin, open research practices are far from routine in the field (Peters et al., 2017). We aim to demonstrate the value of open research practices in Health Psychology by showing how they can be used to facilitate new ways of doing research – such as N-of-1 studies which test theories of health behaviour. Why this Project is Important for Health Continuing to neglect the disconnect between what health behaviour theories

describe and how they are tested will at best have no effect on public health and hamstring the potential of health psychology to accurately describe and understand health behaviours – at worst, it could lead to misinterpretations of data which could lead to harmful interventions or failures to intervene. These problems can be largely avoided by using N-of-1 methods to test theories of health behaviour. Continuing to neglect open research practices will maintain the research practices which led to the reproducibility crisis and thus hinder the translation of accurate findings from Health Psychology into practice which can create positive societal impact. Through this project both open research practices and N-of-1 methods will be promoted within Health Psychology. This is a desirable and synergistic combination. Monitoring and Evaluation
A qualitative study will be conducted with the initial health psychology users to explore usability, feasibility and acceptability of the resource. Telephone interviews will be conducted by the post-doctoral researchers under the supervision of Dr. Mc Sharry an experienced qualitative researcher and following the COnsolidated criteria for REporting Qualitative research (COREQ) checklist. Telephone interviews will be transcribed and analysed to identify barriers and facilitators to the use of the resource, and to suggest potential improvements or changed needed.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal had good potential to impact health research in psychology. However, the level of innovation proposed was limited and it was not clear which parts of the proposal would significantly advance open research.

Title

Deeper Insights from Primary Data (DIPD)

Lead Applicant

Dr Nicholas Ilott

Details of proposal

Our vision is to change the culture in science to promote community access to primary microbiome sequencing data that through collaborative endeavour will enhance biological insight and accelerate the path from data to health benefit. In contrast to the current paradigm for conducting scientific research, the proposed model consists of the following route from data to outputs (Fig. 1): Primary microbiome sequencing data (metagenomics, metatranscriptomics, 16S rRNA sequencing) are uploaded to MGnify (https://www.ebi.ac.uk/metagenomics/) which contains a data view specifically for DIPD projects (Dr Robert Finn). These data are analysed utilising MGnify resources (pipelines, databases etc.) for immediate contextualisation among existing data. The design of the study and data that are generated are published as a Data Note (Wellcome Trust Open Research or equivalent for non-Wellcome Trust funded researchers). Peer review at this stage constitutes the first checkpoint for re-evaluation of the data. The publication also creates a discoverable set of data and acknowledges contributions of individuals involved in study design and data generation.

Interested parties visit MGnify and download the data. At this stage these parties are notified that the data are part of DIPD and encouraged to contact the data submitter for collaboration. Data are analysed according to collaborators' interests and reports are generated and shared amongst collaborators. Collaborators agree on final analyses and a paper is produced. A manuscript is submitted to an open access journal with detailed descriptions of author contributions. The current scientific publishing paradigm often promotes exclusivity, suboptimal use of resources and a long lag between data generation and scientific output. Our model aims to fast-track the process of data generation to research outputs whilst simultaneously increasing scientific rigour, broadening scientific insight, strengthening collaboration and improving cost-effectiveness. Increasing scientific rigour: Collaboration between multiple teams from raw to processed data will draw on unique individual experiences of data quality control that will facilitate the identification of issues that may have gone unnoticed by any individual group. Broadening scientific insight: Collaborative research as outlined brings the potential to draw on diverse interests within a field and maximise new knowledge. Strengthening collaboration: The DIPD model aims to provide a platform for networks to form in the microbiome community that may not otherwise have been considered. This will provide more diversity in microbiome collaborations in terms of interests and geography. Improving cost-effectiveness: The DIPD model utilises a "publish early, publish often" philosophy that is akin to the "release early, release often" philosophy in software development. The ability to garner important feedback from peer review early in the study through publication of a Data Note will save time and resources should there be an overlooked flaw in study design. Further savings are expected due to a reduction in redundant data generation and increasing data reused through meta analysis and enlarging cohorts to increase statistical rigour. Leading by example: A pilot project for DIPD To pilot the initiative we will generate a rich microbiome dataset of broad interest. We will generate a combined metagenomic/metatranscriptomic dataset from stool samples in patients with Primary Sclerosing Cholangitis-associated Inflammatory Bowel Disease, Ulcerative Colitis and healthy controls (n=20/group, Dr Alessandra Geremia). We will also collect a rich set of metadata that includes clinical, demographic and dietary information. These primary data will be made searchable within the MGnify resource, with the rich metadata ensuring that these data are highly discoverable and usable. Indeed, part of the initiative will focus on capturing and visualising metadata within MGnify and flagging missing but desirable fields. Availability of the data to potential collaborators will be publicised through a Data Note, publication of an article describing the DIPD initiative (Wellcome Trust Open Research) and via social media. Publicising the model and data resource

We will publicise the model to the microbiome community in the following ways (as well as publications listed in the Leading by example section): Approach Institutions involved in microbiome research and hold seminars to explain the rationale for the proposal. Establish a DIPD twitter account (@DIPDannouncements) that will tweet when primary datasets are uploaded to MGnify/DIPD. Metrics for assessing impact/success of initiative: We have developed a survey (https://goo.gl/yJUCvs) to assess current data sharing practices (summaries at https://goo.gl/8Fx637). Preliminary results are very encouraging for the DIPD initiative. While there exist perceived barriers to uploading data (Fig. 2a), 53% of respondents would upload their data at the start of a project (Fig. 2b), with 68% believing that this would accelerate research (Fig. 2c). These data support an initiative to kick-start changes to data sharing practices. Success of the DIPD initiative will be reflected in a change in attitude around data sharing. We will utilise metrics that are collected as standard at MGnify including number of data downloads and number of DIPD projects. A comparison to non-DIPD public projects will inform us of how visible and usable DIPD data are. We will further gather information on the willingness of people to collaborate on primary projects (pilot study), assessing number of collaborators and demographics. A successful project will be reflected in multiple international collaborators taking the project to completion. Testimonials from collaborators will be sought to assess perceived benefit and challenges of DIPD.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The proposal described a potentially useful resource to generate a collaborative IBD network. However, the level of innovation proposed was considered limited and it was unclear why a new dataset needs to be generated for the resource. The application would have benefited from including a framework to deal with retroactively available datasets.

Title

dait-c - Data analysis in the cloud.

Lead Applicant

Mr Barry Rowlingson

Details of proposal

In our group we are often developing statistical methods for disease modelling that have immediate real-world applications. Getting the outputs from those models to the people who supply the data and who want to make decisions based on the outputs has been a difficult hurdle to jump. For example, in monitoring Loaloa incidence in areas of lymphatic filariasis (LF) in Africa, we currently receive new data periodically by email, extract it, clean it, run it through our modelling software and finally send back map images for the LF treatment programme decision makers. Several other projects in our group work in a similar manner, including: those investigating the effect of malaria control measures in Malawi; identifying hotspots from real-time surveillance monitoring of health data from Public Health England; producing daily updated maps of ongoing infectious disease outbreaks such as foot-and-mouth in cattle or ebola virus in humans; and producing sampling locations for mosquito insecticide resistance in a sequential sampling framework. Information on these projects is available on the CHICAS website [http://chicas.lancaster-university.uk/projects/]. Some of these projects have already built ad-hoc solutions to the problem of doing regular analysis of updating data and producing new results. These solutions include emailing of data sets, shared online cloud storage (such as Dropbox), development of custom web interfaces, and using local FTP or SCP repositories for file transfer. These solutions all have drawbacks in terms of security, usability and reproducibility, as well as being reinventions of the same basic pattern. This pattern is Data analysis as a service (Daaas). One party is responsible for collecting data which is then processed through an analysis pipeline developed by a second party, and the results are fed back to a third party. We aim to produce Dait-c, a working Daaas platform, as an open-source system which can be installed by anyone on compatible hardware. The target audience for our work includes applied scientists and developers of methodology for those sciences, including statisticians and data scientists. Outputs from production-grade projects run on Dait-c systems can provide outputs for decision-making non-technical staff. Dait-c will use a lightweight web interface to a cloud computing resource (which could be private or public) to make the analysis accessible to anyone with the cloud computing budget, paying only for what they use. This can be an attractive model in low and middle-income country (LMIC) settings since it does not require capital outlay on data centres and hardware. A modular design means that a particular Dait-c installation can have a range of possible data analysis methods available. Each method is implemented by an analysis module, which is code written to conform to the Dait-c analysis module specification and loaded into a Dait-c installation. Analysis modules could do something as simple as fit a straight line through data points in an uploaded file, or as complex as running multiple statistical simulations over a cloud of compute nodes. In both cases the user experience is the same - upload data, set some parameters, then run the analysis module. The user, and anyone else subscribed to notifications for that project, will receive status messages from the project. For example when an analysis run finishes a project statistician might want to check the model fit by interpreting a model diagnostics page. If there is a problem, the statistician might change some parameters and re-start the run. The data supplier will get notified of this. A built-in system similar to a chat room or forum will facilitate discussion of the process until everyone is agreed on the validity of the results. We will run a demonstration instance of Dait-c and allow access to people interested in the system and its operation. We will document the system, run a workshop, and produce tutorials for developing compatible custom analysis modules. Having a demonstration system running will be our main success indicator. The progress along the path to the demonstration system will consist of milestones in the development and implementation phase to be decided in the initial stages of the project. As a fully open-source platform we will build a community of researchers, programmers, and users, and will accept contributions of code, documentation, and other components to the core of the platform under an open-source license. We hope that developers of additional analysis modules will release their code under open-source licenses. People can then search for and freely re-use modules in their own scientific workflow. This involvement in the project from the wider community represents an additional set of success indicators. We would also consider as a success indicator if we find that other people or groups have installed and run their own private or public instance of Dait-c. These indicators will demonstrate the potential for the use of the system in real-world applications and the possibility of building a sustainable development infrastructure. The attached additional information contains wireframe mockup pages for Dait-c - these are illustrations of possible pages for the system.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting and innovative proposal. However, there were concerns over the level of feasibility in practice, and whether user uptake could be successfully achieved.

Title

Resource Watch: Monitoring the Scientific Food Chain

Lead Applicant

Dr Monique Surles-Zeigler

Details of proposal

Although science rests on chains of evidence and the ability to reproduce the results of experiments, to date, we did not have a system in place that lets us do for scientific tools what we can do for food, i.e., track the use of ingredients through the food chain. Research resources or product codes include resources like antibodies, organisms such as genetically modified mice, cell lines, and software tools. However, practitioners of science know that these resources are not perfect; antibodies may be non-specific, cell lines contaminated, knockouts incomplete and software tools to give faulty results. Such issues are reported in the literature, but due to inadequate identification and dissemination practices, discredited resources such as contaminated cell lines continue to be used, well after their contamination is uncovered (Neimark, 2013). Readers of these papers have no easy way of knowing that the results they are reading are based on a discredited cell line. For instance, we know that reports of poor performance do not always have the required impact, which is to caution future users of those resources and to alert readers. Consequently, as has been documented, hundreds of papers continued to be published with discredited cell lines (Neimark, 2013). Over the last couple of years, tools have been designed to address this the issue of inadequate resource identification through the RRID (Research Resource Identifier) Initiative, a standardized and machine-friendly system for identifying research resources within published papers (Bandrowski et al., 2015). This effort was imperative due to various reports documenting what many scientists already knew; scientists do not report enough detail about the resources they enable proper identification (Vasilevsky et al.2013). RRIDs provide each antibody/organism/cell line with a persistent and globally unique identifier and ask authors to add this to their paper, to assist with the goal of reproducibility. Current information about research resource performance is scattered across millions of articles and websites, which limits and prevents access to this vital information. There is only one resource type, cell lines, which has an organization that tracks reports of problem cell lines and updates a central list. However, for other resources, no centralized list exists. Specialized centers will often conduct validation tests and post recommendations for their domain (e.g., ENCODE for Cancer Research). Researchers may also mention such information in passing within an article. In all, there are thousands of papers in PubMed alone that report problems like "nonspecific antibody" which need to be evaluated. What is clear is that we need a better means to discover whether a problem has been reported with a given resource, and a means to disseminate that information during the publication process and across the places where scientists are reading articles that use it. Towards that end, we propose Resource Watch. Resource Watch is a database and service that will aggregate these reports of problems with research resources, tie them to their RRID's, and provide updated information to authors, reviewers, editors, and readers through a web-based user interface. Initial efforts will focus on identifying resources within the Neuroscience domain, concentrating on neuroscience-related open access journals. Any paper that contains RRIDs or with sufficient data about a resource to assign an RRID will display a product code report if this information is available. Integration of Resource Watch data into the RRID information portal will serve the data to authors. Integration into various tools including an independent of publisher web annotation tool will enable reviewers, editors and other readers of the paper to have the same set of information presented to the author and act accordingly. This study will be completed with the following two aims: Aim 1 – To build a database and associated user interface that will store the product Research Resource Identifier (RRID) code and the associated notes of concern. A SQL relational database will be used to store all data, including but not limited to RRID's, paper citation and notes of concern. This database will dynamically interact with the user-interface to allow immediate access to the user. Aim 2 – To use data mining techniques to identify PubMed papers, RRIDs, and external databases (e.g., ENCODE)with reports about the performance of a neuroscience-related resource. Resource Watch will interact with PubMed API, RRID and Scibot technologies to identify papers of interest from PubMed Central. Additionally, we anticipate the need to create a set of structured tags for the notes of concern to allow users to quickly determine the severity of these notes. Regarding outcomes, Resource Watch is expected to give a reason to adopt RRIDs and the Open Science mandates for transparency. The performance of a research resource will directly impact the validity and therefore integrity and cost of science. The success of the tool will be reviewed by the number of clicks(on the note of concern) per month. Resource Watch will for the first time be a means to monitor the performance of neuroscience-related research resources and disseminated information in a form which can be directly acted upon by researchers at the time of the study, reviewers at the time of publication and readers at the time of consumption.

Decision

Not shortlisted

Comment on decision from Wellcome

The was an interesting proposal, which would have value if taken up by users. However, application would have benefitted from more information about how the team would monitor, evaluate and disseminate the resource.

Title

Increasing open data and code by providing technical support to researchers

Lead Applicant

Prof Adrian Barnett

Details of proposal

BACKGROUND Sharing code and data greatly increases value of research, because it strengthens the reproducibility of the results, allows readers to better understand what methods were used, and allows other researchers to investigate new questions using already collected data. Increasing data and code sharing is therefore a relatively simple way to increase the return on the huge current investment in health and medical research. It also has public support, and a recent survey of clinical trial patients found that the great majority (82 to 93%) supported their data being shared with other researchers [2]. Data and code sharing is not part of routine practice for many health and medical researchers, and most journals do not require or request sharing. There are multiple barriers that prevent researchers from sharing, including: political and legal barriers; commercialisation and IP; a concern about negative career impacts; a lack of reward for taking the time to share data and code; and a lack of knowledge and resources on how best to share [3]. Strong policies from journals and funders that aimed to increase data sharing have been partially successful. The BMJ have been a stand-out journal for encouraging data sharing through strong policies, but our research group were only able to get 4.5% of data sets after contacting researchers [4]. Similarly for a funder with a policy of public access to collected data, researchers were only able to access 26% of datasets from published papers [5]. These researchers suggested that, "funding agencies [...] provide data-sharing tools and technical support to awardees". AIM Our aim is to test whether providing technical support to researchers will increase data and code sharing. We will approach researchers immediately after they post a health-related paper to a preprint server, and ask them to share their data and code, and offer them e-mail support from staff members with expertise and experience in data and code sharing. We will only include papers that make no mention of sharing data and code. Preprint servers are an ideal place to recruit, because researchers are still actively engaged in the data, code and drafting. So we believe they will be more likely to build-in sharing, whereas once final publication has occurred, they may have moved on. Using preprint servers will reduce our generalisability, as researchers posting to preprint servers will likely be different to the wider population of researchers. However, preprint servers are becoming more widely used. We have a team that can provide the support we believe researchers are most likely to need. Based on our experience and consulting with QUT research librarians, we anticipate common questions from researchers will include:

Policy and permissions. Whether sharing conflicts with their funders' or institutions' policy on data sharing; whether they have the necessary ethical clearance to share their data.

Time and technical issues. What sharing platforms are best to use, how long sharing will take, how much sharing will cost them in time and money. Harms and privacy. What the potential drawbacks are; whether it will harm their prospect of publishing the results; how the data can be scrambled/anonymised prior to sharing. The great range of potential questions is why we believe dedicated experts are needed to help researchers with sharing. We will recruit 100 researchers and offer them support (see Figure 1). We will exclude papers that do not contain original data (e.g., meta-analyses, editorials, etc). The primary outcome will be whether these papers shared their data and code with the final published paper. We will compare each paper to three control preprints, where the authors did not receive our help. The cost of including controls is low, as we simply need to check their final data and code sharing. This will give us a 91% power to detect an increase in the percent of sharing from 5% to 15%. How we will influence open practices Our key potential to influence practice arises from investigating a policy change that is practical and is currently being considered by institutions and has already been implemented in some. We will discuss the study design with institutions and funders before finalising the protocol, in order to incorporate their ideas and increase their interest in the study’s results. We are currently conducting a randomised trial with the journal BMJ Open to investigate whether badges increases data sharing by providing a reward to researchers. However, to our knowledge, no study has investigated the barrier of providing technical assistance. Therefore we have the ability to influence practice by providing novel insight to an important question using a rigorous study design. How we will monitor and evaluate our proposal, including success indicators. A success indicator will be the difference in the rate of papers that shared their data and/or code after receiving our technical assistance, compared with the rate amongst papers that were not contacted. We will use an ordinal outcome of sharing of: 1) unrestricted sharing, 2) restricted sharing (e.g., conditional on further ethics applications), 3) no sharing. Another success indicator will be the interest we receive from institutions and funders looking to implement this policy.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was from a strong team and the methodology was robust and clearly described. However, the level of innovation proposed was limited, as was the potential for wider impact of the results.

Title

Mobistudy: a multi-centre, multi-study mobile-health research platform

Lead Applicant

Dr Carmelo Velardo

Details of proposal

Recently, many mobile-health applications have been developed for clinical and research purposes. Even when data are used for medical research, a per-institution or per-study approach is usually pursued. This causes fragmentation, duplication, unintentional variation in governance and ethical approaches and wastes resources. Research studies can only be conducted through the involvement of technically able researchers. Designed solutions are usually bespoke and not open to reuse, modification, or extension.   Personal health record systems like HealthKit, GoogleFit or Samsung Health allow the secure storage of health-related data in smart mobile devices with the consent of the user, and to share those data with others. This makes it possible to reuse apps in different settings and for different studies (e.g. a fitness app can be used for both weight loss and rehabilitation). Medical researchers are still unable to access these rich data in a simple way. Researchers need a platform that can exploit the potential of personal health records to facilitate the design of mobile health research studies. This platform combines those functionalities that are most common between mobile-health applications and is therefore reusable in different studies (by the same or different institutions). The main objective of this proposal is to build and test an open platform consisting of a smartphone app and website that will streamline the design of research studies involving smart mobile devices. Mobistudy will have a large target audience that will include clinical researchers as well as engineers. Our main goal will be to make the platform independent of coding experience so that, even for those with limited technology literacy, it will be possible to design a mobile-health intervention. We will rely on generic, extensible modules that will result in reusable components that researchers will use to design and initiate mobile-health studies. Our open-source system will accommodate most common requirements of mobile-health studies like a well-structured dynamic informed consent interface, validated patient reported questionnaires, forms for self-reporting of symptoms, medications, and other clinical events. We anticipate that the proposed platform will strongly contribute to: Providing better access to mobile-health research for those researchers or institution who find it difficult or cannot afford technical collaborators   Easing data collection from and engagement with large patient populations   Creating an active community of researchers that will extend the open-source code base Helping promote experiment repeatability and reproducibility by providing a standard platform to design mobile-health interventions.   Our platform will benefit academics in a number of ways. Firstly, by providing an easy-to-use, accessible way of designing mobile health intervention. Secondly, by having a standard to follow for the implementation of mobile-health interventions, that will make data capture, data sharing, and methodology sharing easier. Thirdly, Mobistudy will help participants' engagement and recruitment, allowing for quicker and more affordable clinical trials. The project will be evaluated together with clinical collaborators. Two pilot studies have already been planned, one that will integrate Mobistudy with RUDY (www.rudystudy.org), an online platform for people with rare musculoskeletal diseases, and one that will help the Oxford Biobank Study (www.oxfordbiobank.org.uk/) to collect questionnaires from patients participating in their long-term research into common diseases like diabetes, obesity and cardiovascular disease. Our team have already started the scoping work with the team that developed RUDY to understand the key aspects that we can generalise into Mobistudy. Regular meetings with the clinical team and other stakeholders (e.g. RUDY's PPI group) are planned so that advancements in the development of Mobistudy can be monitored and disseminated. An open-source GIT repository has been designed to keep track of progress, a specification document has been produced and is kept up-to-date as the project progresses towards its final milestone. The methodology that will be followed – underpinned by a patient-centred approach – will guarantee that the developed platform will be usable, engaging and thus highly suitable for long-term use. The dissemination of the research outputs to the academic community will involve publications in high-quality peer-reviewed biomedical engineering and medical informatics journals (e.g. Journal of Medical Internet Research) as well as high-impact clinical journals (e.g. BMJ Open). Programme stakeholders, including participants and healthcare professionals, will be regularly engaged at meetings, co-design workshops, and by means of a quarterly newsletter. A website will be used to coordinate this effort in a centralised manner. The open-source nature of the project allows scrutiny by public and research peers. It will provide the opportunity for inter-institutional collaborations, and the chance to develop a community of interested open-source developers which, in turn, could expand the functionalities and capabilities of the platform itself (e.g. by providing new scenarios or new integrations with existing platforms). We have already attracted funds from NIHR from both the BRC and the Oxford CLAHRC and we now seek to accelerate our development with further funds from the Wellcome Trust Open Fund. We have ongoing, informal discussions with other academic institutions at the forefront of mobile-health (Cornell University, Stanford University) and industrial partners (Apple, Google). We believe that the time has come for a platform made for clinical researchers rather than technologists, and that the Wellcome Trust Open Fund is the ideal environment to further implement and test the Mobistudy platform.

Decision
Not shortlisted

Comment on decision from Wellcome
The applicant opted not to share this information

Title

Development of a Minimum Information for Self-Monitoring Experiments reporting guideline (MISME)

Lead Applicant

Dr Guillermo Lopez Campos

Details of proposal

The reduction in costs of new individual digital sensors and wearable technologies (e.g. fitbit, apple watch) has popularised their use for different purposes including research or disease management. The use of these devices to monitor a broad range of activities or parameters is known as self-monitoring, self-tracking or self-quantification. The use of these activities in research is framed across different disciplines and areas as digital health, participatory medicine and citizen science. The continuously growing market of apps, devices and solutions available to researchers and consumers offers different solutions and different data and measurements for many different parameters. However, often the use of proprietary data platforms associated with the self-monitoring devices or apps make it difficult to retrieve or share the generated data. This results in the publication of aggregated data rather than the raw data, which would be useful for future meta-analyses, and often in an insufficient amount of metadata associated with the experiment in key aspects that could facilitate its reproducibility. For these reasons, we consider necessary to develop new practices oriented towards getting a greater adherence to FAIR principles in this area that will engage with the relevant stakeholders, namely digital health researchers, biomedical informaticians, community and public health agencies, industry (including consumer health tech developers), clinicians and self-experimenters. In this project, we will focus on refining the initial formulation and developing of a 'Minimum Information about a Self-Monitoring Experiment (MISME)' reporting guideline, testing it and disseminating this reporting guideline across relevant stakeholders. For these purposes, the specific aims of this proposal will be: Collection of new use cases for the refinement of the formulation of MISME Development of MISME elaborating the descriptors required in the guideline, soliciting feedback on the elaborated drafts and testing its application. Dissemination of the MISME concept across the different stakeholders. For these purposes we will develop a series of activities that will target these three aims and the different stakeholders to disseminate the concept of MISME and its underlying elements (an anticipated timeline is available in the additional information materials). Development of an online survey to identify and collect new use cases and needs of the different stakeholders that could benefit from the use and development of MISME. Organising open workshops where we will collect data from members of the general public (self-experimenters) as they engage with biomedical informaticians about self-experimenting and MISME. Organise open workshops for the digital health and informatics research community where we will communicate the concept of MISME to potential users and collaborators for the advancement and improvement of MISME. Map MISME into the Investigation, Study, Assay (ISA) framework. Dissemination across the biomedical informatics community organising a panels at relevant conferences inviting to participate in them representatives from the different stakeholders Organise a meeting with the editors in chief of relevant biomedical informatics open access journals (JAMIA Open, JMIR, MIM Open) to discuss the use of MISME in their journals. Development of a pilot project to report mHealth and exposome-related data and projects in a MISME compliant format. In contrast with other related initiatives in this area, focused in the development of data sharing standards, MISME extends its focus to the development of a reporting guideline for the description of the relevant metadata associated with the self-monitoring processes that makes use of other tools and approaches such as ontologies for experimental annotation. The potential benefits associated with the development and implementation of MISME expand beyond the BMI or digital health research communities to other areas and disciplines such as exposome research. With this development we aim not only to improve the reproducibility of the research in this area promoting FAIR principles, as a key initiative to bring greater evidence-based rigor to informatics-based research in this area, but also to develop a tool that would enable new crowdsourcing approaches providing an annotation framework for the contribution to research from citizen science approaches. Finally, this idea could potentially be applied in clinical environments either for the secondary use of medically collected data or even to facilitate the uptake of these data for medical purposes. The expected outcomes and success indicators from this proposal will include: Definition of new use cases leading to an improved and updated version of MISME Development of supporting documentation for MISME describing the aspects covered in this reporting guideline and its potential applications Pilot the use of MISME to describe self-monitoring experiments under different circumstances. The success indicators for this project are The design, running and analysis of the online survey to capture and analyse new use cases and user needs. Publication of MISME supporting documents. The successful organisation of at least one of each of the proposed workshops and the elaboration of a report of the outcomes and conclusions gathered about the use of MISME An analysis of the current use of standards and FAIR principles in self-monitoring research as a literature review. Two examples of how MISME is applied in two different examples of self-monitoring in mHealth or Exposome research. Submission of at least one panel proposal to an international conference in biomedical informatics discussing the potential application of MISME in different areas.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a clear proposal for development of a reporting standard, which could hold significant potential value in the field of wearables. However, there were concerns over its feasibility, and in particular the ability of the team to secure wider community buy-in and uptake.

Title **Open data and notebooks for studying laboratory animal activity: Removing barriers to routine sharing of data**
Lead Applicant **Dr Laurence Brown**
Details of proposal The physiology and behaviour of all life on Earth is tailored to the daily rotation of the planet and the resulting 24h changes in light and temperature. These circadian rhythms (from Latin circa-diem, approximately a day) have profound implications for health, and the disruption of sleep and circadian biology is associated with a wide range of disorders. As a result, studying patterns of activity and rest over multiple days (actigraphy) is an essential approach for the study of circadian rhythms in model organisms such as laboratory mice. The vision for this proposal is based on solving an important but simple problem. Without transparent and open data to accompany publications in circadian biology, our field risks missing out on the additional insights gained by deposition of data. Without such data, papers are limited to the interpretation of the data by the lab upon publication. The majority of studies into the effects of particular genes or environmental conditions are expertly carried out but are individual experiments on small numbers of animals. Routinely, other researchers cannot access the raw data on which research conclusions are made. Moreover, when new improved analytical methods are developed, studies must be replicated. Finally, data cannot be aggregated, making combining data or meta-analysis impossible. The benefits of data deposition are highlighted by the field of transcriptomics. Here the effort that has gone into every published experiment featuring microarrays or similar technologies (since deposition with methodology was agreed in 2001) has the potential for reuse and citation. A good example is our recent meta-analysis paper that revealed novel transcripts enriched in the circadian master pacemaker (Brown, et al. Nucleic Acids Research 2017) made possible with public data. We believe that circadian biology will similarly benefit from data deposition, but we need to identify and address the barriers to this becoming the norm and provide data for others to explore. We recently produced an affordable and efficient system for the collection and analysis of activity and sleep data in mice, made open at all levels. Uptake of the system has been good in our group in the UK and worldwide, including major International phenotyping centres such as MRC Harwell. This system (COMPASS: Continuous Open Mouse Phenotyping of Activity and Sleep Status), is not the only method of collecting information on activity and sleep in rodents, but provides the only open research tool of this nature. Basing discussions about open science and data deposition around an open system ensures that there is no barrier to researchers engaging with this research approach. Communication with the scientific community. The research communities for circadian biology in the UK (UK Clock Club), and worldwide (the Society for Research in Biological Rhythms, https://srbr.org/) and the European Biological Rhythms Society, https://www.ebrs-online.org/) are well-established and inclusion of interested parties would start via these bodies. An anonymous survey (with ethical approval obtained before the work commences) distributed via the mailing list for the UK Clock Club will be used at first to establish to opinions and potential barriers to routine publication of data, as well as the range of data types and formats and feedback on possible meta-data requirements. This will also offer a chance to see where existing data formats may prevent open practices in the future. As with many laboratory measurements, actigraphy has relied on a wide range of different equipment, often hand-crafted and unique to each research group. Conversely, commercial options are often not open to scrutiny (or improvement) and require proprietary software to analyse. The possible use of other professional bodies and review/option articles, plus blogging and social media will also be explored to generate discussion and widen participation in the survey. Leading by example. Alongside defining the contents of the survey and the options for distribution we will begin a detailed review of the data available within the SCNi in Oxford and at MRC Harwell as part of the IMPC screens for sleep and circadian activity phenotypes. By the end of the project we will have

made as much of this data as possible public, with a consistent set of meta-data and guidance to speed up deposition of future data generated at both sites and elsewhere. An estimation based on recent capacity in both centres suggest this will be a minimum of hundreds of individual records from COMPASS in the first instance (and potentially thousands of activity files involving running wheels). Communication with bodies such as fairsharing.org and existing tools (e.g. the Biodare project, https://biodare2.ed.ac.uk) will ensure data is as usable as possible and under permissive licences. Using new data to highlight teaching resources. As data becomes available it will be possible to demonstrate how open data changes a number of behaviours. Providing open data resources as part of an interactive teaching platform is one way we will highlight its utility. Using interactive notebooks a web-based textbook with data to explain practicalities and concepts in analysing actigraphy will be a new teaching resource for circadian biologists, as well as explaining handling time-series data in general. Online reporting (blogging) of the process (akin to https://openlabnotebooks.org/) will also help shape these tools. (Additional details attached)

Decision

Not shortlisted

Comment on decision from Wellcome

This application proposed a consultation on open data and open lab notebooks, which would be of value to those study laboratory animal activity. However there was no detail provided on how how the outcomes and impact of the tools built following this consultation would be evaluated.

Title

Wellcome OpenEvidence: Full acquisition, curation, dissemination of extracted trial data from Cochrane reviews

Lead Applicant

Prof Clive Adams

Details of proposal

1. Vision a. This proposal This proposal will harness the crowd-sourced efforts of production of populated XML files and create from these a widely accessible data set of trial data. These data, created from pre-analyses pre-publication RevMan XML files will be parsed, curated and automatically formatted for: feeding directly back into the Cochrane reviews to improve clarity; and for cloud-based storage, querying and downloading Data, extracted and re-tabulated from the published papers, are not subject to copyright but are currently trapped in the pre-publication RevMan XML files. The Cochrane Collaboration recognise that these are public domain data and are committed to making them more accessible (Addendum—http://bit.ly/2pbNC2J) - and are encouraged to do so by funders (UK's NIHR) - but progress has been slow - hence our call to accelerate this vision (https://doi.org/10.1136/bmj.k3229). Our development of proof-of-concept open source software (RAPTOR, https://goo.gl/gnxB75) illustrates how the vision of making these data widely available is possible. b. Overall This is creation of the definitive repository of reliably extracted healthcare trial data to be available for re-analysis, novel use and live reviewing. Delays in the production of systematic reviews undermine use of evidence in clinical care (https://doi.org/10.1371/journal.pone.0003684). Tailoring of evidence to personal values is not possible. Currently, auto-data extraction from PDF (also pioneered by this group https://doi.org/10.1186/s13643-016-0207-7) can only be machine-assisted – but linking to large crowd-sourced maintained data sets enables future data extraction to become swift. The data repository can be queried by researchers for novel work. Such a 'live' repository would also be available for simple queries by an end-user interface designed to assist question setting for clinical problems - allowing prioritising of outcomes of interest to the reader, automatic analyses and writing of the review in any chosen language [auto-analysis has existed for some time; automatic writing in multiple languages has been pioneered by this group https://doi.org/10.1186/s13643-017-0421-y]. Analyses of frequency and type of public querying of the data set would help prioritise efforts for full complex reviews. 2. Impact This project will: greatly reduce waste in systematic reviewing, democratise access to and use of data from trials, accelerate the pace of best evidence to bedside. 3. Aims This project will: refine through software development and wide testing our existing proof-of-concept software to extract trial data from XML tables and files created by volunteer reviewers; create more systems to curate and improve these data; seed those improved data back into the XML tables and files to improve the source review; create a cloud-based repository of trial data; populate that repository; make a public interface for querying the repository; and publicise this work. 4. Target audiences Reviewers – aiming to reduce workload by providing high-grade already extracted study data, and a system by which new data can be uploaded and automatically improved within their review. Researchers – those wishing to undertake new reviews, or novel work on trial data, will have reliably extracted data instantly available Funders – by having an expanding repository of data funders will be able to estimate what efforts are needed – or not needed – when calling for reviewing projects The public – who should be able to query the data for 'live' evidence to their tailored, prioritised, clinical questions (Additional file - Figure 1). 5. Activities These will be programming and testing. High-grade programming will refine our existing software but this will run in parallel with testing across a selection of Cochrane groups [testing file upload, curation and formatting], and with researchers [testing the public GUI]. Testing across different Cochrane groups will ensure generalisability of the output across health care. 6. Influence open research practices This is an audacious project with broad vision but we know from experience that progress along this road is taken

incrementally. Members of this research group pioneered auto-data extraction from PDF (https://doi.org/10.1186/s13643-016-0207-7), public access to full trial data with each piece of data traceable to source within the PDF (https://doi.org/10.13140/RG.2.2.28907.95529), automatic drafting of reviews in multiple languages (https://doi.org/10.1186/s13643-017-0421-y) and automatic seeding of Wikipedia pages with good evidence (https://doi.org/10.1186/s13643-017-0607-3). Members of this group have been instrumental in drawing up of the Vienna Principles of automated systematic reviewing and the progress of the International Collaboration for the Automation of Systematic Reviews (https://doi.org/10.1186/s13643-018-0740-7). This project, creating the Wellcome open repository of trial data, is one further step in the evolution of fully open trial data and their wide use. 7. Monitoring and evaluation This project comes out of the experience of Cochrane Schizophrenia and must initially show that influence. Even if that were to remain the case the thousands of trials from which data have been meticulously extracted over 25 years will be liberated for wide use off the ‘Cloud’. Monitoring of accrual of trial data will be continuous and success indicators will not only be this accrual but also diversity of accrual from other non-schizophrenia sources. Within months of creation of a Wellcome repository of trial data online download activity is not likely to be great but can be monitored as indicator of future activity. Further success indicators will be in the qualitative and quantitative testing activity from end-users and other Cochrane groups.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting and ambitious proposal in an important area, with the potential for impact. However, it's relationship to existing resources and initiatives was unclear and the evaluation plan would have benefited from more detail.

Title

AttachmentOpen: An online interactive tool for accessing meta-analytic attachment research

Lead Applicant

Dr Marije Verhage

Details of proposal

The overall goal of our proposal is to develop an online tool for meta-analytic data sharing in developmental science and provide solid, easily accessible conclusions for researchers, practitioners who work with families, and policy staff based on all evidence currently available in the field. While meta-analytic data may, on first sight, already appear to be open, this is not true. Meta-analysts' code and retrieved data are beyond what is reported in published manuscripts. The costs of not sharing data, and the benefits of doing so, have recently extended to meta-analytic data in a BMJ paper calling for sharing of full meta-analytic data of Cochrane reviews for re-use. Researchers in the field of language acquisition have led the way with "Metalab" (http://metalab.stanford.edu/index.html), a tool for community-augmented meta-analysis in their field. The "Metalab" tool elegantly demonstrates that standardized collections of meta-analyses could help diagnose issues and outstanding questions in a research field. In the field of attachment research, meta-analysis has been used for decades to test claims on outcomes and precursors of attachment, interventions in attachment, and intergenerational transmission of attachment. This research has convincingly demonstrated that secure parent-child attachment relationships are a key factor for long-term well-being and mental health. This knowledge has had wide-reaching consequences for everyday family life as well as help for families, practices of child-care centers, and family court decisions. Now it is vital to keep the results up to date, and also to make the analyses more accessible and usable for the community of researchers and professionals. To achieve this, we aim to apply and extend the basic functionality of Metalab to the field of attachment and develop an online interactive tool, "AttachmentOpen", adapted to our field's modeling requirements (e.g., categorical data, field-specific variables). Our vision is that AttachmentOpen will serve as the open resource for researchers and mental health professionals with the information they require to set up new studies, provide an evidence base for intervention strategies, and inform court decisions. We envision that AttachmentOpen will serve as an example of meta-analytic data sharing and evidence-based medicine for other fields of study. Specifically, AttachmentOpen will: provide a repository of all existing meta-analyses in the developmental attachment field, which will serve as a knowledge base for the research community, practitioners, intervention developers, and policy staff

update and extend existing data continuously with results of more recent original studies by building a community of curators who are in charge of individual data-sets

make meta-analytic data more visible and more easily accessible by providing interactive features that enable users to easily modify parameters, zoom in on specific populations, compare effects sizes between different groups, and conduct power analyses for funding applications, but also to provide scientific underpinning for interventions and policies include a registration module for authors who intend to write articles based on specific analyses within AttachmentOpen, which will simplify collaboration and prevent the duplication of work

include an analysis module (based on R package 'metafor') to carry out new meta-analyses. To use this module, authors will have to pre-register their research questions and hypotheses within AttachmentOpen. Extracted metadata and analysis scripts will be saved in a repository to prevent it from ending up in the "file-drawer", but instead provide an opportunity for using these data in new analyses, especially when new studies keep being added.

Within the grant period we aim to: Complete the AttachmentOpen tool. A beta version of the basic functionality is currently being built by a scientific programmer of the Research-IT department of our faculty, available Autumn 2018. Make meta-analytic data of 38 existing meta-analyses available that have been conducted in the field of developmental attachment research in the tool (see Additional information). As 20 of these meta-analyses were performed by our project team, we consider this feasible. Update existing meta-analyses with data of more recent studies using a catalog established by our collaborator Sheri Madigan, which contains fully coded sample and study-level information for all studies with a measure of child attachment. Devise guidelines for curatorship of included meta-analyses. Devise guidelines for authorship of new research-papers using AttachmentOpen. Advertise the use of AttachmentOpen at two important conferences (March/July 2019) in the field and on Twitter and Researchgate. Make the R-base code used for AttachmentOpen freely available and publish information on the development and decision-making process in a design paper in Attachment & Human Development. This project will be successful if by the end of the project period: all proposed functions of AttachmentOpen are implemented 90% of the 38 meta-analyses are included in AttachmentOpen at least 100 computations have been made in AttachmentOpen and at least 2 registrations of new planned articles (use monitoring from the launch in July 2019 onwards) user-satisfaction of the tool (measured by Customer Satisfaction Score and Net Promoter Score) is over 80% In line with the vision of reaching solid answers to critical questions through collaborative science, AttachmentOpen will change meta-analysis from a laborious project to a one-click calculation. Researchers and mental health professionals gain easy access to the information they require to set up new studies, choose between intervention strategies, and inform guidelines and policy.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting, important and feasible proposal with a strong evaluation plan. However, the potential impact was unclear due to the narrowness of the research field and the level of methodological innovation proposed was limited.

Title

Developing In Silico Tools To Evaluate Combination Compounds in Pre-Clinical Models

Lead Applicant

Dr Jaine Blayney

Details of proposal

BACKGROUND In the laboratory setting, drug-screening assays have proved invaluable in identifying novel compound applications [i]. To identify synergistic (versus antagonistic or additive) drug combinations one can either employ a hypothesis-driven approach or unbiased pairwise combinatorial screening assays (two compounds per well). The former can miss unanticipated interactions between targets, whereas the latter can be too unwieldy for most academic labs. Multiplex screening has emerged as an alternative approach, having been applied in the identification of drug pairs for HIV [ii]. Using 1,000 compounds, this method used combinations of 10 drugs per well, reducing the number of wells from 499,500 (all pairs) to a manageable ~13,100. Although this approach reduces the number of experiments, isolating key compound interactions becomes more complex. PROPOSED DEVELOPMENT To facilitate not-for-profit/academic biomedical research into novel patient treatments, the two key aspects of our in silico framework are therefore: i) Screen Study Design (SSD): determining an optimum solution to cover all compound pairs across a minimum number of wells, and ii) Screen Deconvolution Analysis (SDA): deconvolving intensity signals from each well to extract combination pairs (or triplets etc) that drive cell viability. We will draw on software/hardware engineering for solutions. The SSD complexity increases with the number of compounds [iii], similar to all-pairs testing in software design [iv]. Mathematical solutions can either be exact (global optimum) [v], or relaxed, in which repeats of pairs are tolerated (estimate of global optimum) [vi]. The latter is more suitable for scaling to larger numbers of compounds. The SDA stage of extracting interacting compound signals is comparable with blind source signal separation, which has been used in cell-type stratification of heterogeneous tumour samples [vii]. Using a relaxed greedy algorithm with iterative local optimal solutions [viii], we have developed an in silico prototype to carry out preliminary work in paediatric acute myeloid leukaemia (AML) (see attached). Using 80 novel apoptosis-inducing compounds (3,160 possible pairs), a minimum solution of 168 wells of ten compounds was achieved (SSD). A branch and bound linear regression subset analysis of combinations across three cell lines at three time points (SDA), ranked ABT-737, a Bcl-2 inhibitor, in combination with Purvalanol A, a CDK inhibitor, in the MV4-11 cell line highly, agreeing with human curation. However while regression methods can include statistical interactions and adjustments, these may not mirror biological/chemical processes. Furthermore human assessment, 'by-eye', may omit more complex patterns. We propose to further develop our prototype, with application to a 300+ multiplex compound screen in AML. Algorithmic Development (Months 1-8) SSD: Evaluation of all-pairs testing algorithmic solutions eg greedy algorithm variations etc; SSD/SDA: Assessment of parallelisation solutions; Calibration of algorithmic/parallelisation solutions with respect to scale (max 1,000 compounds); SDA: Comparison of human decision-making with blind source separation approaches eg non-negative matrix factorisation to extract key combinations. Software Development (Months 3-8): Analysis and design/user testing; graphical user interface; error handling; establishment of web-host (email registration/sign-in); optional forum for users to post screening summaries. Screening Design/Data Analysis/Evaluation (Months 9-11): SSD: Optimise design of 300+ compound screen; SDA: Analysis of multiplex screen. In Vitro Testing (Months 9-11): Testing of individual and combination compound intensities across three AML cell lines at three time points; in vitro validation of in silico predictions. Dissemination (Month 12): Evaluation/preparation of manuscript/s; Release of in silico framework via QUB-hosted server. Locally, the results of the in silico framework will be assessed by the optimum solutions produced (SSD), scalability to larger number of compounds and computational speed. The recovery of compound combination signals (SDA) will also be assessed, with a focus on identifying/validating known/novel interactions.

Globally, we will measure our reach through publications/citations and the number of registered users accessing our site, downloading code and uploading summaries of their own combination screens to encourage collaboration. The online framework will be free-to-access for academic/non-profit researchers. Once registered, users can access the online tool or download the open-source code for use/adaptation locally. The code will be developed in the statistical programming language R [https://www.r-project.org/], written in a parallel framework, to facilitate execution on a user's multi-core desktop/laptop. To enable adaptation/development by other users our software will be designed along object-oriented principles, using the unified modelling language. Code will be fully annotated. Code-sharing platforms such as GitHub and SourceForge can be used to collaborate with active developers. A graphical user interface will minimise a (less technically-skilled) user's need to run complex scripts. Example datasets and 'walk-throughs' will be used to illustrate how the software can be used.   [i] R Macarron, MN Banks, D Bojanic, et al, Nat Rev Drug Discov, 10:188–195. doi: 10.1038/nrd3368, 2011; [ii] X Tan, L Hu, LJ Luquette III, et al. Nature Biotechnology, 30:1125, 2012; [iii] Y Lei and KC Tai, Proceedings of the Third IEEE International High-Assurance Systems Engineering Symposium, 254–261, 1998; [iv] J Czerwonka, Microsoft Corporation, 2008, https://msdn.microsoft.com/en-us/library/cc150619.aspx; [v] R Mandl, Communications of the ACM, 28:1054–1058, 1985; [vi] K Tatsumi, Proceedings of the International Conference on Quality Control, Tokyo, 615–620, 1987; [vii] D Repsilber, S Kern, A Telaar, et al, BMC Bioinformatics, 11:27, https://doi.org/10.1186/1471-2105-11-27, 2010 [viii] TM Chan, SIAM Journal on Computing, 39 (5): 2075–2089, doi:10.1137/08071990x, 2010.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The applicant opted not to share this information

Title

Enhancing reproducibility and validation of automated spike sorting workflows

Lead Applicant

Dr Matthias Hennig

Details of proposal

(i) Vision and aims Recently developed high-density complementary metal-oxide-semiconductor (CMOS) microelectrode arrays allow extracellular recording of activity from thousands of neurons with high precision. While this offers unprecedented opportunities for understanding brain function in health and disease, major challenges arise from the substantial data volume and complexity. In particular, inferring the activity of single neurons from raw recorded signal, a blind source separation process called spike sorting, poses major conceptual and computational challenges. High-quality, scalable spike sorting software is essential for accurately inferring the interactions of neural circuitry in large-scale extracellular recordings. It is well known that mistakes and bias in spike sorting can lead to erroneous conclusions and that these known issues will only be exacerbated by increasing the volume of recorded extracellular data. Currently, there is no consistent methodology to assess spike sorting performance and to perform automated post-hoc assessment of sorted extracellular data. The complexity of this process, often based on lab-specific, handcrafted scripts, complicates reproducible analysis even when code is available. These issues were discussed at a workshop during the 2017 CNS Conference which was attended by many groups developing modern spike sorting solutions. There was broad agreement that established standards and methods are necessary and should be developed collaboratively, leveraging substantial community expertise. This project proposal is a direct result of these discussions and has the following aims: 1. The development of software for spike sorting validation and quality control which will implement both existing and novel community-accepted metrics and algorithms. The key features will be ease of use, extendibility, easy inclusion into existing workflows and interoperability. We will make use of, and contribute to, existing projects such as the Neo library for data conversion, the NBF format by Neurodata Without Borders consortium, and the phy library for visualisation. 2. The development of a framework to track the provenance of extracellular data sets and enable reproducibility of analysis workflows. Provenance refers to the lineage or history of the data, including the processing steps it went through and the parameters and implementation details of the algorithms that were used on it. We will implement this for quality control metrics which will help to assess the required level of provenance necessary for a complete workflow. Our provenance capture technology will use the NoWorkflow system and be documented using the W3C PROV standard which has already been successfully used for the Neuroimaging Data Model (NIDM). In the longer term, we aim to extend our software for complete analysis workflows. 3. The organisation of a workshop to provide a platform to discuss and prioritise community needs and to agree on standards for spike sorting. This will bring together groups that develop methodology and groups that primarily run experiments (the named collaborators would also attend). This event will include training sessions and workshops for widely used software and also a session for the introduction and discussion of novel ideas and methods related to spike sorting. 4. The collection and dissemination of information to facilitate the use of available open source tools and validation data sets generated by project collaborators and other groups. This would make up a large portion of the project's online presence. The end users of this project will be labs using modern electrophysiological recording technology and cutting edge analysis software for high-density MEA data. With the increasing availability of published data sets, our project will also facilitate re-analysis and further interpretation of existing data, a vital step in solving the reproducibility crisis in neuroscience. (ii) Open research practices This proposal is part of a growing effort to transform neuroscience research into open, community-driven science, with increased availability of data, software and pre-prints of key publications. The reproducibility and transferability of data sets, however, requires careful provenance, a problem that is well researched in computer science and has been

successfully solved in other domains. Supported by modern tools developed by the global open source community, this project will focus on the achievable goal of moving towards full provenance of the algorithms and parameters used for data processing, adding much needed transparency to the field. Development of standards, project documents and software will be based on GitHub, a platform that enables consistent version tracking, transparent, democratic decisions, and fair attribution of all work done by contributors. (iii) Monitoring and success indicators The two key success indicators are community adoption of the developed toolkit, and a solid community engagement to broaden its scope beyond the lifetime of the grant. We aim for collaborative development from the start, with regular interactions among collaborators to coordinate and prioritise activities. The workshop will be held in June 2019, where we plan to debut the first release of our planned software framework to the community so that we can solicit feedback and encourage contributions. At this stage, we also aim to present a roadmap for provenance, with a first demonstration. By the end of the project, a fully functional library for quality control will be released which includes data provenance. This will form the basis to support complete workflows which we aim to demonstrate by the end of the grant.
Decision **Funded**
Comment on decision from Wellcome The application was from a strong team, proposing to generate important software for the standardisation of spike sorting. The application showed a strong commitment to advancing openness and had the potential to greatly impact the international community of neuroscientists.

Title

Open platform for cross disease network analysis with Specific Betweenness

Lead Applicant

Prof Francisco Pinto

Details of proposal

We have recently published a network analysis algorithm to predict proteins associated in common with two diseases (García-Vaquero ML, Gama-Carvalho M, De Las Rivas J, Pinto FR. (2018) Searching the overlap between network modules with specific betweenness (S2B) and its application to cross-disease analysis. Sci Rep, 8(1):11555). This algorithm takes advantage of the current knowledge about proteins related with two diseases (separately), and expands it with candidate proteins associated with the two diseases simultaneously through their relative location in protein interaction networks. S2B relies on the assumption that interactors more commonly found on shortest paths linking proteins encoded by genes associated to two diseases must appear in the disease modules overlap. To identify and rank these proteins, S2B employs a specific version of betweenness centrality, which measures how many times a node is involved in a shortest path, focusing specifically on shortest paths linking proteins associated with the two diseases. In this publication, we showed through simulated data and a motor-neuron diseases study-case that S2B candidates were enriched in proteins simultaneously associated with the two diseases. The S2B algorithm was implemented as an R package, freely available at https://github.com/frpinto/S2B. This format is adequate for computational biologists but not for other biomedical scientists. Even for computational biologists, besides the R package, lists of proteins associated genes and networks of protein-protein interaction are needed to perform the S2B analysis. This proposal aims to build the Cross-disease network analysis web portal, where users can 1) perform S2B analysis to predict proteins associated simultaneously with two diseases, and 2) query a database of S2B candidate proteins resulting from previous analyses. The portal will be public and provide a user-friendly interface for the S2B algorithm. This will allow biomedical scientists and clinical researchers to easily apply our cross-disease analysis to their diseases of interest. To reach our aims, we have organised this project into 5 tasks. Adapt S2B algorithm to a web tool: The first step will be to adapt the S2B algorithm to optimize its performance as a web-tool. This work will be co-supervised by João André Carriço, capitalising on his experience in developing open web applications (http://www.patlas.site; http://www.phyloviz.net/). Link web tool with other public resources for input import and for a richer annotation of results: For a greater usability of our tool, we will implement a simple input form were users can select the diseases to study, the databases from where to import disease associated genes and the source of interaction networks. For the latter, it will be possible to filter for interaction type and tissue specificity. A single disease mode of analysis will be developed, were a first run of analysis will provide a list of other diseases with greater overlaps with the input disease. We will also develop a visualisation interface to explore the network of interactions between S2B candidates and known disease associated genes. To facilitate the interpretability of the output network, we will associate links to annotation databases characterising the interactions and the nodes. Set up database to collect results from individual analysis: Besides the S2B web tool, the cross-disease network analysis portal will also have a companion database, accumulating the lists of S2B candidates generated from individual analysis. This database will avoid repetition of the same analysis by different users. It will also provide a way to compare results of S2B analysis for larger groups of diseases. Produce documentation and online tutorials with usage examples: A very good tool can fail if the users do not know how to use it or do it in a wrong way. To avoid this, we will strongly invest in producing a clear and complete documentation, complemented with several tutorial examples that users can follow online. Plan tutorial training sessions and prepare associated materials: With the dual function of advertising our cross-disease network analysis portal and training users on how to use it, we will develop plans for tutorial training sessions and produce companion material for the attending students. Training sessions will be performed on our home institutions (BioISI/FCUL and IMM/FML) and at some collaborator institutions (Univ. Salamanca (Spain), Aachen University (Germany)). In parallel, we will propose to present our tool at international conferences and, if possible to offer this training sessions as satellite events. To evaluate the success of our project we will: Perform questionnaires after tutorial training sessions   Monitor the number of people attending training sessions   Monitor the number of analysis performed by users in the S2B tool   Monitor the number of visits/queries on the S2B results database   Our tool potentiates the knowledge extraction from data that is already public. Therefore, it maximizes the benefit that can be taken from the investment that produced those public datasets. It will also foster the use of cross-disease approaches to better understand molecular pathophysiology. Researchers can compare their results focused on one disease, or on a phenotype that is common in two or more diseases, with our cross-disease predictions. Overlaps will increase the relevance of their experimental approaches and guide them to formulate new hypothesis about disease mechanisms or therapy.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a feasible proposal outlining the development of a tool that could potentially add value to the feild. However, there was little evidence about the level of user demand meaning the impact of the proposal was unclear.

Title

Reproducible and freely-available clustering of bacterial sequences for genomic epidemiology and surveillance

Lead Applicant

Dr John Lees

Details of proposal

Overview Our recent genomic epidemiology clustering tool PopPUNK (see additional information) showed excellent results on ten bacterial species-wide datasets, offering improvements in quality and speed over current approaches. The underlying data structure is unique to our method, allowing clusters to be expanded and remain consistent when new datasets are added in. These enhancements give PopPUNK potentially very broad applications to bacterial disease surveillance: outbreak detection, determining patterns of global dissemination, and defining the distribution of phenotypes such as antimicrobial resistance with respect to population structure. We propose to add to our tool in four ways: improving its speed; curating and maintaining clustering databases for four pathogens; creating interfaces suitable for a range of potential users; maintaining a browser-based deployment available to all researchers. 1) Methodological changes Creating a finalised version of our method is crucial for reproducibility. We will first consider significant performance enhancements, before choosing the stabilised version of the method used for later steps: We will add the use of Bloom filters so that raw sequencing reads can be used as input rather than sequence assemblies. We will increase speed through either the use of the newly developed kmer-db software (doi: 10.1093/bioinformatics/bty610), or large-scale parallelism using GPUs. 2) Creation and distribution of curated databases The use of existing large, high-quality collections of genomes greatly enhances the inferences possible from cluster assignment. Accurate knowledge of the prevalence of resistance or the locations from which a cluster is drawn may have implications for local management of a potential outbreak. Most datasets are generated by research institutes, with little integration into public health research. We will leverage our familiarity with these datasets to integrate these databases into our tool, so any researcher can compare their samples against large, high-quality samples of the existing population. Clustering databases can expand and improve as new genomes are added, but a lack of sharing between studies prevents this in practise. To allow for continuous updating while maintaining reproducibility of results we will use a blockchain-like approach. A public record of current and historical cluster assignments will be automatically distributed to users of the tool; when genomes are added, a new version of the database will be added to the chain. Users will be able to use either the latest database or a stable version which has been curated to remove false positives. Uptake of our approach would result in: The same clustering database, underlying model and specific implementation being used by a wide range of research groups. Cluster names would have a definition that is consistent between studies, retaining one of the key advantages of typing schemes, but using a more sensitive and scalable underlying method. We will focus our database curation efforts on Streptococcus pneumoniae, Streptococcus agalactiae, Klebseilla pneumoniae, Staphylococcus aureus and Escherichia coli, four of which are WHO priority pathogens. We are including collaborators with expertise and available large datasets for each these species. 3) New interfaces It is important to provide our tools so they can be used by those running national and regional surveillance, especially when less local bioinformatics support is available. Creating an interface that doesn’t require specialist hardware or pre-existing bioinformatics expertise will expand our user-base to those who generate data and are closest to outbreak management. This extension also provides the opportunity to standardise analysis steps for all users and share outputs with other researchers. For bioinformaticians, we will create a standardised interface through R as a RECON package (https://www.repidemicsconsortium.org/). For general users we will use Javascript to create a browser-based interface to our method. Users will drag-and-drop raw sequencing data onto the page. It will be possible to include existing large global datasets, which will provide more context and relevant data for small collections. Results will be presented as an interactive visualisation (with a permanent link) allowing dissemination of any analysis performed. The specific command will be available to the user so that they can accurately report their methods. 4) Stable and maintained deployment This proposal will further enhance reproducibility of results by maintaining a deployment of the web-interface to the tool for the next five years. We will utilise a cloud service so that anyone can perform bacterial genomic epidemiology using our consistent, fully documented and open source method. This will provide extra stability for researchers without access to high performance computing facilities. Evaluation The underlying method has already been implemented and extensively tested. This proposal has four clear activities to extend the software's functionality, decentralisation and reproducibility. Completion of these activities first measure of success. A further aim is increasing uptake and audience of the method. We will be able to monitor this through the web service use and standalone package downloads. We will create and distribute a user survey; direct feedback will be possible through GitHub and the creation of a Slack channel. Whether databases continue to be expanded and improved by other users will be a clear indication of uptake, both overall and by species. We have included collaborators who perform nationwide public health surveillance and local clinical surveillance: they will evaluate the capability of our new method to replace their existing approaches.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was from a strong team and had good potential to impact health research in bacterial disease surveillance. However, the level of innovation proposed was limited.

Title

S3X: a novel paradigm for Sustainable Scientific Software publication

Lead Applicant

Prof Cedric Notredame

Details of proposal

I-Vision. Software publication remains a very empirical process in biology with a vast continuum between very structured communities built around established benchmarking standards and regular community-wide exercises – CASP being probably the best-known example – and smaller unstructured activities carried out in an ad-hoc fashion. Multiple Sequence Alignment development is probably the best example of a small-scale community with fragmented reference infrastructure. The rapid shift in methodologies for the acquisition of primary data and their computational processing means that an increasing number of in silico data processing methods are being developed in biology. Unfortunately, the deployment of published methods can be a frustrating process. The packaging is often highly heterogeneous, the reference datasets can be hard to deploy and the benchmark analysis may have to be re-implemented. As a consequence, no simple programmatic way exists to objectively decide among a group of alternative methods the one that is best suited to the task at hand. The goal of this proposal is to produce a proof of feasibility for the establishment of a meta-framework designed for software publication, deployment and live benchmarking. Within this framework, a software is systematically made of four components: the software itself, a reference dataset, an operator that operates the software and is able to quantify the quality of its output and an electronic notebook that bundles the three components and comes along with documentation. An important aspect of the project will be to evaluate the possibility of having these components distributed across suitable repositories so as to ensure they can be programmatically deployed without any need for a centralized repository other than the ones currently available. By the time the project starts the repositories may change but so far, we have considered an infrastructure built around the following components: GitHub, Jupyter, BioHub, BioTools, Zenodoo, Travis. These repositories will systematically be used following the recommended Elixir good practices. II-Influence on Open Research Practices. Our approach is built on the paradigm that "if you build it they will come". We think that imposing novel standards to the community will not be as productive as allowing the community to be drawn towards any existing standard because of the benefits they entail. We think that we will be able to build on our experience in Nextflow so as to set up a framework that will bring immediate benefits to a large share of the scientific community working on methods. Because reference datasets will be bundled the right way, anyone wishing to develop a novel method able to process these datasets into a prediction will have access to the whole system. As a method developer, it means that if you package your method the right way, it becomes immediately benchmarkable and usable by the whole community. The benefits are huge, and the over-heads neglectable with respect with the work needed in any case when publishing a new method. In our opinion, the burden on requirements should be kept at a minimum and should simply require a clear identification of the four compulsory components (software, reference, benchmarker, bundler). By doing so, one will make it possible for a standard to emerge naturally, simply because it will be the most useful. We plan to initiate a new journal based on this paradigm, but we would see it as an even clearer success if existing editors were willing to implement the paradigm outlined in this proposal and eventually help the establishment of an intraoperative community of methods and benchmarks across journals. The benefits for users will be many. First of all, they will have a proper procedure established for software publication, secondly, the standardized procedures will make it very easy to rapidly deploy existing software. The standardization of evaluation procedures will allow users to focus their efforts on novel predictive methods, novel datasets or novel benchmarking metrics. Standardization will allow these three core components to rapidly grow independently from each other. This same standardization will be a driving force for the interoperability of novel methods. It is important to insist that the standardization will occur as a bottom-up rather than a top-down process. In other words, the standards will emerge through the competitive use of alternative datasets, benchmarking procedures and benchmarking metrics. It is also important to insist that this proposal will not be dedicated to establishing a novel standard. Rather than doing so we will show how existing standards can be combined so as to fulfill the requirement of publishing four interoperable components in a FAIR way. III-Monitoring and Evaluation. The first criteria of success will be the implementation of at least four pilot entries: two transcriptome quantifier and two multiple sequence aligner, all featuring different benchmarks, different reference datasets, and different bundling procedures). These four pilots will be published along with guidelines on how to use them as a template for future submissions. The guidelines will include specific keywords and ontological terms that will allow the systematic monitoring of similar entries in public rep. Monitoring of these keywords will allow a quantification of the uptake.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal about sustainable software development. However, the methodology was not clearly described, and the evaluation plan would have benefited from more detail, for example identifying targets that would indicate success.

Title

Increasing the impact of open research by enabling crowd curation of figures and linking of preprint data

Lead Applicant

Dr Girija Goyal

Details of proposal

In a survey conducted by us, echoed by others, 50% of research results are not openly available (Fig. 1 in attached document). A fraction of these are submitted to data repositories but remain undiscoverable. This negatively impacts early career researchers who may have spent months producing these data. It pressures them to ignore openness in a drive to publish multi-figure papers. Now consider this from the view of the readers who are presented with a story with 50% of the data left out. Despite that, each individual figure in a paper is a story unto itself with multiple experiments. Thus, individual figures are read by each reader with unique insight. A particular figure may be of interest to me because of the reagents used, or the technique or a very specific part of the results. With the help of the eLife Innovation Initiative, we created ReFigure to allow users to connect incremental research results to previously published work or compile individual figures to create open remixes of papers. https://refigure.org/collections/item/d1a18740-5243-11e7-a161-1b348ae109e4/ is an examplar where a postdoctoral researcher identified a better diagnositc paradigm for Zika and made it openly available by curating case studies. A traditional paper showing the same results was published one year later and remains paywalled. The Zika ReFigure has been viewed >400 times. Imagine the impact if all 5 million articles on EuropePMC could be remixed with reader's insights and incremental data. Specific Aim I: Expansion of ReFigure capability to >7000 OA journals ReFigure beta only linked figures from PubMedCentral, PLOS, eLife and Figshare. Despite the restriction to just 4 sites, and just over 200 users, we refigured >1200 figures from ~250 journals (via PMC) from 87 different publishers (Fig. 2 in attached document). Reproducibility and cancer were among the top keywords (Fig. 3 in attached document). Now we would like to expand ReFigure capabilities to all open access journals and repositories. This requires a concerted programming effort as there is no clear demarcation between an open access article and a paywalled article for image harvesting. Lack of HTML/CSS standards for online articles further complicates this issue. Specific Aim II: Case studies on the use of ReFigure to improve access to and impact of open research. Case study on reproducibility with the Center for Open Science (COS): A high-impact use case for ReFigure is the curation of replicates from different sources. Reproducibility Project in Cancer Biology (RPCB) from COS is a project where 28 cancer studies were systematically reproduced and fueled the debate on scientific reproducibility. Currently the data for reproductions is on different platforms. Readers would find it valuable to see the replications side-by-side as on ReFigure.org. Further, in a majority of the cases, the studies were reproduced partly suggesting that a more granular interpretation would be helpful. An "agnostic" platform like ReFigure which can connect individual replications is ideal for inviting community review. Readers, who may be expert practitioners in the field, are often aware of the reproducibility of findings as illustrated by current Refigures on RPCB papers which provide complementary interpretations. To engage the community further, we will implement a commenting system on ReFigure. COS will collaborate on the design of the study and reader interface, creation of ReFigures and outreach to seek community feedback. We anticipate that engagement with RPCB ReFigures will encourage readers to create their own replication ReFigures. We will monitor the use of the commenting platform, altmetrics, increase in the number of replication ReFigures and their views, the diversity of journals connected, increase in searches on reproducibility related terms on ReFigure and finally outlinks to cited journals . Case study on connecting micropublications to peer-reviewed data: We will collaborate with micropublication.org, a publicly funded, non-profit site to link peer-reviewed micropublications from C.elegans researchers to relevant figures from OA journals and make them discoverable on

ReFigure.org. ReFigure also increases discoverability by letting users know that a paper has been refigured even when they are on a publisher website. User feedback will be assessed, paying close attention to whether they were able to use their open data to engage funders, employers, supervisors and colleagues. We will also monitor the increase in views and altmetrics of the micropublication.org. Case study on assisting undergraduate literature review with ReFigure: After informally surveying undergrads, we suspected that they were unaware of the distinction between open science and current paradigms, sources of open access literature and often learnt about finding and reading papers by trial and error (See Box 1 in attached document). As a pilot, we worked with a community college intern through the Mass Life Sciences Internship Challenge. Her feedback about the experience and the increase in her interim research products is recorded in Box.1. We will use the funding from the Open Research Fund to create an online/in-person guided course with a cohort of 10 students, to be run thrice during the funding period. Impact assessment will be conducted by entry and exit survey, increase in open interim research products.
Decision **Not shortlisted**
Comment on decision from Wellcome The proposal was innovative and potentially impactful. However, there was some concern on the feasibility of the approach and the evaluation plan would have benefited from more detail.

Title

Genome.zone: Development of an online portal for clinical microbial diagnostics using next-generation sequence data

Lead Applicant

Dr Rhys Farrer

Details of proposal

Genetic characterisation for a pathogenic bacteria or fungus in a hospital can take weeks if not months, with the main bottleneck being the analysis time, cost and lack of specific expertise. Our vision is to provide a free, and rapid (analysis completed under an hour, with less than one working hour intended) online tool for assessing the biological origins and other clinically relevant information from WGS data, without the need for a dedicated bioinformatician. This online tool is intended primarily for clinical microbiologists and healthcare scientists whose objectives are to provide in-depth patient management information linked to virulence and AMR as well as infection prevention control information for cross transmission investigation. Genome.zone will rapidly confirm the presence or absence of infectious agents, including typing information, AMR profiles, and other similarities/differences between samples. Genome.zone may also be useful for wet-lab scientists and bioinformaticians looking for the closest reference genome available in public databases. Whilst similar tools exist, current alternatives are either directed at high Bioinformatics literacy (e.g. Galaxy) or charge a premium for most functionalities and searches (e.g. One Codex). Our tool would be free, and designed with the needs of clinical teams in mind including both language, visual presentation and data inclusion requirements. Genome.zone would include a secure login/user area, where potentially confidential datasets can be uploaded. Once uploaded (using an integrated tool such as Aspera – developed for high-speed uploads of large files), the user can run an online implementation of a recently developed distance estimation algorithm (such as MinHash or Kmer-db), which purport searches of tens of thousands of species from next-generation sequence data in under 4 minutes on a modern workstation. Users can then visualise the results with interactive JavaScript charts such as those implemented by Highcharts, including hits per species/genus, dendrograms to support outbreak investigation, charts of common resistance genes, and virulence genes detected, as well as details of their accuracy. Reporting will be similar to content that is currently frequently provided to clinical teams, such as those provided by Public Health England for continuity within clinical records. For example, there will be an option to download a report designed closely to support the use of results within ISO 15189 accredited diagnostic environments. Uploaded datasets will be available for re-analysis for a month which equates to the length of time for a standard outbreak investigation. Comparisons between multiple samples will also be permitted (for example to detect outbreaks). The downloadable reports will be stored long term within dedicated controlled user accounts. Over the course of this grant, we will benchmark the fastest algorithms and implementations available for rapid diagnostics and test our tool with a variety of datasets, testing for accuracy, speed and robustness. We have already developed a proof of principle (see http://www.genome.zone, and accompanying screenshots), based around the PHP web framework Laravel and JavaScript libraries such as Bootstrap, which is hosted by Dreamhost at Genome.zone. However, numerous limitations and features require further development for it to be an effective tool. Firstly, we need to model and validate the tool with clinical data to determine how it can be used to change patient monitoring and support IPC actions. Secondly, our current hosting plans prevent memory intensive tasks (e.g. via daemons) and have limited bandwidth. We aim to host Genome.zone on a dedicated web-server, which would be cost-effective (cheaper than web-hosting from Amazon, Google, Dreamhost etc., with the higher specs necessary in terms of memory and storage), allow root-access for implementing custom workflows and pipelines, and will be maintained by Aberdeen University as part of their computer cluster (thus dealing with set-up and running expertise, power surges/cuts, server restarts and service longevity etc.). The development and success of Genome.zone will stem from the co-design and validation of the reporting tool based on collating of available clinical report formats, ISO 15189 guidance and discussion with clinical teams about information may result in changes in patient management. Suggestions that may change management of an outbreak will be validated by running retrospective outbreak WGS through the tool with a clinical team to determine which information would have supported changes in patient management or Infection Prevention Control Interventions. We will publish a short communication to an open-access journal to demonstrate the potential cost and timesaving of rapid WGS diagnostics through this case study, and the validated reporting system will be used prospectively by collaborators and evaluated over the course of the project. Our team brings together a clinician, data scientists and a dedicate programmer, which will have regular contact via online services, but also at a one day 'hackathon' to discuss plans for new functionalities and implementations. To that end, we aim to have three meetings over the year. We will also run a one-day workshop aimed at clinicians and Health Care Scientists at London Hospitals to showcase our tool and discuss active joint development. Further advertising will occur on university and hospital internal mailing lists, posters for the workshop, and a dedicated Twitter account and mailing list/help email address for updates and assistance. Usage will be tracked and evaluated with a web-tracker, and database metrics.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The proposal was from a strong team, and the methodology was clearly described. However, the level of innovation proposed was considered limited. The application would have benefitted from more information about how the outcomes and impact would be evaluated.

Title

Improving registration and reporting of university-sponsored clinical trials in seven key countries

Lead Applicant

Ms Priscilla Li Ying

Details of proposal

VISION Open research principles are firmly embedded in university-run clinical trials worldwide AIMS (1) Universities and researchers will begin the lengthy process of retrospectively reporting the outcomes of past clinical trials that have so far remained hidden to accelerate scientific progress, improve the quantity and quality of the medical evidence base, and reduce research waste. (2) Universities will embed global best practices in clinical trial transparency in their policies and standard operating procedures, and researchers will apply them in their day-to-day work, ensuring that in the future, trials are registered and reported in line with open research principles. AUDIENCES (1) Primary audience: Universities' senior managers, research governance staff, current and future medical researchers. (2) Secondary audiences (reached through media coverage): Research funders, policy makers, and the general public, including patients and trial participants. Phase I (Dec 2018-May 2019) Selection: 40 universities will be selected from 7 countries among the top 10 trial locations worldwide. Universities within these countries will be selected based on the World University Ranking, with caps per country to ensure national balance. A further 10+ universities will be selected from UAEM's network of chapters, some of which are likely to be based in additional countries. Training of trainers: 10+ student leaders from participating chapters will be trained in open research principles related to clinical trials, and in advocacy. Pre-registration of project's research protocol. Assessment of universities' trial transparency performance using two metrics, summary results posting and overall outcome reporting. Summary results will be assessed by combining data from the EU trial registry (using the EUCTR Tracker, due to be launched in Sept 2018) and Clinicaltrials.gov (using an existing tool developed by UAEM-UK and TranspariMED). Overall outcome reporting will be assessed manually using a research protocol drawn from the Charite's IntoValue#1 project (which has completed data collection and thus demonstrated feasibility) to identify 5 pre-2016 unreported trials per university. Data sets will be shared with all 50+ universities for their review, feedback and validation. Phase II (June-Nov 2019) Outreach: 100+ principal investigators of trials flagged as unreported will be emailed individually, reminded of obligation to share outcomes, and invited to commit to making trial results public; responses will be tracked. Guide: Writing, expert review, publication and dissemination of a concise how-to guide on best practice trial registration and reporting. Providing technical assistance and advice to universities seeking to strengthen their policies and processes. Reports: Launching one report for each of the 7 focus countries, quantifying performance per university and across universities, and listing all unreported trials. All underlying data sets made accessible online. Launch coordinated with local UAEM chapters and exclusive national media partners. Engagement and dissemination: Student-led public events at universities to promote open research principles and advocate for trial transparency. Proactive dissemination of reports and visuals via Twitter, Facebook, and LinkedIn. Closeout. Documenting project performance against target indicators; narrative and financial reporting. (II) Influence on open research practices We believe this project is highly likely to lead numerous researchers to retrospectively report the outcomes of old clinical trials, and universities to put into place policies, procedures and systems that improve the quantity and quality of outcome reporting for future trials. Currently, around half of clinical trials never report results. Research consistently shows that in this regard, university-sponsored trials perform even worse than industry-sponsored trials. In addition, prior research by TranspariMED shows that trial registry entries by universities are often incomplete, inaccurate, and outdated: https://docs.wixstatic.com/ugd/01f35d_15c506da05e4463ca8bd70c2b45bb359.pdf Most universities have so far had few incentives – positive or negative – to tackle this problem as their

performance has remained largely invisible and unchallenged. Even today, many prominent universities still do not centrally track their clinical trials or monitor their registry entries: https://www.transparimed.org/single-post/2018/07/02/London-School-of-Hygiene-and-Tropical-Medicine-fails-to-tackle-research-waste However, experience consistently shows that when universities' performance is made visible and they are called to account over their performance, many of them take significant steps to address the issue. Examples include a 2015 STAT News investigation, two 2017 TranspariMED audits of individual universities' registry entries, and the EBM Data Lab's Trials Tracker and FDAAA Trials Tracker: https://www.statnews.com/2018/01/09/clinical-trials-reporting-nih/ Research also indicates that prompting those responsible to report past trial outcomes can have significant impact: https://www.bmj.com/content/349/bmj.g5579   A key reason for the remarkable successes noted above is that while the non-reporting of trial results is clearly unethical and scientifically unsound (and sometimes illegal), preventative countermeasures – centrally tracking trials and strengthening policies – cost universities little to implement. This project will boost incentives further by mobilizing 10+ UAEM student groups to apply pressure from within their universities. (III) Monitoring and Evaluation The project will be monitored throughout and a mid-term assessment will inform the project's progress, against the following indicators: # of universities assessed and contacted (target: 50+) # unreported trials identified (target: 250+) # of principal investigators of 'missing' trials d
Decision **Not shortlisted**
Comment on decision from Wellcome This proposal aimed to embed open practices into institutions. However, the level of innovation proposed was limited.

Title **Rapid PREreview: A rapid preprint review platform to support outbreak science**
Lead Applicant **Dr Monica Granados**
Details of proposal Outbreaks of infectious diseases are a global concern. Responding to them effectively requires a global effort informed by scientific evidence, but the traditional scientific communication system is not designed for rapid dissemination of data, methods, or results. In practice, manuscripts can take months, even years to progress from submission to publication. During epidemics, months can mean lives lost. Recently, preprints, scientific manuscripts posted online prior to peer review, have emerged as a tool to accelerate and democratize the dissemination of scientific information. In the recent Zika epidemic, while only a minority of all publications had associated preprints, the majority were available approximately five months before their peer-reviewed versions (https://doi.org/10.1371/journal.pmed.1002549), accelerating the accessibility of this research to anyone with internet access. While editorial peer review plays a vital role in filtering scientific content, it is a slow and subjective process. In the context of outbreaks, where timely decision making is critical, rapid reviews become even more important. Policy makers, journalists, and scientists need to be able to quickly identify the most important advances while avoiding basing decisions on poorly designed studies and weak evidence. Here, PREreview will join forces with Outbreak Science to develop an open tool, Rapid PREreview, to enable the rapid assessment of scientific contributions during outbreaks and beyond. PREreview (https://prereview.org/), an open project that seeks to diversify peer review in the academic community by promoting the crowdsourcing of pre-publication feedback, is developing tools to support peer review of preprints. Outbreak Science is a new non-profit organization founded by scientists with a wealth of experience in responding to epidemics and dedicated to advancing the use of science to inform outbreak responses. Rapid PREreview will be developed as an extension of PREreview 2.0, a separately funded, new open source platform designed to engage the whole scientific community with collaborative and constructive peer review. Many of the key features of PREreview 2.0 will be extended to Rapid PREreview, including users' ability to sign-in with their ORCID IDs (with optional pseudonymity), solicit preprint feedback, leave comments, and endorse others' reviews, all in accordance with PREreview's code of conduct. Rapid PREreview will provide mechanisms for scientists to rapidly identify the quality, novelty, and importance of preprints via a short and structured review requiring only a few minutes to complete. We envision this interface to be comprised of a series of multiple choice questions which have a low barrier to participation (see beta version of this questionnaire in Additional Information). Data summarizing aggregated community feedback will be made openly available both as an interactive visualization of the contributions related to a given topic, and in the form of raw data, stripped of any personal identification for anyone to access. Just as we did for PREreview 2.0 (https://airtable.com/shrlwN9DcyLD2GyUE), for this extension we will follow the design thinking approach, an iterative framework for problem-solving that involves assessing known aspects of a problem, challenging assumptions, testing ideas, and integrating feedback. Assessment protocols will include tracking the number of users and reviews, and monitoring how many communities we engage, including scientific societies, institutions and community groups across the world. While researchers in the field of outbreak science constitute the primary target audience for the proposed tool, we believe other stakeholders will benefit from its implementation. For example, rapid reviews can help journal editors make more objective decisions about the soundness and novelty of the research. They can help policy makers quickly navigate through the complex emerging evidence in the midst of outbreaks. Journalists can benefit from the aggregated visualization of rapid reviews, facilitating the identification of robust recent discoveries. Finally, rapid reviews could reveal useful information to health providers working in areas afflicted by the outbreaks, as well as inform the very people who are in immediate danger of being affected by

the infectious disease. Outbreaks are a context in which rapid reviews are critical, but there are many other scientific areas where rapid reviews may meet important needs. To facilitate broader use, this tool will be developed as an open, scalable project that could be adopted and adapted to any number of scientific disciplines and other interested stakeholders. Both PREreview and Outbreak Science recognize the important role community plays in the success and adoption of new tools. To this end, we propose organizing and facilitating sprint-like events across the world – similarly to how Mozilla leads their annual Global Sprint events. These sprints will allow us to build awareness of the tool, generate new content, and receive important feedback so that we can make further improvements. As scientist-led, open science initiatives, both PREreview and Outbreak Science are well-positioned to build an extensive community of rapid reviewers. Since its infancy, PREreview has received the support of well established open science communities, such as Mozilla and OpenCon, attracting the attention of various stakeholders. Together with Outbreak Science's key subject matter expertise, we have a unique opportunity to move beyond solely providing author feedback; here we can complement our existing work by harnessing the collective knowledge of a community of researchers to provide critical data necessary to tackle global concerns.

Decision

Funded

Comment on decision from Wellcome

The application was from a strong team, proposing to generate an important tool. The proposal was innovative and had potentially wide-reaching impact.

Title

Integrating ARTiFACTS Attribution Services with Open Science Software Systems

Lead Applicant

Mr David Kochalko

Details of proposal

(i) Vision While programatic access to open science resources exist, these nearly all focus on transmitting and accessing data from articles published in traditional journals; a narrow view of academic output that limits scientific progress and inhibits the free and timely exchange of research-based ideas. Our vision is to leverage emerging Web 3.0 technologies including A.I., Big Data, and blockchain in support of improving timely and secure access to emerging scientific ideas and data and to integrate these technologies into other established tools in the open science software ecosystem. A key step toward this goal, which is the essence of this proposal, is to develop and deploy an API to enable direct interaction between our Artifacts.ai system and other open science software tools. We have identified ORCiD, and OSF as initial partners and will develop our API to facilitate integration with all interested parties. Welcome Open Research represents a viable potential partner. See Participants section for more information about Collaboration Partners. Artifacts.ai addresses persistent barriers in scholarly communications which delay access to research findings, timely recognition of the contributions of scientists, and slows scientific advancements for society. Publishing bias for novel findings prevents confirming or negative results from appearing in the literature, and prevents valuable intelligence from reaching potential beneficiaries. Web-based indices of published findings capture only 20% of all published science and scholarship, hiding most of this content from view and rendering it inaccessible. Virtually none of the supporting artifacts which are instrumental to published works are reliably linked or accessible for peer reviewers or other research teams, creating a vast corpus of hidden knowledge. By facilitating sharing of early stage findings and enabling these creative works to receive citations, Artifacts.ai will speed the research engines of teams, their funding organizations, and improve collaboration. Artifacts.ai delivers three simple, yet powerful services that facilitate sharing and enable real-time recognition for all discoveries and materials created throughout the entire research process: Establish proof-of-existence and confirm provenance at any time, Protect and manage intellectual property while concurrently facilitating knowledge and content sharing, Provide and receive valid, break-proof attribution and assignment of credit. In doing so, researchers will: Have the ability to acknowledge the works of others by giving a citation and to receive credit and recognition for their own creative work products. Be able to share their research findings, of all types and throughout all stages of the research process, while retaining control over the provenance of their IP and the parties able to view and build upon their works. Rely on a distributed ledger and transaction engine that records a permanent, valid, and immutable chain of records for all research artifact 'transactions' including citing transactions. Building the connective linkages between Artifacts.ai and open systems such as ORCiD and OSF will further enhance the value of these platforms for their research communities. Making the ARTiFACTS attribution engine available to these communities will provide creators with the confidence and incentive to share their work products earlier and eliminate delays imposed by publishing cycles. Primary activities will include: Development and documentation of an Artifacts.ai API. Creation of plug-ins that would expose the Artifacts.ai functionality within ORCiD and the OSF application. Coordination with ORCiD and OSF teams for incorporating Artifacts.ai as an additional service on their platforms. Testing and quality management process to assess design usability and debug known issues. Deployment and support of new services. (ii) Impact Accelerate research velocity: Enabling the ORCiD and OSF researcher communities (5.1 million and growing) to establish proof of existence and authorship over their creative works will reduce barriers to reporting outcomes. Findings will be able to be reported, under the control of their creators, well in advance of formal publishing cycles. Recognition for contributions: Researchers will be able to receive attribution in the form of formal citations to all of their creative works, not merely the published manuscript. Their profiles will reflect a more complete picture of their contributions, views into how their work has progressed, and the knowledge they have created, all of which are force multipliers for their ability to win grant awards, appointments, and career advancement. Improved decision making and research quality: With deeper insight into how research has progressed where the supporting evidence is visible, funding organizations will have qualitative information available when setting priorities and awarding specific grants. Researchers will have access to the necessary details to conduct confirming analyses or other evaluative work regarding reproducibility. Researchers will have incentives for such work knowing they will receive recognition (iii) Success Indicators Development success will be achieved upon completing the build, test, and deployment of the Artifacts.ai API service. Metrics we will monitor include: Number of users   Frequency of use   Artifacts shared   Proof of existence transactions   Artifacts cited (ie, Attributions made)   Along with general analysis of the uptake and use of the API, we will undertake a detailed use-case analysis on an existing Population Health dataset (https://pophealth.discoverresearch.ai/). This is explained further later in this proposal.

Decision

Not shortlisted

Comment on decision from Wellcome

The applicant opted not to share this information

Title

Open Source Period: a model for co-created collaborative open research?

Lead Applicant

Dr Siouxsie Wiles

Details of proposal

(i) the vision for your proposal, including aims, target audiences, activities; Our vision is to develop a model for co-created collaborative research that embeds within it the principles of open research and the philosophy that such projects must be carried out in an inclusive, culturally-appropriate way, that does not ignore and/or exploit indigenous and/or under-represented communities. The aim of this project is to develop a set of guidance documents that will allow diverse communities, be they patient groups, curious citizens, or academic researchers, to establish their processes and 'codes of conduct' for each stage of the research process. To achieve this aim, our project is divided into two objectives: Develop a set of guidance documents for a co-created collaborative open research project based on the study of menstrual cups, Modify the guide documents so they can be used as a template for others wishing to embark on a co-created collaborative research project. Objective 1: Develop a set of guidance documents for a co-created collaborative open research project based on the study of menstrual cups Menstrual cups are marketed as a cheap and environmentally-friendly alternative to tampons and are growing in popularity. They are also marketed as extremely safe, but a recent study suggests this may not be the case. Social media has allowed a community of interested researchers, menstrual cup users, and social enterprises, from around the world to begin to form. This community is keen to answer a range of questions that its members have about menstrual cups, from what bacteria may be able to grow on menstrual cups to whether menstrual cup users are more likely to suffer from urinary tract infections and thrush (see additional information section). We will use a combination of literature reviews (to identify examples of current best practice) and stakeholder workshops to develop a set of guidance documents to enable our community of researchers, menstrual cup users, and social enterprises to co-create and carry out their research project (see schematic of our proposed model in additional information section). Our documents will cover: Community-building: How can a community of stakeholders be built that will ensure the entire project is carried out in an inclusive, culturally-appropriate way, that does not exploit indigenous and/or under-represented communities? How can we ensure that the language we use is inclusive and that we reach stakeholders from all walks of life around the globe? Identification of research questions and setting a research agenda: How will research questions be identified and how will the community decide which are the important questions to focus on? How will these questions be turned into a research agenda? How will suitable projects be developed? How can best practice in open research be embedded into the research agenda to ensure that the methodology is reproducible and that all the outputs that are generated are findable, accessible, interoperable and reusable (FAIR)? Carrying out a research agenda: How do communities set up their own citizen/participatory science projects? Could communities use platforms such as Science Exchange (https://www.scienceexchange.com/) to access some of the world's leading scientific service providers and most innovative scientific technologies? How can communities fund their research agenda? What codes of conduct should apply to those involved in carrying out the research agenda? Data analysis and dissemination: How will data be analysed and disseminated? Objective 2: Modify the guide documents so they can be used as a template for others wishing to embark on a co-created collaborative research project. Once Objective 1 has successfully been completed, we will solicit engagement with a wider group of stakeholders, including charitable funders, patient groups, and open research and citizen science advocates to convert our documents into an adaptable template suitable for use by others in a wider context of settings. (ii) how your proposal will influence open research practices in your field or more broadly; We anticipate that by embedding open research practices into our model, those researchers who engage with the menstrual cup community to carry out the research agenda will see the benefit of such practices and act as exemplars and advocates for open research within their wider research communities. We also anticipate that the documentation we develop will provide practical guidance for any researcher wishing to make their data outputs findable, accessible, interoperable and reusable (FAIR). (iii) how you will monitor and evaluate your proposal, including success indicators. The outcome of this project will be the development of two sets of documents for the co-creation of collaborative open research projects, the first for use by a community interested in studying menstrual cups, and the second by the wider community. We will measure success in the first instance by whether the documents we create are fit for purpose and enable a collaborative open research project centred on menstrual cups to begin. Broader success will be measured by the adaptation and use of our documents and model by other communities. This will be monitored by downloads of the documents and tracking where their DOI's are cited.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The application was from a team with a strong track record in community building. However, the proposal was unfocused and level of innovation was limited. The application would have benefited from the inclusion of collaborators from like-minded organisations to share and publicise the work beyond open science silos.

Title

Harvesting lost gene signatures with improved open-source analysis to accelerate biomedical discovery

Lead Applicant

Dr Matthew Ritchie

Details of proposal

Specific goals Aim 1. To unlock thousands of new gene signatures and improve base infrastructure for gene set enrichment analysis. Through collaboration with genomics data curators at the EMBL-EBI's Gene Expression Team (Dr Papatheodorou), we will collect signatures generated through the Expression Atlas re-processing pipeline (iRAP and RNASeq-er API [3]) which has produced over 4,000 signatures to date that are not readily accessible outside this platform (Figure 2A). Methods available in the mixOmics R package [4] by Dr Lê Cao to integrate expression data across multiple platforms will be applied to create more robust signatures in selected disease-specific but related data sets (as previously adopted in the Stemformatics atlas resource [5]). Similar signature collections from the ENCODE gene set hub [6] and TCGA will also be gathered. Together with Prof Morgan, we will ensure these signatures are conveniently accessible through Bioconductor's AnnotationHub web service [7] to allow dynamic querying and re-use by other software. Infrastructure improvements will involve updates to the GSEABase Bioconductor package [8] to accommodate new information, such as interactions between genes and the directionality of expression changes in order to allow more sophisticated gene enrichment analyses. Work towards a consistent standard will maximise the interoperability of open source tools. To reduce the loss of signatures at the publication stage, we will trial the collection of genomics signatures alongside the raw data upon submission to ArrayExpress. Direct follow-up of contributors who recently publish their work will collect signatures in human readable format that can be added directly to the Expression Atlas collection without the need to reprocess raw data. We will also engage the editors of 2 open-access journals (F1000Research and GigaScience) to assess any barriers from the publisher's perspective to making the gene signatures they publish easier to re-use and discover. Aim 2. Improve ensemble based methods for gene set testing and add new capabilities for network based methods for gene signature analysis that will be incorporated in the recently published EGSEA Bioconductor package [9,10]. Our Ensemble of Gene Set Enrichment Analysis (EGSEA) approach (Figure 3) was originally tailored to RNA-seq data and combined 12 methods together with tens of thousands of gene signatures from a diverse range of sources, including MSigDB, KEGG and GeneSigDB. We recently added support for microarray data and will further improve these method by developing a more sophisticated approach to combine results from different algorithms. Rather than giving results from each algorithm equal weight, as per the current implementation, a weighted analysis that is either fixed, based on performance rankings from previous studies, or is adaptive for each analysis could be envisaged. We are also planning to explore the use of molecular interactions to improve gene set testing performance, especially in the context of ensemble approaches. Finally, we will enhance EGSEA's reporting capabilities by adding the ability to interactively search gene set plots such as barcode plots and summary plots for entire gene signature collections to make exploration of the results a more intuitive experience for end users. This will build upon the plotting functionality in our Glimma package [11] and use plotly. All of these features need to be implemented and tested using suitable benchmarking data sets to ensure they are performing optimally. New resources to enhance the usability of these packages at both the command line (MetaGeneSig123 Bioconductor workflow package) and in a point-and-click manner (Galaxy tool) will lower the barrier for researchers to capitalise on our developments (Figure 2B-C). This aim will be pursued in collaboration with Dr Alhamdoosh and Dr Ng. Outcomes A timeline for this proposal is given in Figure 2D. Our work will benefit biomedical researchers world-wide, as the signatures curated and infrastructure developed will be freely available for re-use. Success in the short-term will be measured by the number of high quality signatures that can be curated (> 16,000) and the availability of new versions of existing open-source software packages (GSEABase and EGSEA) and entirely new packages (MetaGeneSig123 and a point-and-click Galaxy tool). A high quality pre-print outlining our achievements will be made available on bioRxiv and submitted to an open-science journal (F1000Research or GigaScience). Over the longer term, success will be measured by standard evaluation metrics that include the number of software downloads and citations. Through our interactions with Bioconductor software developers, it is our endeavour to propose new standards for representing gene signatures that will be widely adopted across open-source software platforms. Insights gained from our interactions with data contributors and journals will transform the way gene signatures are collected and provided to researchers more generally, thereby increasing knowledge flow into the public domain. Risks Our research team has a well-established track record in methods development across a range of popular R-based software packages. The risks associated with this project are therefore very low. In Aim 1, there is a risk that the improved standards we develop may not be widely adopted by others, since developers of existing software have already made choices on how to represent gene signatures. We will minimise this risk through engagement of developers during the improvement process and at annual Bioconductor conferences in the US, Europe and Asia.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The proposal has clear and wide-reaching potential impact and is from a team with a strong track record in open research. However, the proposal would have benefited from more detail on how the impact of the resource would be evaluated.

Title
Distributed Ledger Technology Based Health Alliance Platform for Enhanced Research Collaboration

Lead Applicant
Prof Vallipuram Muthukkumarasamy

Details of proposal

The main objective of this research is to build a common platform for various stakeholders in health sector where all participants would be able to securely access, exchange and reuse sensitive health data without compromising authenticity, privacy and integrity. A smart contract driven Distributed Ledger Technology (DLT) based system will be designed and developed to provide an innovative platform for collaboration among the researchers, clinicians, patients, and health service providers. The proposed system has the potential to increase privacy, safety, transparency, auditability and accountability for a number of use-cases such as clinical trials, data sharing and management of patient records. Distributed Ledger Technology may provide a potential solution for the most common type of clinical fraud e.g., editing records, outcome switching, selective reporting, and falsifying patient consent. An immutable and irrefutable log of patient consent can overcome the problem by providing clinical trial subjects ownership of their own information while giving an audit trial for clinical staff, regulatory bodies and researchers. Similarly, through the use of a distributed ledger of medical records, DLT-based system may have the potential to offer interoperability since all healthcare providers will have the same copy of this ledger. In addition, patients can have full control of their medical data as they have ability to accept, delete or modify relationships with healthcare providers such as hospitals, insurers, and clinics. In the proposed system, patients may also be able to anonymously share their medical information to research community and thus could decrease the burden on research subjects though allowing the reuse of the existing data. The DLT-based system may not only prevent researchers from changing data or endpoints, but also facilitates collaboration with confidence that nobody can take anyone's 'credit' away. At present, the stakeholders of healthcare systems are working as isolated entities in silos rather than as parts of an integrated care system in most of the countries. With an integrated system, information would be available for use by approved third parties, insurers, hospitals, clinics, researchers and patients as necessary. This will speed up the manual or traditional isolated procedures of finding and sharing data through facilitating real-time collaboration. One of the key goals of this project is to develop a common collaborative platform where patients would be able to share their sensitive health data anonymously for secondary purposes such as research. As an example, researchers use artificial intelligence enabled screening algorithms to detect breast cancer which requires millions of mammograms to train the system first. However, it is quite difficult to create such a massive database due to privacy laws and regulations. To protect the privacy of individual patient data, the best way is to anonymise the data, wherever possible, before sharing with other parties. Thus, the proposed project will provide an open platform where the researchers will have access to a large volume of research data and they will also have the opportunity to make their research outputs available for others. To ensure the active participation of patients in building the large dataset, the project also aims to introduce tokens and incentivise the patients for sharing their health data. In a nutshell, the outputs of the project will develop a proof of concept model to ensure the following requirements: Findable- stakeholders of health sector would be able to use the platform for clinical trials, sharing information, and monitoring and tracking activities; Accessible- the platform and dataset is open and accessible to everyone; Interoperable- the patient electronic health record can be shared across multiple hospitals or health service providers; Reusable- the collected data with patients’ consent is reusable for research purposes. We will set a number of milestones to evaluate the progress and test the outcomes of each step. Initially, the requirements for the proposed system will be collected through interviewing the stakeholders and inspecting the existing systems. Then, the collected requirements will be reviewed and analysed by the health experts to eliminate ambiguity and redundancy. After the requirement analysis, a prototype will be developed to visualise the system functionalities and to understand user expectations. The stakeholders will be asked to evaluate the prototype and their feedback will be collected to further investigate the system model. The experts will also be involved in this phase to give their feedback on the users' demands. Once the prototype is finalised, the technical team will develop the platform according to the requirements which will go through a number of testing procedures. In this stage, the platform will be tested and evaluated by the stakeholders to see whether the system fulfills their expectations. Since user experience is the key to success for any interactive and collaborative system, we aim to conduct a thorough testing procedure on usability, reliability, efficiency and performance of the developed system. On the basis of the evaluation outcomes and expert judgement, the required changes will be identified and the system will be re-tuned. This is an iterative process which can continue until the system meets a certain level of user expectations.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal aiming to facilitate patient data sharing. However, the methodology was not clearly described and potential impact of this proposal to transform health research through openness was limited.

Title

Developing and Publicising an Online Platform for Facilitating Research Collaboration and Resource Sharing

Lead Applicant

Dr Balazs Aczel

Details of proposal

The Vision of the Proposal The goal of this project is to foster openness, efficiency, and quality of scientific research by developing an online platform that facilitates collaboration and resource sharing amongst scientists. Researchers in the health and behavioural sciences often struggle with resource constraints such as inadequate access to study participants, specialised equipment, or collaborators with particular expertise. These limitations hinder the advance of science by preventing scientific projects from reaching their full potential. For example, across multiple scientific disciplines, the high prevalence of studies with small sample sizes has potentially contributed to generating large bodies of published literature that are uninformative, misleading, and often irreplicable (Fraley & Vazire, 2014; Ioannidis, 2005; Button et al., 2013). One solution to this inefficient use of resources is to facilitate greater collaboration and resource sharing among research laboratories. Often resources such as study participants, equipment, and expertise that are required by one laboratory (NEEDS) are in surplus at another laboratory (HAVES). Unfortunately, in the present scientific practice, encounters between HAVEs and NEEDs are mainly incidental and rely on fragmented local networks. Our proposed solution is to develop a sophisticated online platform where researchers can advertise their diverse HAVES and NEEDS and make exchange agreements to facilitate the transfer of resources to areas where they are needed most. Ultimately, such a platform may help shift the scientific ecosystem away from the model of siloed laboratories that monopolises scarce resources towards a collaborative network of laboratories that openly exchange key resources in a mutually beneficial scientific commons. In our recent survey at the Psychological Science Accelerator network, 92% of the respondents expressed interest in sharing their laboratory capacities. A nascent online platform, StudySwap (Chartier & McCarthy, 2017), has demonstrated the viability of such a solution in the context of exchanging study participants between laboratories, mainly in the field of psychology. In a typical successful collaboration, researchers formulate a mutually beneficial "exchange agreement" outlining technical and ethical issues and relevant incentives, such as authorship on the resulting publication(s). While StudySwap provides a proof of concept, its technical infrastructure limits its potential to diversify into a broader array of resource exchanges that could benefit the health and behavioural sciences. We are proposing an ambitious upgrade of this platform that provides extensive new functionality to support diverse resource exchanges. Our platform will reduce waste, maximise efficiency, and enhance the quality of research outputs across multiple scientific disciplines. Our team proposes to develop an open resource exchange platform with the following key features: Advertising HAVES and NEEDS for a diverse range of scientific resources, including: study participants (especially hard-to-reach clinical populations) specialised equipment (e.g., eye-trackers, flow cytometry, microarrays) research materials (e.g., software codes, experimental stimuli, biological reagents, cell-lines, plasmids) skills (e.g., programming experience, translators) Filtering search functions Platform profiles linked to professional profiles Email notifications system for matching HAVES and NEEDS Tagging requests by area (e.g., clinical research), expertise, and request type (e.g., consultation). A reputation/credit system for users. A demo-version of our platform is available here: https://collabora.herokuapp.com/ Target Audience The primary target audience of the proposed platform is researchers in the health and behavioural sciences. However, the recurring need for scarce resources, specialised equipment, and collaborators with particular expertise are

common across research domains. Therefore, our ambition is to facilitate resource sharing across a broad range of disciplines. To this end, we are in contact with the Center of Open Science (COS), which expressed its support of our goals, and its openness to a potential collaboration.
Proposed Activities In Stage 1 of the project, we will develop open-source code to support a professional and user-friendly collaborative resource sharing platform. In Stage 2, we will pilot test the site within three open-science networks of which we are active participants, including the COS Ambassadors. In Stage 3, we will publicise the platform to the broadest researcher community. For a detailed Action Plan see Supplementary Table 1. Influencing Open Research Practices Our platform will be global in scope and open to all. This will increase diversity by connecting researchers across geographic and cultural boundaries. Collaboration inherently promotes open research practices, especially when it takes place outside of existing personal networks. Our platform will facilitate a range of collaborative and resource exchange practices that promote and often require transparency. For example:

When laboratories agree to share research participants this not only increases statistical power, it also leads to sharing materials, expertise, analysis code, and data.

Analysis exchange (independent verification of data analysis) increases the validity of reported results, but also requires data and code sharing.

Multi-site studies, such as Registered Replication Reports, can bring together researchers who adopt differing theoretical positions within an open collaborative framework that encourages pre-registration, open data, open materials, and open analysis code.

Agreement templates in which they can commit to sharing their data, code, and study materials. Monitoring and Evaluating the Proposal The success of the project will be monitored in 3 stages. Each stage will be considered complete when its pre-established requirements have been fully met. For a detailed assessment plan and success indicators see Supplementary Table 2.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a clear proposal based on a compelling idea, which could have real impact if adopted by the community. However, there were concerns over the feasibility of this in practice and the likelihood of this achieving widespread uptake.

Title

Creating an innovative country-specific Open Research Data Hub for health researchers: Ghana as a case study

Lead Applicant

Dr Michael Head

Details of proposal

(i) the vision for your proposal, including aims, target audiences, activities; Overall vision: Our vision is to utilise dynamic, near real-time data to promote successful health research endeavours through an open-access hub providing a country-specific evidence base, here relevant to Ghana. Aim: The aim of this project is to create a simple-to-use, online, and open-access hub that allows researchers to easily access data from a multitude of difference sources to inform thinking, strategy and project development around research and health in Ghana. Target audiences: The hub will support Ghanaian and global researchers with an interest in health in Ghana. This includes, for example, university academics, policymakers, and clinicians. The hub will be useful for individuals to easily find national information all in one place, update knowledge on the health landscape in Ghana, identify data to answer a knowledge gap, identify individuals and institutions with specific skills or capacities such as to host and run a clinical trial, and provide background information to support research endeavours. Activities: University of Ghana postdoctoral researchers will lead on the identification of in-country data sources, liaise with local stakeholders to identify specific knowledge gaps and priority areas for Ghana. They will use this feedback to perform analyses on secondary data, and incorporate findings about or of relevance to Ghana into the hub. University of Southampton colleagues will lead on the development of the hub infrastructure and perform a review of the research landscape in Ghana. This will include mapping investment trends, research institutions and infrastructure, and sites of research activity. University of Southampton colleagues will lead on the engagement with international stakeholders for example: at the WHO, the World Bank, and the Wellcome Trust. The hub will be disseminated through a half-day meeting in Accra at the end of the study with around 30 stakeholders in attendance. This will raise awareness and promote the hub, solicit feedback on appropriate ways forward, and allow discussion on future uses of the hub. Training – The Ghanaian-based researchers will travel to Southampton for training in freely-available software such as Microsoft Power BI (for interactive report generation), qGIS (for geospatial methods), and Stata (for statistical analysis). This will be facilitated by the existing high-quality support infrastructure at the University of Southampton, including the WorldPop geospatial research group and Public Policy@Southampton, who support researchers in translating findings into policy, capacity building, and research impact. The researchers will receive sufficient training to enhance their professional development and train further Ghanaian colleagues on how to review the research landscape with appropriate software for visualisation and analysis. This will build in-country capacity, upskill, and increase the profile of the researchers, and promote sustainability of the hub and surrounding methodology beyond the end of the funded project. (ii) how your proposal will influence open research practices in your field or more broadly; This hub can provide an innovative model of how best to bring country-relevant data together to inform (for example) future research questions, research priorities and to find local expertise. It will also emphasise the need for not just open data but also near real-time data to drive evidence-informed decision-making. Future activity can include producing a 'version 2' of this hub, approaches to automatically update sections of the hub and automatically run analyses, and scaling up the methods to other countries or settings. This analysis can support development of recommended minimum standards/datasets for analysing research portfolio, develop guidelines for countries to better review and manage these portfolios to identify priority questions and research gaps, and enhance local PI-driven hypotheses to take forward these priority areas. (iii) how you will monitor and evaluate your proposal, including success indicators. We will evaluate the proposal through the half-day stakeholder workshop, through further in-person meetings, and

ask for written feedback from the wider research and policy community. Portal usage and study website hits will be (anonymously) tracked using Google Analytics or similar tool. The hub will have a ‘feedback’ and 'bug report' button. Success indicators will include responses to questions around the usefulness of this current tool, and the receipt of suggestions for future use and improvements. A further measure of success will be the Ghanaian study staff having received training to boost their skillset, and to be able to train local colleagues in the sustainable use of reviewing the research landscape to find strengths and knowledge gaps, and use of appropriate software.
Decision **Not shortlisted**
Comment on decision from Wellcome This proposal was for a national-level resource that could be of value. However, the level of innovation proposed was limited. The application would also have benefited from a more detailed evaluation plan, for example identifying targets that would indicate success.

Title **OpenHeart Project – An Open-Source Research Community in the Field of Mechanical Circulatory Support**
Lead Applicant **Dr Jo Pauls**
Details of proposal (i) The research field of mechanical circulatory support (MCS) consists of researchers from research laboratories, universities, and companies around the globe. Research is often undertaken in isolation within each laboratory, limiting inter-laboratory collaboration and thus limiting the full potential of research data and outcomes. Additionally, many good research projects end abruptly upon graduation of the PhD candidate carrying out the research; it has been reported that an estimated 85% of all global research resources (not only in MCS) are wasted (Fig. 1). There is a clear need for improved collaboration and data sharing and subsequent improvement in research quality and outcomes within the field of MCS. The OpenHeart Project is an open-source research project which aims to improve collaboration among researchers, standardisation of research practices, open data sharing and education of emerging researchers within the field of MCS. It is targeted at undergraduate students with an interest in MCS, HDR and honours students undertaking research in MCS, and early career and experienced researchers, clinicians, surgeons, cardiologists and nurses also working in the field. Key entities of the OpenHeart Project platform are the OpenHeart website, data repositories, educational tools and networking capabilities (Fig. 2). The project team employed Atlassian's software solutions Stride, Confluence and Bitbucket to provide the main structure of OpenHeart. The core entity for data sharing within the MCS community are the open-source data repositories hosted within Bitbucket, which allow researchers to create unlimited repositories to upload their existing data. Specifically, current and former PhD candidates are invited to share their data to make PhD research outputs reusable. To incentivise data sharing and improve data findability and citeability, researchers will be able to create digital object identifiers (DOIs) for their data that has been utilised in peer-reviewed publications previously, through University repositories (e.g. Griffith University and UQ). Bitbucket allows researchers to collaborate in near real-time on new projects, while the in-built version control and option to work within multiple branches allows for co-design and development of research projects. To improve education and training of HDR students and early career researchers around the globe and especially in developing countries, a free series of MOOCs will be developed. The MOOCs will guide the integration and education of the next generation of researchers. Furthermore, an MCS Wiki will be developed as a knowledge base where most common terms in the field, equations and information can be defined and expanded on by the community. A tool will be developed to mine statistical data, with the resultant data available on the OpenHeart website. This openly displayed data will save research time when searching for relevant statistical information individually, while big data sets may be utilised for predicting future research needs. (ii) Currently a research community in the field of MCS consists through ISMCS, along with kindred societies. It is through this research community that we have been made aware of the many discontinued research projects and associated datasets that are stored in university archives. By incorporating silo-free research, resources will be combined, instead of wasted. However, until now, no online platform existed to facilitate access to these datasets, and improve collaboration and data sharing between these physical entities. Thus, OpenHeart uniquely supports an already existing research community by providing a platform to interact and share information on multiple levels. Through active participation within the OpenHeart Project, students and researchers will have the opportunity to shape a new generation of research within the field of MCS. They will be able to showcase their research and expertise through the creation of data repositories and content within the MCS Wiki, while being able to interact with their peers. To further incentivise active participation, OpenHeart will explore options to credit active members with continuous professional development hours. By sharing

existing solutions, expertise and equipment it will be possible to save research time and money while giving emerging researchers a head start. It is envisaged that commercial entities in the MCS space will become familiar with OpenHeart and will look to connect with researchers and projects via the platform. The OpenHeart Project also has strong potential to impact the broader research community, as its approach could be easily adopted in other research fields. (iii) With the support of the Wellcome Trust, three postdoctoral research fellows and twelve HDR students will be trained in how to work open, to ultimately support the creation of an open-research culture within the OpenHeart community. Project success will be monitored by tracking the number of sign-ups to the community via the website, and the number of community contributions (e.g. within the MCS Wiki pages and data repositories). Furthermore, we will monitor the number of repositories and DOIs created, with a target of 30 DOIs by the end of 2019. A first MOOC will be successfully designed by Q2 2019, with the first cohort of students (target of 20 students) completed by the end of 2019. Concepts for additional MOOCs will be demonstrated at ISMCS 2019 conference in Bologna, Italy. To further expand the reach and impact of the project, the progress and experience of OpenHeart will be written up for a peer-reviewed open-access publication.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal sought to extend an open source platform. The activities and methodology were not clearly described, and so the potential impact of this proposal to transform health research through openness was unclear.

Title

Development of an open source, accessible platform for large-scale hallucinations research

Lead Applicant

Dr Peter Moseley

Details of proposal

Hallucinations are one of the defining features of psychosis, can be extremely distressing and socially isolating experiences(2), and are refractory to anti-psychotic medication in around 25% of cases(3). However, as many as 13% of individuals in the general population also report hallucinatory experiences(4), suggesting a continuum of hallucination-proneness (or psychosis-proneness) across the population, with only some experiences becoming pathological. Yet, the evidence regarding the underlying neurocognitive processes of hallucinations is hampered by small sample sizes (a mean of 45 in non-clinical studies(5)), non-standardized methods, lack of replication, and few openly available measures and datasets(6). Improvement in this field is crucial for understanding what can make some experiences pathological, and whether cognitive interventions are therefore likely to be successful(7). As part of the International Consortium on Hallucination Research (ICHR), a working group led by the principal applicant, including leading researchers and clinicians around the world, has set up a multi-site study to counteract these problems. We are collecting the largest ever general population sample of participants tested on hallucination-specific neurocognitive measures, including source-monitoring, language lateralization, cognitive control, and verbal working memory, and psychopathology measures including depression, anxiety, trauma, and hallucination-proneness, programmed in jspsych (an open source javascript toolbox). Data collection is currently underway in 11 laboratories, based across five countries. The protocol, exclusion criteria, and analysis plans are pre-registered here: osf.io/cyu6j. We plan to roll our platform out to further research groups in subsequent phases of data collection. However, to achieve the greatest impact upon this research area – and change hallucinations research globally – further development to increase accessibility and openness is needed. The aims of the proposed project are, therefore, twofold: To drastically improve reproducibility and replicability in hallucinations research There has been recent concern regarding replicability and reproducibility in the entire field of psychology(8), and hallucinations research is no exception(6). The present proposal would allow us to vastly improve reproducibility by making our current internal platform public; that is, all tools and datasets openly available. We are planning to achieve this by using github and the Open Science Framework, and making all analysis available by developing detailed codebooks and R Markdown documents. This will allow for all measures and analysis to be fully reproducible by the wider research community, with open materials allowing direct replications by other teams. A fully open dataset would promote complete transparency within this research field, allow other teams to test novel hypotheses from the dataset, as well as allowing researchers to compare findings from clinical populations to a standardized, general population sample. To provide a standardized set of validated measures easily accessible to hallucinations researchers across the world A crucial aspect of this project is to make the central tools of hallucinations research accessible to a wide range of researchers at zero cost. Existing research into the neurocognitive mechanisms underlying hallucinations has largely been conducted in westernized, rich, well-educated countries where design and analysis tools are assumed to be accessible due to institutional resources. With an open and accessible platform, we would aim to use collaborative networks within the ICHR to extend research into low- and middle-income countries around the world who would only need a computer and internet access to use the platform. While currently the platform includes basic instructions and versions of the tasks, further development would aim to turn this platform into a 'one-stop shop' for hallucinations research in cognitive psychology, including: 1. a full online instruction manual regarding use of the platform and data analysis (both text and video 2. the functionality to choose specific measures to be used in an experimental session rather than running a pre-set order 3. the ability to access open datasets directly from the platform, with the option to customize variables 4. fully accessible and editable analysis scripts, with examples in R Markdown. All relevant documentation would be made freely available, including ethics applications. Successful accomplishment of these aims would be evaluated by monitoring uptake (number of research groups involved, size of resulting datasets), as well as downloaded datasets. We would develop a feedback form in which researchers could log issues and suggest improvements and additions. We would seek to influence practices in the wider research area by: running a 1-day workshop on ‘open science in psychosis research’, inviting researchers, clinicians, and service-users, including tutorials on open practices proposing symposia on ‘open science in psychosis research’ at national and international conferences (e.g., British Psychological Society, Schizophrenia International Research Society). We would communicate the availability of our platform through the Early Career Hallucinations Research network and feed our work back to the ICHR conference which will be recorded and made available online. Finally, a successfully delivered project would pave the way for similar multi-site approaches to understand other clinically relevant unusual experiences (e.g., delusional beliefs) and psychopathology (e.g., self-esteem, anxiety). By making this type of research available via our open platform we will make access to research more democratic and increase the reach of the way we gather information. This will be particularly advantageous for research in mental health, an area which has been slow to adopt open science methods(6).

Decision

Not shortlisted

Comment on decision from Wellcome

This was a well-written proposal from a strong team. However, the level of innovation proposed was limited and the evaluation plan would have benefited from more detail.

Title

Clinical Code List Inventory and Citation (CCLIC) Portal

Lead Applicant

Dr Jennifer Quint

Details of proposal

Our vision In the field of healthcare data, identifying suitable datasets to answer research questions and test hypotheses is difficult; data sources are many and metadata can be lacking. One specific challenge with utilising routinely collected data for research surrounds clinical code lists. For example, to identify a disease or condition, researchers must consider not only the terminology specific to the classification system they are using, but also, if using multiple or linked datasets, reconcile several overlapping clinical term dictionaries, and differences in published code lists (if these are published at all). In addition, each database may have specific code list capabilities unique to their provider. Currently, there are few incentives for academics to make code lists public. Code lists are infrequently added as supplementary material in publications, as researchers either don't see value in publishing them, or view them as their own intellectual property. Common resources such as Hospital Episode Statistics (HES) or Clinical Practice Research Datalink (CPRD) do not necessarily provide robust or open-access utilities for building code lists. Code lists, where published, are often fragmented, outdated, or difficult to find. Rather than publishing code lists with individual publications, there is an opportunity for code lists and disease specification methodology to be openly accessible, stored, referenced and/or indexed in one place, and independently citable as a standalone re-usable resource or starting point. We are in the process of undertaking a scoping exercise of existing resources – identifying current code list repositories (such as the Manchester clinical codes repository), both independent repositories and those that form part of publication processes, and we will evaluate them based on availability, accessibility, and current and future use cases. This work will inform this proposed project and illustrate the need to further develop a flexible open-access and open-source code list portal for academics across disciplines. Where existing resources are usable, we intend to create an index instead of importing code lists directly, thereby creating a unique DOI for code list assets.   Our aim To identify the facilitators and barriers for code list publication in an open access repository, and to assess the perceived value among the scientific community and data vendors for open-access code list publication. The long-term goal is to improve the quality and validity of electronic health record data research with appropriate academic credit being ascribed to incentivise the rapid dissemination of key domain knowledge to both research and industry users.   Target Audiences Primarily academic researchers from all disciplines who utilise healthcare data. Additionally, data providers, database managers and programmers using similar data.   Activities 1. Polling of subject matter experts – assess the facilitators and barriers to code list sharing and re-use among subject matter experts both from a research and data provider perspective. 2. Linking code lists to clinical conditions – link existing code lists to disease specifications, allowing clinical conditions to be searched for and related code lists accessed. Code list versioning will allow for the rapid prototyping of new and updated code lists across data sources.   3. Development of graph database – develop a graph database forming the back-end of our clinical code list portal that provides an innovative method of associating code lists with publications and dataset resources. Indexing and referencing existing data in place avoids the overhead of importing this ourselves. Leveraging Natural Language Processing (NLP) we can index topics of publications to create unique code list linkages. 4. Pilot portal development – develop a publicly available pilot clinical code list portal built upon a version-controlled online repository that resolves the barriers identified in step one and facilitates: reproducible research, easy to access data (FAIR), multidisciplinary collaboration and evaluation using new metrics. The portal will have the ability to correlate new and existing code lists, linking them to specific clinical conditions, publications, and other resources such as publicly available datasets. Publications resulting from updated code lists can then be linked to the original list(s) and citations updated. 5. Pilot portal evaluation – evaluate the portal's use among researchers at Imperial College London.   How your proposal will influence open research practices in your field or more broadly Creation of clinical code lists is a laborious task and time consuming task, requiring in depth research to specify even small code lists. This project will provide a unique means and incentives for clinical code list collaboration with an intent to modify the typically reserved culture of code list sharing. By collaborating with existing initiatives, such as the Manchester clinical codes repository and the Oxford metadata catalogue, we will develop a community around the portal, spreading unique domain knowledge and accelerating multi-disciplinary research.   How you will monitor and evaluate your proposal, including success indicators At the end of the development cycle we will consult focus groups and conduct questionnaire-based surveys to evaluate our portal. A key measure of success will be if we are able to overcome the barriers and emphasise the facilitators highlighted in our scoping exercise. Uptake and impact will then be monitored through randomly selected web-based surveys on users, along with upload, download and page view metrics.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The application was from a strong team, proposing to generate an important resource. However, it would have benefitted from more information about the planned incentives to encourage clinician involvement. It was also not clear to what extent the proposal would meaningfully advance open research.

Title

An algorithm-based browser plug-in to reveal if scientific articles have been refuted or confirmed.

Lead Applicant

Mr Peter Grabitz

Details of proposal

The vision Using Natural Language Processing and Deep Learning, Scite Inc. developed an algorithm that automatically classifies citations and indicates whether an article has been refuted, confirmed or merely mentioned by subsequent papers. Our vision is to reveal the impact of research through automated qualitative citation analysis. We therefore propose building a browser-based plug-in with the help of the Wellcome Open Research Fund that makes this novel technology and approach widely available and user friendly. We aim to develop an open source browser plug-in, that is consolidated with a web-based service and will smoothly integrate into researchers' workflow. Other existing plug-ins like Unpaywall or the Open Access Button are further proof of this approach. In the biomedical field the "PubMed" search engine is the central search tool. Hence, we will start optimizing the proposed plug-in for Google Chrome in PubMed. Extensive user-interviews and focus group meetings of scientists in different biomedical sub-fields will ensure a high degree of usability that is essential for a new tool such as ours to be widely adopted. Already conducted, preliminary user-feedback has shaped the details of our vision and the plug-in will have the following features (also see additional information): Reports showing the context of each citation in form of a text snippet and an automated classification into refuting/confirming/mentioning or unclear compromise the central feature. Key visualizations that show the distribution of mentioning/confirming/refuting citations for each article in a badge-like way next to each article. User driven feedback function: whenever an automated classification is deemed inadequate by a user, s/he will be able to flag and annotate the citation in question. Downloading reports in .csv and other open formats to ensure interoperability. The primary target audience for the tool include academics, scholars and clinicians in the research community. Moreover, drug developers, policy makers and science journalists as well as the general public have a vested interest in evaluating the trustworthiness of scientific results. We will approach the development of the plug-in in the following way (see additional information for timeline): Phase 1: Baseline user testing (2 weeks) We will conduct mock-up-based user testing (1-on-1s) with researchers of different biomedical subfields to verify our current ideas and sharpen details of our vision. Phase 2: Prototype development (3 weeks) Together with a consulting Product Manager we will draft the vision for a first prototype based on user testing. Together with an experienced web-developer we will develop a prototype plug-in. Phase 3: Continued testing (3 weeks) In a second round of user-testing we will trial the prototype and get first-hand user feedback on the implemented features. This testing cycle may be repeated several times whenever new features are integrated. Phase 4: Preparation of public launch (4 weeks) Design stage: Web design, creation of documentation on usage, public outreach, graphics and extended How-Tos. Finalization of development process. Phase 5 Launch of Public Beta-Version, dissemination and evaluation (12 weeks) Public launch of the Chrome plug-In. Web-based trainings run by project lead. Continued evaluation and monitoring. Influence on open research practices The suggested plug-in will influence open research practices in the following ways: a) Currently, the research process stops with a given publication after an article is accepted by a journal. Our academic system incentivizes researchers to publish a lot, regardless of reproducibility. Gernerally, the metric by which academic success is measured for a researcher is number of publications, impact factor of journals, and number of citations. Thus, there is little direct incentive to publish reproducible results. The proposed plug-in will make reproducible research more transparent and easily identifiable, thus generating a new way by which research can be evaluated. This could lead to a paradigm shift in research practices that focuses on reproducible results. Researchers will be e.g. incentivized to make data available and be more

specific in methods-section, because the uptake of their publication will be available for everyone to see. b) The plug-in will alter literature research practices, as it will allow searching "forward" from an article of interest. Currently, scrolling through references lists of an article to identify relevant studies is a common practice (searching "backwards"). Our approach will allow for identification of relevant studies that cite the article of interest (i.e. forwards) and these results will be categorized as confirming, refuting or mentioning the article. c) Through increased transparency of citation contexts, identification of citation coersion will be facilitated. Self- and circular citation patterns will be discovered more easily. Monitoring and evaluation In order to monitor the success of the proposal we will have a strict timeline (see Additional Information) in terms of development and product management. We aim to create a sustainable and user friendly tool. In service of this goal, we will integrate user-testing and focus groups into each step of our timeline. Focus groups will include researchers, drug developers and science journalists. Our success indicators include the following: a) During development: Total number of conducted user-testing and focus.groups (aim: minimum of 12 user-test per cycle) b) Individual downloads of the plug-in (aim: 5.000 within the first month after public beta launch) c) Report-feedback rate d) Number of exported reports through the plug-in
<u>Decision</u> **Not shortlisted**
<u>Comment on decision from Wellcome</u> This was an interesting proposal with potential to impact health research across multiple fields. However, there were concerns over the feasibility of the approach set out.

Title

Creating COSECs – the largest dataset of COmmercial SExual Contacts in the UK

Lead Applicant

Dr Luca Giommoni

Details of proposal

Vision Sex workers and buyers have disproportionate risks and burdens of sexually transmitted infections (STIs). They can create a core group for the diffusion of STIs and connect distant actors in the sexual network through higher rates of concurrency and risky sexual behaviours. A focus on the health of sex workers is especially relevant given the development of the inclusion health agenda in the UK. Sex workers continue to face social stigmas, discrimination, criminalization, violence and persistent health inequalities. Standard surveillance and outreach programmes for STIs prevention in sex workers and their clients are then likely to be ineffective and unsustainable. For better health promotion outcomes, more empirical work is crucial. Data on socio-sexual network, in fact, can inform effective interventions to reduce STIs diffusion. For instance, research has shown that targeted interventions on core-groups (a few very active persons) are more effective at preventing the emergence of epidemics. However, data on commercial sexual networks are almost non-existent. This proposal creates and openly shares the largest dataset of sexual contacts between sex workers and clients. We will extract this information from the largest UK online community dedicated to reviewing sex workers' services. These data provide new opportunities for understanding sexual contact patterns and developing STIs policy and practice. We can develop a picture of British commercial sex networks using client reviews. Specifically, we can establish a link between Client A and Sex Worker B, every time Client A posts a review about Sex Worker B (see Additional Information for an example of the final network). We can also identify sex workers and clients partaking in high-risk sexual behaviours such as unprotected intercourses, anal sex, etc. From the preliminary analysis of 1-years worth of data, we have already identified geographical clusters, mobile sex buyers and popular sex workers vulnerable to risky behaviours. We can then analyse the network and characterise key players for public health interventions. For instance, we can provide empirically-driven information for producing safe-sex campaigns targeting mobile sex buyers who might play a prominent role in the diffusion of STIs across distant parts of the country. The dataset contains over 50,000 reviews in more than 300 different locations between 2003 and 2017. Each review contains information about client and sex worker usernames, date and time, city, provider (e.g. escort agency, massage parlour), duration of the encounter, price paid, and three written accounts describing the venue, the sex worker and the intercourse. This interdisciplinary project will bring together experts in digital research methods, public health, social network analysis and sexuality. It will use new forms of data collection (i.e. web scraping and crawling) and innovative research methods (i.e. social network analysis) to deepen our understanding of commercial sex networks with implications for STIs diffusion. Objectives This proposal will add to the current discussion on STIs in three ways: It will form the empirical basis for researching commercial socio-sexual networks by collecting, cleaning, coding, integrating and openly sharing the largest dataset of commercial sexual contacts; It will develop recommendations for tailored interventions to prevent STIs diffusion based on the analysis of this new dataset; It will engage potential users and stakeholders to show how to use digital data to study and respond to STIs within the commercial sex work community. Open research We will make openly available to the research community via the SDSL's website this newly assembled dataset. While the data that we collected are public, researchers might not have the computational skills for developing software that extracts and compiles this information into a structured dataset. We have already developed the software and collected the data. We will also clean and code the dataset so that the researcher community can analyse it without investing time and resources in these tasks. By openly sharing this dataset we will enable other researchers to reanalyse these data, leading to new research discoveries and outcomes. We will also share the results of our analysis and make them available through two

different channels: Publishing pre-prints in socarxiv Publishing a 1-page research briefing for each publication. This will provide a user-friendly description of our findings with a focus on policy implications. We will ask researchers using these data to do the same. Monitoring of the proposal We will monitor the success of the proposal by: Counting how many times users download the dataset Counting the number of visualisations of the webpage Counting the number of citations of the dataset Conducting a survey among potential data users and stakeholders about the usefulness of the dataset in the end of project event. Monitoring ongoing public awareness campaigns driven by this data Activities: A1: Clean data from typographical errors and inconsistencies; A2: Develop a framework and coding over 50,000 reviews; A3: Integrate the dataset with other variables; A4: Build the British commercial sex networks through the data collected; A5: Analyse the role that different players have within the commercial sex network; A6: Use network models to explore the structure of commercial socio-sexual networks and STIs transmission patterns across the population; A7: Develop policy guidance on effective STIs responses; A8: Share data and results via the SDSL website.
Decision **Not shortlisted**
Comment on decision from Wellcome *The applicant opted not to share this information*

Title
GEM: translational software for outbreak analysis

Lead Applicant
Dr Chris Jewell

Details of proposal

The problem Responding to epidemics is a grave challenge of the 21st century as new diseases emerge alongside climate, environment, and social change1. Recent innovations in statistical approaches to epidemic models have provided cutting-edge decision support solutions for outbreaks2,3. However, these demand advanced mathematics, statistics, computing, and epidemiology from the researcher, putting state-of-the-art analysis out of immediate reach of many in epidemics researchers. In other cognate disciplines, domain-specific software has transformed the rate of applied research. Examples such as STAN for general Bayesian statistics, BEAST for phylogenetics, and Gromacs for chemistry, demonstrate the transformational benefit of discipline-norm software for research communities to utilise and build upon. For epidemics, no such software exists as a single resource: code is written anew for each new outbreak application. Ad-hoc software fails to make modelling assumptions explicit, and is difficult to modify and assess for correctness4. This development cost impedes reproducibility and adoption of new analytic innovations, consequently slowing down scientific research.   Vision Our vision is to create an effective software tool, combining flexible model specification with modern inference and simulation algorithms which bridges the gap between epidemiology, mathematics/statistics, and computer science.   As shown in the accompanying figure, our aim is that the epidemiologist should focus on design, critique, and interpretation of epidemic models.   The specifics of selection and implementation of statistical algorithms appropriate to the model should be left to the computer, much as in areas such as generalised linear models. In this project, we will develop "GEM": a prototype "General Epidemic Modelling" language containing core functionality to allow rapid development and communication of epidemic models. GEM will: be intuitive to engage experts and non-experts with basic training;   be expressive, allowing rapid and flexible model construction;   have in-built parameter estimation and simulation algorithms;   make best use of available hardware for maximum performance;   be extensible for future development.   Activities We will address our aim through three main activities: Epidemic modelling language. Achieving a clear, concise language to specify epidemic models is key to maximising the GEM's utility. With our existing collaborators (PHE, APHA; Cambridge, Warwick, Melbourne, Massey Universities), we will determine a basic core of language features necessary for expressing individual-level, meta-population level, and homogeneous population representations.   Algorithm implementation. We will build on the highly successful open-source Bayesian probabilistic programming library PyMC3 (soon to be PyMC4)5. This will allow feasible implementation of GEM within the project time-frame, augmenting PyMC3's existing algorithms with additional epidemic-specific code only where necessary.   Dissemination. Ahead of our start date, we will use IDDConf-2018 (UK 140-delegate epidemic modelling conference) to present our project, conduct a confirmatory requirements survey, and recruit software testers. As an extension to PyMC3, we will announce GEM via the PyMC3 discussion forums allowing us to reach a large and active community of applied epidemiology and biostatistics researchers.   Target audiences We target 3 user classes: Research epidemiologists: an epidemic modelling language will allow domain-researchers to focus on model design, critique, and interpretation without needing to acquire advanced mathematics and programming skills.   Methods researchers: a modular architecture will allow methodologists to provide new features and improved algorithms. This will accelerate translation of methodology into high-impact applications, and allow extensive code review and testing.   Teachers and students: GEM will provide an epidemics teaching platform that focuses on modern epidemic modelling principles, crossing the computer programming and mathematics skills barriers. Influence on open research practice Transparent communication and trust is key between these communities, and open research is central to this. GEM aims to influence open research practice in epidemic analysis through 5 key principles: Transparency of epidemic modelling by providing an intuitive, concise way to describe a model, immediately exposing all modelling assumptions to cross disciplinary barriers; Increasing reliability and reproducibility of results by using a consistent set of fully tested inference and simulation algorithms that operate on user-specified models; Fully open source (MIT or similar licence) code which is publicly available via source-code repository at all stages of development promotes engagement from both user and developer communities; Facilitating a community of applied users and methods developers engaged in dialog increases productivity and innovation in scientific research; Sharing of model descriptions for comparison between countries, diseases, and research groups is promoted through a clear model description language, separate from the complexity of algorithmic code. Success evaluation Our key success indicators are: A modelling language expressing a test set of epidemic models: individual-level, meta-population, homogeneously mixing; Inference and simulation algorithms comparable to complex hand-coded implementations of the models in (1); Engagement by the epidemics and biostatistics community through online software version control and discussion platform, and IDDConf, a UK workshop of 140 epidemic modelling experts held in September. At IDDConf-2019, we will deliver a presentation on our progress, with a software tester questionnaire measuring success in epidemiological and productivity terms6. This will of course influence how we take GEM forward after this initial development phase. References Dye et al. 2013. WHO Tech. report. Jewell et al. 2009. Bayes Anal. 4:465–496. Kypraios et al. 2017. Math. Biosci. 287:42–53. Muscatello et al. 2017. Emerg. Infec. Dis. 23:e161720. Salvatier et al. 2016. PeerJ Comput. Sci. 2:e55. K. Kennedy et al. 2004. IJHPCA 8:14

Decision

Funded

Comment on decision from Wellcome

The was an ambitious application to generate innovative software. The commitment to advancing openness was clear throughout the proposal.

Title

An Open-Source Database for Predicting Pharmacokinetics

Lead Applicant

Dr Joseph Standing

Details of proposal

Vision: This proposal aims to create a web resource to host standardised datasets and model code to bring a big data approach to modelling of dose-concentration-response relationships (pharmacokinetics and pharmacodynamics (PKPD)). The applications of this knowledge will range from artificial intelligence and machine learning to evaluation and further development of clinical pharmacological models. This will contribute to reduced drug attrition through candidate selection based on predicted PK and reduced animal use in preclinical PK. The web resource will grow from this initial work to encompass extensive PKPD big data resources. Background to first project: Model-based approaches are crucial in drug development. Getting the dose right for personalised therapies is vital: First in man dose (even biologics, TGN1412 Phase I failure was in part due to a 100-fold over-dose) and predicting the dose in special populations (children, organ dysfunction) are of particular importance. In 1973 Prof Malcolm Rowland of the University of Manchester defined clearance (CL), which has become the most important concept in PK determining steady-state concentration and area under the curve (AUC). Predicting CL for a new compound determines whether it will be possible to achieve therapeutic concentrations in humans. CL prediction based on chemistry or systems pharmacology models have been tried. The most widely used physiologically-based PK (PBPK) model is SIMCYP (University of Sheffield spin-out). PBPK models are useful to a point, but often do not predict CL of a new drug well (within 2-fold is often considered "good"), may only provide a point estimate, and can fail to predict PK in special populations. PBPK models are data hungry (require experimental inputs over and above drug chemistry), and users cannot access source code. We argue PBPK models have reached close to their maximum potential and there are unlikely to be significant further improvements to PBPK CL prediction. Throughout science, as systems become more complex, new phenomena emerge that are not predicted by reducing the system to its component parts. There are now tens of thousands of published human CL values, giving the opportunity to study emergent phenomena on these big data. We will collate all published CL values along with patient covariates to use as a learning dataset to predict CL of new compounds or in special populations. Aims: To create the platform and populate it with its first dataset, a list of CL values from papers identified in pubmed, covariates associated with predictions, number of patients, number of samples per patient, data analysis approach (popPK or non-compartmental), and the drug's physicochemical properties. There are too many published PK papers to do this manually, and the database must grow automatically. Hence artificial intelligence scientists at BenevolentAI will collaborate to automate this text-mining task using supervised machine learning. Target audience: Academic clinical pharmacologists and applied computer scientists, the pharmaceutical industry, and medicines regulators. Activities: 1. Develop a learning database of CL values. We have already collated 300 such values, this will be manually augmented to circa 3000. 2. With the help of experts at BenevolentAI, develop an automated text mining algorithm (supervised learning) using the dataset in part 1 to automatically extract CL and covariate information from published papers. 3. Add chemistry information 4. Develop website 5. Publicise: workshop at PAGE and ASCPT meetings, email through ISoP, NMUsers, ddmore, PKUK and pharmPK networks. 6. Kaggle competition to develop algorithms to predict CL with data from AstraZeneca. Influence open research practices: Availability of standardised big data is inseparably connected to successful development and validation of machine learning algorithms. Creating a standardised big dataset for CL will introduce a new paradigm in model-based clinical pharmacology that has to-date been the preserve of chemical/drug discovery focussed databases. The community of quantitative pharmacologists interested in machine learning will be fostered through the a searchable list of names of registered users, with the aim of kick-starting collaborations. A project showcase will link to GitHub algorithm/model code, the first of which will be the text mining and natural language processing algorithm, developed in partnership with BenevolentAI to create and automatically update the CL database (beyond this grant). Future development: Expansion of the database to other PK parameters such as (apparent) distribution volume and absorption parameters and potentially PD of specific diseases. The community to be naturally synthesised around the project (i.e. through Kaggle competition) will generate novel insights in PK processes and novel analysis algorithms resulting in new grant applications. Monitoring and evaluation: Citations to the article describing the database. Free registration allows users to download the database/access plotting/basic analytics, and thus generate a record of name, location and organisation, along with monitoring user activity. By registering for the website and logging on, users will consent to be on a list of other users (name, institution and country) and via the website, send messages to other users. This fosters collaboration whilst retaining user privicy (same system operates for PAGE meeting attendees). Success indicators: 1. Website fully functional with pubmed updating system working 2. Number of citations to website and accompanying papers 3. Number of registered users, site accesses and data downloads 4. Attracting further funding/collaborations to extend data content and algorithm development.

Decision

Funded

Comment on decision from Wellcome

This was an ambitious application to accelerate drug discovery and development through an open research approach. The assembled team was strong and the proposal had good potential to impact health research.

Title

Repurposing genetic diversity information from gnomAD for a better representation and analysis of the human proteome in neXtProt

Lead Applicant

Dr Lydie Lane

Details of proposal

neXtProt (www.nextprot.org) supports applications relevant to human proteins and constitutes the reference knowledgebase for the HUPO Human Proteome Project (HPP) (PMCID:PMC5872831). It combines manually curated data from UniProtKB/Swiss-Prot with quality-filtered data at the genomic, transcriptomic and proteomic levels with fully traceable data provenance. It provides a powerful advanced query system based on semantic technologies that allows to perform precise queries on data both in neXtProt and in other SPARQL endpoints. Over 150 customisable sample queries are provided. neXtProt annotations are freely available under the Creative Commons Attribution (CC BY 4.0) and software code is open source and available on GitHub (PMCID:PMC5210547). neXtProt currently reports 5.7 million single amino acid variants (SAAV) including somatic (30%) and germline (70%) mutations. They are displayed in the sequence view of protein entries, as shown for MSH6 in Figure 1. Figure 2 presents examples of complex queries that combine variant information with other data in neXtProt. Using literature reports, we started capturing structured information on the phenotypic effects of SAAVs and we display this information in a dedicated neXtProt view (PMCID: PMC5413847, PMCID:PMC6042458), as shown for the MSH6 Ser144Ile variant in Figure 3. Such information is currently available for 103 human proteins (sodium channels and familial cancer genes) of interest to the clinical community. Expanding our annotation activities to other gene families of clinical interest will be the focus of specific grants. Combined with other information available on these genes in neXtProt, these annotations serve as the foundation to develop predictors of pathogenicity for novel variants identified in these genes or their close paralogs. Robust frequency data would allow these predictions to be checked and refined. Variant information is also used by the neXtProt peptide uniqueness checker we developed for the HPP community (~500 researchers world-wide) to validate the uniqueness of peptide to protein matches when analyzing mass spectrometry data (PMCID:PMC5860159), as illustrated in Figure 4. In the absence of frequency data, this tool considers all SAAVs as equiprobable, resulting in a high number of rejected matches. The Genome Aggregation Database (gnomAD) (http://gnomad.broadinstitute.org), developed by an international coalition, provides robust population frequency estimates (PMCID:PMC5018207), as shown for the MSH6 gene in Figure 5. Integrating gnomAD information in neXtProt would allow to propose new functionalities useful for life scientists, proteomicians and clinical geneticists. More specifically, we propose to: Convert gnomAD frequency data to RDF format so that it becomes interoperable with neXtProt and other RDF-based resources, and make it available on the neXtProt platform. Should the data already be available in RDF from another source such as med2rdf (http://med2rdf.org/), we will use that source. Publish at least 5 new SPARQL query examples illustrating the use of gnomAD frequency information. For example: "Variants with a frequency of more than 1/500 affecting known phosphorylation sites, with their associated frequencies". Adapt the neXtProt uniqueness checker to use the gnomAD frequency data. Display the gnomAD frequency data in the neXtProt sequence and phenotype views in an interactive manner. Users will be involved at several steps of the project: At LS2 annual meeting (Zurich, Feb 2019), we will install a booth and get detailed feedback and suggestions from users from various fields

At HUPO2019 (Australia, Sept 2019), we will discuss potential necessary adjustments before public release of the new functionalities of the peptide uniqueness checker

At a clinical genomics conference (TBD, late 2019), we will discuss the potential use of neXtProt data to improve variant pathogenicity predictors The new functionalities will be made available on neXtProt at the February 2020 release at the latest, and presented as a paper

in the 2020 NAR database issue. Usage of the repurposed data in neXtProt platform will then be monitored using Google analytics. Figure 6 shows the envisioned timeline for the project, starting Jan 1 2019.
Decision **Shortlisted, not funded**
Comment on decision from Wellcome *The applicant opted not to share this information*

Title

Improving collaboration using findable, accessible, interoperable and reusable (FAIR) functions for scalable and reproducible research computing.

Lead Applicant

Dr Aleksandra Pawlik

Details of proposal

The aim of this proposal is to develop infrastructure and templates FAIR functions for programmatical data analysis and visualisation. FAIR functions will apply the FAIR principles (https://www.go-fair.org/fair-principles/) to research computing functions. The functions and supporting tools will enhance collaboration between researchers using primarily spreadsheets and their colleagues using programmatical open research methods, such as data manipulation using Python or R. The spreadsheet-based researchers will be able to apply functions written in programming languages commonly used in open research. We will develop infrastructure and a process for other researchers to contribute and curate functions for data analysis. We are partnering in this project with the group of microbiology researchers who primarily use spreadsheet-type proprietary software for their work. They are aware and appreciative of open science approaches, however adapting these existing tools into their workflows is a steep learning curve. The FAIR functions are a solution not just for them but for many researchers who are in a similar setting. We also target the communities who want to share their best practices around open and reproducible research by sharing the source code which they develop. Our activities will involve both technical work, user-focused workshops, community engagement, and outreach. The project will start with a workshop run with the microbiology researchers to define the scope of functions most applicable for their work. The outcomes of the workshop will directly inform the development of the infrastructure for the FAIR functions in Stencila, as well as the library of functions itself. We will keep a tight feedback loop during the development keeping the researchers from the Bioluminescent Superbugs Lab up to date on progress and at the same time supporting them to adapt the FAIR functions and approach in their research practice. The researchers will be also the first beta-testers. We will also run a second workshop which will include members of the the community of contributors of FAIR functions. FAIR functions will help researchers make their work more transparent by exposing the code underpinning their data analysis and visualisation. Research done with FAIR functions and Stencila will be portable and archivable. Based on previous experience (https://o2r.info/results, http://stenci.la/blog/2017-07-docker-with-strace/), we will extend Stencila's capabilities to provide user-level access to different levels of runtime specification and snapshots. Findability of functions will enable users to search for functions via both the Stencila Hub and Desktop. Users may use a built-in search form, point to a local directory or archive, or directly provide a code repository reference (e.g. a GitHub or GitLab project). The latter allows easy collaborative development, access to both stable and latest version, and complete code transparency. The functions registered with Stencila will be loaded automatically together with the relevant documentation. The search and reviewing functionality will leverage the APIs of existing free and stable platforms for code hosting (such as GitHub). We will take advantage of the existing mechanisms for tagging, labelling and release management. Accessibility of functions will be ensured by making their integration into a document or spreadsheet possible from within the Stencila UI. The available functions will be suggested in a meaningful way, i.e. only functions with a suitable input data type (e.g. a cell with a string vs. a range of cells with numbers) will be applicable, and will be configurable with a simple GUI form generated from the function source code and documentation. If needs be, a researcher can inspect the code and functions used. Thanks to Stencila supporting conversion between a variety of formats often used in open research there will be no vendor lock-in, which is crucial for a potential update of FAIR function libraries and Stencila components by third parties. Interoperability of functions has three aspects, the most important one connecting different programming languages via serialization and shared types and via file formats with automated converters. This seamless interoperability of FAIR functions lets scientists stop worrying about technology and focus on applying the most promising workflow, independent of the specific implementation language or data model. Under the hood, function libraries are well-defined extensions of existing software packaging mechanisms of the respective language. Function metadata is used to advertise suitable inputs, which are applied for context-aware suggestions in the Stencila UI. This specification enables a second aspect of interoperability: other editors or platforms can also embed FAIR functions. The execution infrastructure builds upon containerisation (Docker) for controlled reproducible environments and provides system interoperability. The FAIR function libraries are executed in the environments which are well defined and understandable both for machines and humans (Dockerfiles). Reusability of functions will be powered by a simple yet powerful specification of how function libraries are defined within existing packaging mechanisms for the respective programming language. The function specification allows developers of research-oriented R or Python modules to expose existing functionality as FAIR functions, or researchers developing new methods to wrap them in a usable way for collaborators. Stencila function libraries may use a complex infrastructure for packaging and execution, but they can just as well be executed outside of Stencila components on a researchers own machine.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal presented an interesting and elegant idea, which could be useful. However, the level of need and demand was felt to be unclear and so the potential impact of this proposal to transform health research through openness was unclear.

Title
Evaluating and Improving Open Science Practices in Health Research

Lead Applicant
Dr Dermot Lynott

Details of proposal

(i) The benefits of sharing datasets have been widely discussed (National Audit Office, 2017) although issues persist pertaining the legal and ethical ramifications of sharing datasets (Houtkoop et al., 2018; Mostert et al., 2016). Additionally, the role of study pre-registration in improving transparency and reproducibility has been well-documented (Nosek et al., 2018), but recent research highlights issues and inconsistencies with current practice (Hardwicke & Ioannidis, 2018; Panhuis et al., 2014). Thus, there are three primary aims in the current proposal: 1) identify the current state of open science practices (data sharing, registration) in health research globally and in the UK, 2) identify issues arising from the tension between open data initiatives and data privacy requirements and safeguards, including but not limited to General Data Protection Regulation/GDPR), and 3) develop practical guidance that seeks to address gaps identified in stage 1, and how to address perceived barriers related to GDPR. We detail these aims below. (1) We will follow the protocol developed by two of the project team (Towse and Ellis; see also Roche et al, 2015), where we sample a range of journals from across the health sciences and select individual articles for detailed analysis. The work of Towse and Ellis provides a proof of concept for the feasibility of the approach, and findings so far suggest wide variability in data-sharing practices across psychology (see Appendix: Summary, and pre-registration plans here: https://osf.io/pk2hm/ and here: https://osf.io/kqz53/). The proposal builds on this work, assessing the extent that health researchers have engaged in key open science practices, focussing on data sharing practices and use of study pre-registration: Data sharing: For each article we examine the extent that data is made available, the extent that datasets are intelligible - implementing peer-reviewed analytic procedures - and the extent that existing datasets contain information relevant to privacy. Study pre-registrations: Are studies pre-registered or not? Are pre-registrations made available? Are pre-registrations sufficiently restrictive to limit the outcome switching and inflation of false-positive findings? Such a systematic analysis of pre-registration forms will help identify common strengths and weakness in current practice. (2) Through our ongoing work we have identified recurring scenarios reflecting actual and perceived barriers to sharing data openly, specifically related to data privacy legislation. Scenarios include, "Does anonymisation matter?", "How can I share secondary/video/audio/qualitative data?", "What constitutes sensitive data?", "Can I retrospectively share data when I didn’t envisage this within consent forms?" We will conduct a qualitative survey of researchers active in the health sciences to determine a) perceived legislative barriers to open science, b) specific concerns/scenarios experienced by researchers, and c) established legally-compliant solutions developed by researchers which facilitate open science practices. In developing and analysing the questionnaire we will liaise with the university’s Data Information manager, and consult with a legal expert on how legislation applies to health research. (3) Following work packages 1 and 2, we will develop and deliver practical advice/training on how to incorporate open science practices in health research in the UK while also educating about the legal requirements in light of recent changes to EU legislation, whilst recognising the specific derogations already in place for research purposes. We will host a workshop aimed at active health-oriented researchers based in the UK. In line with Wellcome Trust principles, the workshop will be free to attend, with all materials freely provided and accessible (including presentation slides, tutorials, worksheets, worked examples of pre-registration forms etc.). We will also develop an online decision tree that will detail exactly what researchers need to do in order to engage in open science practices for a range of given circumstances (e.g., different types of data, anonymisation, etc). The target audience for these outputs will include researchers across the health sciences, PhD students, NHS and University research administrators, and health professionals who interact with researchers in

their roles. (ii) The project will capture the extent that health researchers are following recommended best practices concerning data sharing and study pre-registrations. We will identify (and address) perceived barriers regarding data protection legislation, providing accessible guidance for researchers. Despite the focus on health research, we expect that the workshop and materials created will be relevant to researchers across a range of empirical sciences and cognate disciplines (medicine, psychology, psychiatry, economics). Hosting a free workshop, and sharing all materials freely online will maximise the potential audience of the project's outputs. Furthermore, the benefits extend to the general public, not only through the potential for improvements to reproducibility of health research, but through ready access to findings and materials. (iii) The success the project will be determined by the completion of our predetermined deliverables, including collation of data, write up of findings, and development/deployment of open science guidance materials (see Appendix: Project Timeline). Outputs will be further evaluated through peer review of publications stemming from data collection activities, and qualitative feedback from workshop attendees and users of our materials. Each output will also be supplied with a DOI or unique URL, which will provide quantitative measures concerning citations, downloads, page views etc.,. Where feasible, we will incorporate simple user feedback of online material.
Decision **Not shortlisted**
Comment on decision from Wellcome This proposal showed commitment to advancing openness in research through creating a training resource. However the potential impact of this proposal to transform health research through openness was limited as there was no piloting proposed.

Title

CODE CHECK: A web service for independent reproduction of computations underlying biomedical research

Lead Applicant

Dr Stephen Eglen

Details of proposal

Background Analysis of data and computational modelling is central to many areas of biomedical research particularly with the explosion of data now available. The underlying computer programs are complex and costly to design. However, these computational techniques are rarely checked during review of the corresponding papers, nor shared upon publication. Instead, the primary method for sharing data and computer programs today is for authors to state data available upon reasonable request. Despite best intentions, these programs and data can quickly disappear from laboratories. A systematic study examining 300 papers published in 2016/17 in Journal of Computational Physics, a journal that promotes sharing of digital artefacts, found that only 5.6% made artefacts available [STODDEN2018]. Given that code and data are rich digital artefacts that can be shared relatively easily in most cases, and that funders and journals increasingly request/mandate sharing of resources, we should be sharing more. Vision We wish to build a computational platform, called CODE CHECK, to enhance the availability, discovery and reproducibility of published computational research. Researchers that provide code and data will have their code independently run to ensure their work can be reproduced. The results from our independent run will then be shared freely post-publication. Our independent runs will act as a certificate of reproducible computation to document that the research outputs could be replicated outside of the researcher's lab. These time-stamped certificates will include key outputs, including figures and tables, and valuable information such as the environment used to evaluate the code. Such certificates will help in the peer review process by showing reviewers that the code is available and works. We will work with several journals to design the system so that it can be used a pilot service by those journals. The long-term vision is that our service will be used by journals in biomedicine (and beyond), with financial support from infrastructure grants and journals. Aims Develop a workflow for receiving code/data from researchers and curating/storing it. Create environments to run code and store key research outputs (figures/tables/results). Build a website to curate the visible outputs from computation. Generate certificates of reproducible computation that capture key outputs, and confirm when and how the computation was run. Work with journal editors to integrate CODE CHECK into their workflows. Establish a pilot scheme to evaluate the usage of CODE CHECK in journals. Working prototype certificates are available at: . Target audience Although the approach is broad enough to apply across many areas of biomedical science, I plan to pilot the work within neuroscience. Beneficiaries include: Neuroscience researchers can test that their work is reproducible and have their results permanently archived. Readers of neuroscience papers. Our website will complement the research paper by showing independent runs of the code. Studying other people's code provides valuable additional insights. Journal publishers. By using our system, funders and journal publishers will not need to duplicate infrastructure. Activities We will build the CODE CHECK system and test it on case studies from previously published work. Feedback from collaborators will ensure that the system is technically appropriate and useful. We will liaise with journal publishers to evaluate case studies for suitable papers that are due for publication. We will host a training workshop for key researchers and editors to learn how to use the system (and provide further feedback). All our materials will be released as Open Source under MIT licence, and we aim to write publications about this project. Influencing open research We hope our system will encourage researchers to share data before publication, as it can confirm that results work outside their lab. In the likely case that programs do not work upon first submission, our early feedback will help reduce the chance of incorrect code/data sharing post publication. We hope to support Diamond OA journals such as the upcoming NBDT

(Neurons, Behavior, Data and Theory,). (Diamond journals are supported by community time and grants, and thus free for authors to publish in and free for all to read). Unlike most other journals, NBDT is committed to mandating all code and data be shared upon publication. Our system should have broad appeal; beyond the specific tools and simulators used, many areas of biomedicine would benefit from CODE CHECK. Monitoring and success indicators will include:
Number of submissions to system. Compute time used to generate certificates.
Number of downloads/accesses to certificates. The appendix lists challenges to overcome in building the system, which will also act as progress indicators. The level of reproducibility in CODE CHECK will vary considerably. At a minimum, we propose one key figure or table from a paper needs to be regenerated from the code, rather than the entire research article. (The choice of how much should be reproducible will ultimately be up to each journal.) Our aim is to provide a quality check that the provided code runs, rather than a full assessment of the correctness of the code. We will not guarantee that the code can be run forever, as this is simply infeasible; time-stamped certificates will stand as independent verification of the results around the time the paper was published.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a potentially very impactful proposal to produce a tool for checking code presented in journal articles. However, concerns were raised over feasbility of this work, and the proposal would have benefited from more detail on the indicators of success for this tool.

Title

Open Biomedical Citations in Context Corpus

Lead Applicant

Dr Silvio Peroni

Details of proposal

Vision of the proposal Citations are primary scholarly data that provide both provenance and an explanation for how we know facts, and are an important vehicle for the discovery, dissemination, and evaluation of scholarly knowledge. However, citation data are not usually freely available to access, they are often subject to inconsistent, hard-to-parse licenses, and they are frequently not machine-readable. Furthermore, current citation indexes contain no information about the number of times a particular work is referenced in the citing work, nor about the structural and semantic context of such references, i.e. from what section(s) of the article they are made, and the textual contexts of these in-text references. The aim of the Open Biomedical Citations in Context Corpus is to create a new open corpus that contains individual in-text references in the biomedical literature. Having data at the level of individual in-text references offers many new opportunities. It will, for instance, make it possible to distinguish between cited works that are referenced just once in a citing publication and those that are referenced multiple times. In addition, it will be possible to see which in-text references occur together (e.g. in the same sentence or the same paragraph), to determine in which specific section of the publication these in-text references occur (e.g. Introduction, Methods, or Results), and potentially, by textual analysis of the citation contexts, to retrieve the functions of citation – i.e. the reason why an author cites another work. The target users of our project are primarily the biomedical researchers themselves, particularly those that need to study the literature in their field systematically. The following activities will be carried out: We will harvest the full text and reference lists of articles within the Open Access Subset of biomedical literature hosted by Europe PubMed Central (letter of support attached as additional information) that are available in XML by using their RESTful API. We will extract the in-text references from the full text of each harvested publication, and store these in an expanded and extended version of the OpenCitations Corpus, which is an open database using Semantic Web technologies to record scholarly bibliographic and citation data in RDF. We will develop a set of web interfaces to explore and query the proposed Open Biomedical Citations in Context Corpus. For this, we will adapt existing OpenCitations technologies to provide search and browse interfaces for humans, and a REST API and a SPARQL endpoint over the RDF data to permit programmatic access. We will adapt the popular VOSviewer software tool for bibliometric visualization to make use of the proposed Open Biomedical Citations in Context Corpus. For example, it will be possible to visualize the strength of citations based on the number of their in-text references. Proposal influence Literature searching, literature reviewing, and bibliometric analyses are currently done mostly by using the data from proprietary closed databases, such as Web of Science, Scopus, and Google Scholar. The proposed Open Biomedical Citations in Context Corpus will enable the above-mentioned tasks to be carried out using an openly available citation corpus. This has several important advantages. First, the tasks can be carried out by anyone worldwide with no fee. Second, bibliometric analyses can be carried out and published in a fully reproducible manner, which is not possible using proprietary databases. Third, fully automated large-scale analyses can be carried out, which is not supported by proprietary databases because only small portions of data can be extracted from these databases at any one time, and even this requires a significant amount of manual work. Fourth, third parties can freely develop analytical services that make use of the proposed Open Biomedical Citations in Context Corpus. Compared to existing citation indices (either proprietary or open) that work at the level of references in the reference lists of publications, the proposed Open Biomedical Citations in Context Corpus will reap important benefits by containing open information at the level of in-text references. For instance, when using citation links to search for relevant literature, it will be possible to filter by the section in the citing publication in which a

reference occurs, enabling researchers to focus specifically on references related to, say, methods or empirical findings. These features will enable researchers to carry out tasks such as literature searching and systematic reviewing in a more effective and more efficient manner. Monitor and evaluation Monitoring and evaluation will be based on a number of indicators: the fraction of the Open Access subset of Pubmed Central from which information on in-text references has been extracted and indexed; the proportion of all publications indexed in PubMed for which in-text reference data have been made available in the proposed Open Biomedical Citations in Context Corpus; the number of tools that make use of the proposed corpus; and the number of times the proposed Open Biomedical Citations in Context Corpus is queried per month. Since the Open Biomedical Citations in Context Corpus will build on existing working services and tools - the OpenCitations technologies and the VOSviewer software tool for bibliometric visualization - its sustainability will be high.
Decision **Funded**
Comment on decision from Wellcome This was an interesting proposal from a strong team. The application was innovative and had clear and potentially wide-reaching impact.

Title

Advancing Openness of Lab Notebooks

Lead Applicant

No title Valerie McCutcheon

Details of proposal

We will take a holistic approach to tackling issues around open research practices by interacting with researchers and practitioners from all disciplines. The information obtained from initial research and workshops will guide our approach to piloting Life Science laboratory notebook software and driving community activity in sharing best practice. (i) Vision To reduce the risk of valuable health research data held in laboratory notebooks being lost to future researchers. To make historical data FAIR – findable, accessible, interoperable and reusable. To make health research data open and accessible to all relevant parties in a secure, easy-to-use format. To further strengthen institutional policies and supporting structures to ensure that research follows the highest standards of integrity. Aims We aim to: Perform a review of literature and current practices around open research data, with a particular focus on the use and storage of laboratory notebooks for recording research data. Consider the perspective of a wide range of stakeholders — within the institution and in the broader community — including researchers, research administrators, institutions, funders, and suppliers. Deliver practical advice and solutions to aid the sharing and preservation of relevant laboratory notebook information. Share our findings with the wider community through various channels. Establish a sustainable network for discussing and promoting best practice associated with the use of laboratory notebooks. Target audiences Wellcome-Trust funded researchers A broader audience of researchers across all disciplines Research administrators Research funders Service providers, such as Jisc Activities Review the current literature that addresses issues and attitudes around paper and electronic laboratory notebook management. Survey peer practice and concerns to determine best practice and to identify areas that would benefit from innovative, novel approaches to health-research-data management. This will include specific collaboration with other Research Organisations, including the Delft University of Technology, the University of Cambridge, and the University of Edinburgh, as well as the wider community as our workshops and communications will be targeted widely. Run workshops to explore current issues and attitudes around paper and electronic laboratory notebook management, identify metadata requirements, and identify possible process solutions. Workshops would include demonstrations and discussion around Electronic Laboratory Notebook tools. Digitise samples of paper laboratory notebooks, considering selection criteria for preserving some materials over others, options for storing, Wellcome Trust technical guidelines for digitisation, Wellcome Trust open tools and services, data protection, Digital Preservation Coalition guidance, and issues around sharing 'orphan' laboratory notebooks to reduce the risk of loss and make them accessible. Examples that we may digitise samples from include both live and prior projects e.g. The Alexander Haddow Zika Collection that was recently brought to prominence via a Wellcome Trust grant https://www.gla.ac.uk/myglasgow/library/collections/medicalhumanities/zika/ Review methods for community discussion and sharing of best practice of laboratory notebook management. Pilot an Electronic Laboratory Notebook tool with particular focus on life sciences: users will be drawn from a number of sub-disciplines and having different levels of experience of open research. Obtain feedback and write up the pros and cons and perceptions. Work with Microsoft Teams to test integration of Lab Notebook software and to support researchers and produce a case study on this in collaboration with Microsoft Teams Team and the Lab Notebook software provider RSpace. This innovation will join up the power of full collaboration (chat, video, audio, live file editing and storage) from anywhere in the world on secure platforms for health research – with the strength of a market-leading electronic lab notebook software. Work with CREATe (the RCUK-funded centre for copyright law at UofG) to clarify legal issues around ownership and copyright of laboratory notebooks. Publish case studies and share findings. (ii) Influencing open research practices This proposal will influence open research practices more broadly by identifying and sharing challenges and solutions to open research. (iii) Monitoring and evaluating the proposal Our project will have a project plan and we will apply PRINCE2 practices. Several of the project participants are registered PRINCE2 practitioners with a wealth of project management experience. We will have a project plan overseen by a Steering Group, and hold monthly Project Team meetings. Success will be indicated by: Delivery of discussions and workshops, and feedback received from attendees on the usefulness of these in fostering community and best-practice sharing Publication of reports and case studies documenting key findings and lessons learned Establishment of a community laboratory notebooks forum (or reinvigoration of existing forums) A post-project evaluation of our impact The project is low risk; the potential benefits will be practical. We have found that cross-stakeholder engagement to be very successful. For example, recent workshops from our project with Jisc and CREATe to explore barriers to dataset licencing led to better understanding, delivery of freely available guidance, and excellent feedback. The project team already has the network of internal and external contacts required to undertake the project as described.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to study electronic lab notebooks; opening these up would have good potential to impact health research across multiple fields. However, the level of innovation proposed was limited, and the proposal would have benefited from more methodological detail about the pilot of electronic lab notebooks.

Title

Reuse SDK for easy discoverability of additional research outputs

Lead Applicant

Mr Aadi Narayana Varma

Details of proposal

Aim & Target audience AIM: To aggregate AROs & curate meta data associated with a research article on to a single platform for easy discovery & reuse. Target audience: Researchers: can easily access all the AROs associated with the research article in a single place & it updates automatically with any new ARO deposition in the respective repositories. Publishers: Publishers can use the APIs to further enrich their content & allowing their readers to discover all AROs alongside of the publication & it will be updated in the real time with new ARO deposition. This can be looked upon as a tool enhancing reader engagement with the content Activities: Initial Three months: To aggregate and curate metadata information of all the additional research outputs(ARO's) associated with an article which are deposited on various platforms into a single easily reusable database which can be queried upon. Next 3 months: A query language for accessing the data by a machine, and human. Releasing an open API Sample Web interfaces and Browser extensions to establish use cases of the database for the community. Next 3 months: Testing the API, production deployment. Reaching out to pilot partners(Publishers and Institutes) for the initial pilot programme. Launching an open web platform for researchers to execute their queries and Browser extensions for finding the ARO's for a journal article. Last month: Publishing initial use cases of the tool by different stakeholders along with usage analytics. Case study on how the tool has benefitted the ECRs(Early Career Researcher) in their research. Influence on Open Research Practices: Easy Reuse and Discoverability: Tool enables easy discovery and reuse of the AROs associated with a research article. Enables an Institute/funder to assess open data compliance and also the impact of additional research outputs that are generated from a single publication/grant. Enable researchers to make their AROs more discoverable even though they are publishing their data/protocols after the original publication. Proposal Evaluation Initial Pilot studies: In our initial pilot study on 6 months data from Zenodo repository, we found that almost ~75% of the data deposited is not being referred in the main journal article. Hence, making it difficult for the readers of the article to discover the data that is deposited in Zenodo. Most of the plasmids and Gene libraries that are made available in Addgene by researchers is made post-publication and there is no reference in the main published article, that plasmids used in the study are available elsewhere. Monitoring & Evaluation FGR (Focus group research) - Engage with researchers in institutions to gain feedback and utility of this tool during initial development. Pilot study with institution - This will help to understand how institutional researchers & librarians are engaging with tool to advance their research & monitor AROs impact respectively. Pilot study with Publisher - This will help us to understand how publishers are valuing the AROs in aiding the reusability of the journal article, by enabling their readers to discover the AROs along side the article. Indicators to monitor success No. of search queries - for discovering AROs associated with a research article/grant/institute. Time spent by researchers on the platform to discover AROs No. of people who have subscribed for real-time updates whenever an ARO is made available associated with a research article. No. of API queries/hits. No. of time AROs are accessed either through a publisher platform/web interfaces No. of unique users using the platform.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal which could generate a potentially valuable and impactful ouput. However, there were concerns over the feasbilitiy of the approach described and the ability of the team to deliver this in the timescale indicated.

Title

Embedding FAIR principles into analysis workflows with data packaging: Costs and best practices

Lead Applicant

Dr Darren Dahly

Details of proposal

VISION, AIMS, and ACTIVITIES We aim to develop a teachable, widely-applicable method for embedding FAIR principles in data workflows using data packaging (please see the attached figure). Thinking of data outputs as packages (or discrete research objects) puts the emphasis on bundling data, code and related artefacts in alongside the minimal metadata needed for someone else to understand and re-use them with ease and confidence. The simplicity of the approach, which often involves simply including a few additional files in an existing folder structure, transforms the daunting, abstract topic of interoperable data stewardship into a problem that can be more easily addressed in familiar terms using known tools. For this project, dIfferent data packaging options will be field-tested by the Statistics and Data Analysis Unit (SDAU) of the Health Research Board (HRB) Clinical Research Facility Cork (CRF-C), in collaboration with the Research and Digital Services at the UCC Library. The SDAU (https://crfcsdau.github.io/post/) currently supports over 30 ongoing academic-led, patient-focused research projects in epidemiology and clinical trials. We are typically involved at every stage of these projects, but our primary contributions are in the areas of data collection, management, and analysis. The SDAU will thus provide an ideal, "real-world" testing arena for evaluating a data packaging approach. Our volume of work will provide opportunities for testing different options, facilitate accurate costings, and allow us to survey a variety of investigators about their views on FAIR and open science. Because the research we support is so varied, ranging from small observational studies to multinational clinical trials, we will also be able to demonstrate the applicability of our findings to the wider health research community. Our expertise in working with data will help us to identify the most practical, cost-effective means of embedding FAIR principles into our established workflows, serving as a useful example for others. Ultimately, we want to help others to bridge the gap between their existing data handling practices and the requirements of FAIR data stewardship, particularly the more esoteric challenge of ensuring minimally sufficient data interoperability. Our aims will be met through the following activities: An expert meeting to selecting relevant data packaging options for evaluation and to establish the degree of metadata description and semantic modelling that is useful and achievable.

Modifying existing SDAU workflows to incorporate the data packaging options.

Hiring a FAIR Data Steward who will be tasked the the day-to-day FAIRification of SDAU research outputs. Updating existing SDAU reporting processes to facilitate monitoring of these changes. Estimating the costs of adding FAIRification to existing data workflows. Developing and deploying a survey on investigator needs, attitudes, and perceived barriers for open science and FAIR.

Developing and publishing (in a variety of formats) a best practices guide for practical FAIR implementation using data packages. Presenting our experiences to the wider research community to enable similar investigations within other units, disciplines and domains.

INFLUENCE The outputs of this project will provide practical guidance, empowering applied-researchers from a variety of fields to incorporate FAIR data stewardship into their daily workflows. This will be disseminated online, through workshops, formal publication, and by engagement with professional health research networks (e.g. Clinical Research Coordination Ireland, HRB Trials Methodology Research Network). More broadly, this work will provide an important case-study to inform adoption of research data packaging approaches. Specifically, the findings will directly inform the work of several Research Data Alliance working groups of which UCC Library is already an active member. Similarly, through direct and open engagement (e.g. improved documentation, gap analysis, issue logging, pull-requests) with the maintainers of key data packaging initiatives, our experiences will influence the development of these projects and broaden their adoption. MONITORING Each packaging format will be evaluated against emerging FAIR Data metrics, and in relation to the following qualitative criteria: ease of learning; ease of teaching; ease of integration with existing statistical analysis workflows; quality of existing documentation; quality of existing tooling (e.g. GUIs, software libraries). Monitoring the effectiveness and limitations of the ‘data packaging’ approach when applied to actual project data is a key element of this project, including clearly demonstrating potential risks and costs. Existing SDAU workflows allow us to support a large amount of research relative to the resources available to us. We are thus risk-averse when faced with the prospect of changing them. Impact on existing SDAU activity will be facilitated by our existing reporting procedures, both within the CRF-C, and to our funder, the HRB. We currently track a number of metrics, including the number of projects initiated, progress on these, and research outputs. We also have an existing processes for opening and closing SDAU projects, that revolves around evaluating the potential for research projects we are asked to collaborate on, clarifying research aims, and assessing how satisfied our collaborators were with the support. We will thus be able to see how the implementation of FAIR practices impact on these.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a sound and feasible proposal, which could have some value in advancing the uptake of FAIR practices. However, there were concerns over achieveing user uptake, and the level of innovation and potential impact of this proposal to transform health research through openness were considered limited.

Title

Educational web-based crowdsourcing suite for neuroradiological assessment

Lead Applicant

Dr Carole Sudre

Details of proposal

Skills learning for neuro-radiologists requires repeated practice and extensive training in order to build automatisms and correctly assess images from a wide range of sources and appearances. In the field of neurology, image assessment may cover a wide variety of activities ranging from visual rating scales, developed in an attempt to systematically quantify the observations and their distance from normalcy to detailed element segmentation (such as lesion) going through counting and identification of pathology markers. All these manual/visual assessments performed daily in clinical practice have the merit to focus the attention of the radiologist onto specific organ properties and have the potential to provide invaluable training ground for the development of automated methods. However, such outcome is strongly dependent on appropriate training, feedback and score recalibration. The overarching aim of this project is to provide a high-quality, modular, web-based platform for neuro-radiologists including both theoretical learning and practical training application leading to crowdsourcing participation. Sections of the web-suite will also be made available to the general population, patients and their carers, informing and explaining subtypes of findings, and thus improving the perception of neurological conditions. Once a practical scheme has been appropriately validated, the trainee will be encouraged to assess additional cases/images in order to contribute to crowdsourcing efforts. Objectives:

Provide a web-based training suite for training radiologists to learn how to appropriately assess neurodegenerative and neurovascular burden through appropriate rating scales. Provide a tool for consultant radiological re-training and rating-scale score recalibration to minimise exposure bias and score drift effects. Provide an educational tool for the general public to learn about different neurological problems and how clinicians assess brain images. Provide a platform for quality-assured crowdsourcing of image labels. Figure 1 (supplementary material) provides the pipeline process from training to crowdsourcing Training avenues: The neurodegenerative part of the training suite will focus on atrophy scales (medial temporal, general atrophy) and provide insight on patterns of atrophy (identification of gradients) and their relationship to different pathologies thus widening the scope of the seeing dementia initiative (http://seeingdementia.ucl.ac.uk). For neurovascular pathology, emphasis will be put on three domains: the assessment of white matter hyperintensity burden severity using four classical scales expanding from the existing Visual Rating Tool platform (http://cmictig.cs.ucl.ac.uk/vrt); the identification and differentiation between markers of cerebral small vessel disease; and the segmentation of pathological findings. Since these assessments are heavily based on MR imaging, the web-based suite will also introduce the relevant MRI sequences and training in the identification of artefacts Training features: In order to optimise training and maximise knowledge retention rate, theoretical learning content will feature memorable example-based templates accompanied with physical and biological explanation of the observed training target. To consolidate the learning, each training topic will be associated with quizzes for the theoretical part and practical exercises of identification and rating. History of the results will be kept for systematic analysis and assessment of progression and provide suggestions on how to improve rater performance. Figure 2 (supplementary material) presents a suggestion of personalised feedback. Crowdsourcing: Once the training of a given skill/task (rating, identification, segmentation) is completed and evaluated, the trainee will be encouraged to contribute to a crowdsourcing effort related to the exact same task. By combining training and research crowdsourcing, the platform will provide a way of making trainees contribute to research after benefitting from free training, while ensuring the quality of the performed task. Planned expansion: This web-based suite is conceived to be modular so that other researchers whose area of expertise require regular training and calibration can also contribute to the platform, exploring aspects such as the assessment of longitudinal change, notions on vascular variants and pathology (e.g. stenosis), grading of tumour in neuro-oncology, imaging-based McDonald criteria for multiple sclerosis, and cases of rare neurological pathology. Influence on open research practice and beyond This web-based suite will allow neuro-radiologists to standardise their clinical practice, offering an easy way to compare visual and manual assessment, and thus evaluate inter and intra-rater variability. This will enhance research reproducibility and assessment quality. Furthemore, by providing practical exercises then translated into label crowdsourcing, this platform will allow for large-scale data collection while mitigating concerns related to individual rater performance. Project Monitoring and evaluation The evaluation of the final product will consider two paths: quality of training and quantity of completed ratings. The quality of the training will be assessed quantitatively from the evaluation learning curves, assessing how quickly a task is learned to reach a satisfactory level of performance, while quantity will measure how many trainees participate and accept to continue rating scans outside of the scope of the calibration phase. Qualitatively, a user survey will ensure that training is of sufficient quality and monitor expected areas of expansion.

Decision

Not shortlisted

Comment on decision from Wellcome

There were several good ideas articulated in this proposal for a community platform. However, as a whole, the proposal lacked focus and the feasibility of delivering this in practice was a concern, and so the potential impact of this proposal to transform health research through openness was unclear.

Title

"bims: Biomed news" from machine learning to expertise sharing

Lead Applicant

Dr Thomas Krichel

Details of proposal

You may think that a system as outlined in the summary is a pipe dream. But it exists. Thomas Krichel created it 20 years ago, in an area few people are aware of: academic economics. It's called NEP: New Economics Papers. It is for working papers only. In conjunction with other parts of RePEc it has done miracles to lift the working paper culture in economics to new heights. To build a similar system for the biomedical sciences, we need a bibliographic database to watch. PubMed gives us a head start. Next we need software that produces all new papers from PubMed every week. Thomas built one starting in 2014. At this point, it runs like a clockwork. Then comes the most complicated part. It is to build an interface that allows selectors to build weekly report issues. Thomas started on this in 2015. He developed the ernad software originally written for NEP. Now, ernad can examine the 22573, on average, papers that are new to PubMed every week. Ernad uses machine learning to provide the selectors with the most relevant papers in a ranked list. In early 2017, Thomas found the first selector, Dr Gavin McStay. He now directs the project. It is called bims: Biomed news at http://biomed.news. In early 2018, we started to recruit selectors. We have about ten. With this few selectors, we can really only say that we have a service prototype. The good news is that selectors like the system. They appreciate that the system is less time-consuming than PubMed searches and more precise than PubMed profiles. The median selector spends about ten minutes on the weekly task of composing a report issue. The not so good news is that we have found it more difficult to convey the open science nature of the project. We do not hide the open science nature, but it is not obvious. The output data is already available by rsync, a protocol not wholly familiar to laboratory-based researchers. Selectors will be encouraged to disseminate information about their reports through personal web pages, social media links and via other scholarly communication outlets. The selectors' participation in Biomed news is intended to be part of the larger community movement to support open access to scientific literature. The objective of the funding application is (1) to turn the prototype to a service, and (2) to morph it from a machine learning tool to an expertise sharing service. To elevate the project from prototype to actual service, we need to expand the number and diversity of selectors. Currently our selectors cover 0.1% of PubMed every week. Gavin thinks that it is best to have highly selective reports. The more selective reports are, the more reports we have to create to cover all of PubMed. For example, if a report brings out seven papers a week, we need at least 3000 reports, but probably many more as there will be overlaps. To recruit selectors, we want to hire an assistant who will look for email addresses of researchers to be potential selectors. A list of grantees provided by the Wellcome Trust would be a good starting point. We will focus on laboratory heads, post-doctoral researchers and experienced laboratory scientists at reputable biomedical research institutions all over the world, especially those who have demonstrated support for the open science movement. We will contact them via email and invite them to become a selector. We believe these individuals would show dedication to the service and spread visibility of the project. The support of the Wellcome Trust would increase credibility of cold-call email communications. Gavin will travel to specific international biomedical research conferences to announce and describe the service providing opportunity to open reports on site. In order to elevate the project to become an expertise-sharing system, we will set up email distribution of reports. We will construct homepages of reports where web visitors will be invited to subscribe. We will encourage selectors to build a subscriber base. We will also monitor selectors to ensure that they are doing the weekly editing on time. Our success indicators will initially be to recruit selectors of diverse topics to raise coverage. The number of recruited selectors from different research areas will increase the coverage of Biomed news selections. After this we will want to recruit as many readers as possible. Critics may suggest that

our project is flaky because we rely on volunteer selectors. This is only partly true. It depends on the selector. If the selector is a patient with a chronic disease, (s)he may be thought of as a volunteer. We will open reports for patients. But they are not our recruiting focus, academics are. Academics have to know the literature anyway. We give them a state-of-the-art tool. As readership of reports increases, name recognition benefits sets in. Selectors will be able to include selectorship as a service item in their CVs. With this in mind, we think of Biomed news as a rare constellation where pure self-interest, aided by sophisticated technology, maintains an information source that will be of great benefit to humanity.
Decision **Not shortlisted**
Comment on decision from Wellcome This application proposed to scale up an existing service, which uses machine learning alongside academic expertise to highlight new research in biomedical sciences. However the level of innovation and the potential impact of this proposal to transform health research through openness were limited, and there was no evaluation plan, for example to identify targets that would indicate success.

Title

The Public Domain Technology Review: shining a spotlight on open research tools for the life sciences

Lead Applicant

Dr Jennifer Molloy

Details of proposal

Our vision is enabling life science researchers to discover useful, open technologies to accelerate discovery and innovation underpinning improved health. This vision is a response to the growing number of expired patents and increasing dedication of technologies to the public domain through initiatives such as OpenPlant, Structural Genomics Consortium, BioBricks Foundation and other organisations generating 'born open' technologies. Many useful technologies such as fluorescent proteins, DNA assembly techniques, and enzymes are in the public domain but are neither easily discoverable nor accessible since they typically appear only on websites of originating institutions, are buried in patent databases or scattered in journal articles. Aims To improve discovery of public domain technologies for life science researchers via an online portal, to be named the Public Domain Technology Review To enhance the value of publicly accessible information through curation and crowdsourced domain expert knowledge of how open technologies can accelerate research and ultimately improve health. To facilitate research on the public domain in the life sciences through providing downloadable datasets and enabling user research. Target Audiences Life sciences researchers and entrepreneurs in academia and industry, especially those developing or using open technologies. Technology transfer professionals. Scholars interested in the value of the public domain or the impact of open technologies in research. We will target contributions making use of the team's extensive networks in these communities. Activities WP1: Participatory design [Months 1-2] A small group of ~10 stakeholders will be convened for a facilitated workshop to design user stories that will guide software development. WP2: Database development [Months 2-8] US and UK patent datasets will be obtained, including expired patents and patents that expired prematurely due to failure to pay maintenance fees. Additionally, the database will be seeded with open technologies developed by the team and focus group members. WP3: Curation interface and curator community building [Months 4-12] Curators will contribute by: Writing summaries for patents and adding metadata to enhance discoverability. Adding open technologies to the database. User stories and focus group feedback will guide development of an easy-to-use interface which attributes the contributors and developers of the tools. We will target communities that already have a interest in this space to reduce risk of low recruitment and retention. WP4: Public interface development [Months 4-8] The goal of this work package is to provide a well-designed and intuitive interface for accessing the database, guided by the user stories from WP1. This work will be informed by related efforts such as the interface created by NASA to promote use of technologies in expired patents and patents made available in the public domain (https://technology.nasa.gov/search/public_domain/). The work will also leverage the search for expired US patents created by researchers at Michigan Tech. Requirements for our interface will be led by user stories and will add functionality through filtering by keywords and IPC classification. WP5: Evaluation and dissemination [Months 10-12] See below for Monitoring and Evaluation activities. Dissemination activities will include creating a user guide and materials for conference presentations, as well as a report to be submitted to a peer-reviewed journal. Influencing open research practices in your field or more broadly Open research practices more commonly encompass ways of working and the sharing of digital, copyrightable research outputs such as text, data and software. However, there is increasing experimentation with opening up patentable outputs and particularly research tools. The Public Domain Technology Review aims to raise awareness of the value of open technologies, initially in the life sciences but with the goal of influencing other domains to adopt the model. We hope to increase use of open technologies, stimulate broader discussion and policy making on open research, highlight the role of patents in building the public domain, and encourage researchers to consider the question, posed by Wellcome Trust, of whether protecting or opening up intellectual property is the best strategy for maximising the impact of their research in particular contexts. Monitoring and Evaluation Monitoring will take place frequently during the project through: Measuring user contributions of open technologies to the collection [Target: 100 items from 50 contributors] Measuring curatorial activity and summaries added to patents [Target: 200 items from 50 contributors] Measuring usage data for the search portal [Target: 1000 users making a search enquiry] Feedback from software focus group, meeting at least once per quarter Evaluation of success in achieving the broader aims of the grant will take place on a longer term basis. Aims 1 & 2: Evaluated by a combination of a user survey, usage data (comparing use of whole vs curated datasets) and focus group feedback [Target: 80%+ to agree that the portal improves discoverability compared to other means or has interested them to look for public domain technologies for the first time. 75%+ to agree that domain expert summaries provided context and information that would not have been apparent on reading the patent abstract.] Aim 3: Evaluated by download data plus feedback from a prominent request to provide a reason for downloading the data [Target: 5+ researchers downloading datasets for research on open technologies or the public domain]

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal described a useful tool with relevance to health research across multiple fields, and a good evaluation plan. However, the identity and role of the curators was not clearly described, and there were concerns over the feasibility of this project.

Title

Metagram: A next generation tool for finding research outputs

Lead Applicant

Dr Cassandra Gould van Praag

Details of proposal

Vision While medical research around the world progresses steadily, the way in which we disseminate that knowledge has stagnated. Typical dissemination tools yield seemingly endless "yellow pages" style listing of results. These results make ineffective use of the human visual system, in its ability to see patterns, spot trends, and identify outliers. They also require researchers to use inadequate proxies to inform their understanding of the field (journal impact factor, etc.). The biases introduced by these proxies often prevent important outputs being incorporated into international efforts to achieve health benefits. In Metagram, we will develop an accessible and unified search space, and seed the incorporation of our methods into professional search platforms. Aims Our aim is to produce an interactive, graphical representation of research output search results, along with flexible and powerful means of interrogating those results (see attached Figures 1 and 2). This online platform will act as a proof-of-concept demonstration to inform the next generation of professional academic search engines, which will leverage metadata and graphical results summaries. The graphical representation will be intuitive to use, convey pertinent information, and support medical researchers in making reliably informed decisions, by reducing the cognitive burden involved in coherently abstracting information from research outputs. This will improve the efficiency in analysing search results, such that reliance on inadequate proxy measures will be alleviated. Target audience Metagram is intended to be used by anyone who has ever performed an online search of medical research outputs, including non-professional researchers. Greatest benefit will be brought to users attempting to synthesise coherent overviews of a topic (e.g. compiling literature reviews or writing introductory text), and those in the experimental planning stage, in identifying relevant material to inform design, identify trends or new directions for research. Activities Work package 1 (WP1): User interface (months 1-12) User-centred development work (UI/UX) will start with specification of user needs: summaries of output content; preferred descriptive statistics; the suitability of visualisation methods (e.g. the dendrogram structure, see attached Figure 2) to navigate results. Development will involve a series of "design sprints", to rapidly prototype and refine possibilities for the UI, with iterative UX feedback. Work package 2 (WP2): Research output search and metadata tagging (months 1-6) Metagram will leverage open Application Programming Interfaces (APIs) (e.g. the Europe PMC RESTful Web Service API, and tools developed by Content Mine http://contentmine.org) to send and manage user searches. Received results will then be screened against a classification database (e.g. the National Library of Medicine Medical Subject Headings https://www.nlm.nih.gov/mesh/) and tagged accordingly. Activities in this work package will focus on identifying and applying relevant existing open tools for this purpose. Work package 3 (WP3): Web-hosting platform (months 1-6) Metagram will be available online, free to access, and have optional log-in features to enable users to save and track search results. WP3 will ensure that the UI developed in WP1 and technologies identified in WP2 can be delivered to users in an accessible and effective manner. Work package 4 (WP4): User engagement and method promotion (months 1-12) We will undertake user engagement activities throughout the project to grow a sustainable user base for developer contributions. We will also actively engage with professional search providers (e.g. via PubMed Labs https://www.ncbi.nlm.nih.gov/labs/pubmed/) to demonstrate the power of our developed methods to improve the experience of search users. Activities will include promotion at conferences and national talks, and the identification of partners for dissemination. Influence on open research practices Metagram aims to make research output more ‘findable’ by adding extended and structured metadata. Our suggested metadata goes over and above the DataCite Metadata schema for the identification of resources (https://schema.datacite.org/) and proposes that a single freehand text descriptor is insufficient for effective retrieval of outputs. Metagram will shoulder the burden of appropriately applying metadata by using existing well defined ontologies, and remove this responsibility from the researcher. We will also showcase the value of community-lead UX research and engagement, along with the potential to expedite progress in research practice through ‘open by design’ principles. Monitoring and evaluation, including success indicators Short term indicators of success (0-2 years) will be inferred from activity on the Metagram website (e.g. the number of active users) and rates of "completion" in the user journey (see attached Figure 1). We have secured the agreement of community developers to provide web-activity metrics to act as benchmarks in this regard. We will also release our code and data at the earliest opportunity, so will have rapid insight of contributor requests and be able to track the early re-use of our methods or data. Medium term indicators of success (2-5 years) will be in the citation of formal publications and our software repositories via ciatbale DOIs, benchmarked against other similar projects. Long term success (6+ years) will be indicated by uptake of our demonstrated methods by existing and newly developed professional search engines.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal to make research literature searching more accessible. However, there were concerns raised about the feasibility of the proposed branching user interface.

Title

The development of a Generic Drug Database with their New Indications- Generic Drugs Repositioning

Lead Applicant

Dr Nibedita Rath

Details of proposal

(i) Vision: Repurposing Generic Drugs for New Indications: The development of a Generic Drug Database with their New Indications. A one-stop shop knowledge -hub comprising data from preclinical medicinal chemistry and biology of the small molecule to the data from the clinical trial and patient usage would enable informed decision making in drug repurposing of generic drugs and hence improve the chances of success. Aims: (1)The development of generic drugs database with their new indications: (2) Focus on Infection and Immunity: that help global neglected diseases and translational research (3) To encourage data sharing in the global neglected disease community by providing a shareable resource. Target Audiences: Our target audiences are (1) Scientists and clinicians doing neglected diseases/ infection drug discovery and development. (2) The non-profit organization working on NDs (such as TB, Malaria, etc.). (3) Patient communities, and (4) Students and scientists at Academic Medical Centres and Research Hospitals. Activities: During the course of this grant, we will collect all generic drugs and we will screen selected generic drugs using nanostring infection and immunity panel at TheraCUES Innovation Pvt. Ltd. The database will include (i) active chemical structure, (ii) Physico-chemical properties (iii) primary indication (iv) preclinical data both in-vitro and in vivo (v) mode of action (vi) clinical trial data from phase I, II and III (viii) safety indication from the clinical usages and (ix) Gene perturbation assay in different cell types to identify its implication in new disease indications. (ii) Influence on the Field: Repositioning opportunities exist because drugs perturb multiple biological entities (on and off-targets) themselves involved in multiple biological processes. As the drug discovery pipelines are focused on one disease of interest, a therapeutic application for a drug to other areas can be missed. This generic drug database would bring in additional implications of generic drugs in other disease indications. This project will be the first of its kind in the neglected diseases field and will provide the basis for further open and collaborative work in this field. (iii) Monitoring and Evaluation This project is very straightforward to monitor and evaluate, both during development, deployment and adoption. In the development phase, success will be measured by the number of data points generation from screening platform in multiplex assays. We are aiming to screen a few hundred compounds using inflammation and immunity panel and if possible would explore oncogene panel. During the deployment phase, success will be measured by the collection of compounds, samples and screening using nanostring platform curation of data and finally identification of few drugs that could be used in other disease indications Finally, during the Adoption Phase, success will be measured by the development of new users for the platform, the adoption by end users (who will either register for user ID. Uses of the database by more than 100 scientists/clinicians/researches in the first year will be considered success, with an expectation of growing this to over 500 scientists/clinicians/researchers in the following year.

Decision

Not shortlisted

Comment on decision from Wellcome

The proposal was felt to be potentially very impactful and targeting an unmet need. However, there were concerns how the information to populate the resource would be sourced - in particular the inclusion of labaratory work which is outside the remit of this scheme.

Title

The Developoment of an Equine Assisted Interventions Research Repository’

Lead Applicant

Dr jill carey

Details of proposal

Vision With members in over 45 countries worldwide, the Federation of Horses in Education and Therapy International AISBL (HETI) is a not-for- profit organization established in 1974. HETI’s mission is to aid in the collaboration and sharing of scientific and educational knowledge between organizations and individuals practicing Equine Assisted Interventions (EAI’s) worldwide. HETI is currently the only international organization acting as an umbrella group for all EAI’s. HETI have held 16 International Triennial Scientific and Educational Congresses and has an annual peer-reviewed journal ‘The Scientific and Educational Journal of Therapeutic Riding’ (SEJThR) with an archive containing over 20 years of publications. Many researchers and providers of EAI’s report inadequate access to peer-reviewed journals (Stroud & Hallberg, 2016) . In the field of EAI’s these are the primary sources for documenting current research and practice trend. The inability to access these contributes to a ‘science to service gap’ (National Implementation Science Research Network, 2016). This has implications for the development of practice and education in the field. The vision of this programme is to make research in the field of EAI’s more transparent, efficient and collaborative. In order to achieve this a 3- phase plan is proposed: 1. Make SEJThR Open Access(OA). 2. Develop a curated online repository that indexes and provides OA, peer-reviewed journals. a. Develop an international categorization system which enables the classification and provision of a digital collection of abstracts and full texts of OA scholarly researched EAI publications and abstracts of EAI publications in subscription journals. 3. Run a collaborative campaign with Federation members, educational and funding institutions to promote and develop the use of open research in the field using the OA repository as the first step. The target audience includes researchers and practitioners in the field of Neuroscience, Psychiatry, Psychology, Physiotherapy, Occupational Therapy, Speech and Language Therapy and third level Education Institutions. The secondary audience includes patients and clients who utilise current services provided as well as potential funders in this field of work. Proposed Implemented Activities: 1. Run an initial online focus group with Federation members, researchers and practitioners in the field to aid in ascertaining an international classification system for publications and methodology materials. 2. Develop an archiving strategy for data obtained, finalize technical requirements for the repository based on activity 1 and develop the web-based interface and database for systematization, storage and easy access to the repository. 3. Launch an open access journal and open access repository. 4. Launch an international social media campaign promoting the use of the OA repository and to explore the possibilities in the field of EAI’s if an Open Research approach was adopted. HETI’s social media campaign has the potential to develop a broad and deep engagement in research in the field. We believe social media has the potential to provide a high-return, low-risk science outreach tool in which researchers and practitioners can play a valuable role in promoting open research and accessibility. Some activities we have planned to generate engagement include: - Marketing and social media campaign to reframe the significance of the SEJThR as an open access journal through questions and answers, forum discussions, competitions, visuals and direct links to the OA repository. - Social Media campaign to generate engagement and discussion in relation to open research. It is envisaged that by adopting an OA approach in this field that it will pave the way to greater collaboration, understanding, and scientific merit as well as funding outcomes for future research. Providing a central base by which researchers can gain access to current research can aid in increasing citation and visibility and provides an intellectual history of the progress and pitfalls of EAI research which can be accessed by individuals worldwide. This could begin bridging the gap between research and practice that currently exists in the field of EAI’s. We believe that HETI is best placed to develop this research repository due to its’ well-established International standing, dedication to promotion of scientific research and education in the field of EAI’s as well as its’ capacity to provide guidance and support for practitioners. Success indicators will be monitored and evaluated in the following ways: 1. International user testing Focus groups in order to gain feedback and evaluation of classification system and ease of access for service users. 2. Monitoring and comparison of citations with open access articles in SEJThR . 3. Survey questionnaires sent to funding institutions, researchers and practitioners to ascertain increased visibility and usage. 4. Qualitative comparative analysis of feedback from HETI International Congress in Korea 2021 compared with feedback received in Ireland 2018 in relation to access and quality of current research. 5. Website analysis to ascertain trends over time Risks Caution needs to be taken in relation to selection bias and control for factors such as research quality, article type ,team and sample size as these are imperative to provide an accurate understanding of citation advantage claims. Controls such as citation advantages that may be associated with particular funders or within particular disciplines need to be taken into consideration.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to open up research in the field of Equine Assisted Interventions. The level of innovation proposed was limited, and so the potential impact of this proposal to transform health research through openness was unclear.

Title

HealthyR Notebooks: Democratising open and reproducible data analysis in resource-poor environments

Lead Applicant

Dr Riinu Ots

Details of proposal

Vision: To democratise open and reproducible data analysis among healthcare professionals and researchers around the world. Introduction: Traditionally, data analysis has required expensive software run on high performance computer equipment. This has limited the opportunities for those working in low-resource settings to drive their own research programme in areas such as clinical surgery. For even simple analyses, support has often been required from institutions in high-income countries and transfer of sensitive patient data across continents has been necessary. A long-held objective for our collaborators in low-resource countries is empowerment to lead the research agenda and work autonomously. GlobalSurg (globalsurg.org) is an international network of 5000 young surgeons working in 100 countries. We perform large-scale international research ranging from cohort studies[1] to randomised controlled trials. A central pillar of this collaborative is the democratisation of healthcare research with the provision of support for low-income country collaborators to analyse their own data[2]. We provide the highly regarded "HealthyR" training programmes in data science; however, these are currently delivered in-person within a computer lab. This proposal couples our development of a new "data notebook" analytics methodology with our capacity to deliver training at scale across an established international network. Aims: To empower health professionals in resource-poor environments to perform their own data analysis. To develop and adapt a lightweight, open-source, and scalable "data notebook" resource that focuses specifically on healthcare data analysis. To deploy, teach and evaluate this in the contrasting environments of Estonia and Ghana. Deliverables: A deployable package which contains all resources required to set-up "data notebooks" A deployable teaching suite in "data notebook" format, providing foundation training and materials for open and reusable healthcare data analysis. Field testing of deliverables 1. and 2. in Estonia and Ghana. Methods: A "Data Notebook"[3] is a software concept that brings together training materials, analysis code, and data analysis results into one document. This is intuitive and coherent, for instance, each graph or results table appears directly beneath the input. Therefore, rather than continually switching between training materials, analysis scripts and output windows, Notebooks bring everything into a single screen view (Figure 1). This makes the data analysis experience immediate, interactive, and rewarding. This has a number of additional benefits. Firstly, Complex data analysis can be performed on smaller screens, such as laptops, rather than full-sized computer monitors. Secondly, the software can be easily run on cloud-based facilities, meaning that where internet coverage is available, only a cheap computer with a web-browser is required to perform state-of the-art data analytics. Thirdly, the delivery of data science training becomes straightforward and easily delivered remotely at low cost. Audience and impact: The development of this technology is aimed primarily at healthcare professionals. They will be based in any country, but our focus is those in low-resource settings. This project has the potential for academic, economic and broader societal impact. By providing a data analytics platform and training researchers, we will contribute towards global academic advancement, building a worldwide community of data literate individuals who will help cultivate a valuable knowledge economy. This community may become self-sustaining, building knowledge to improve health and well-being, and providing the necessary information to help policy makers develop cost effective healthcare systems. Our group are one of the few pioneering patient and public involvement in sub-Saharan Africa, and these individuals help us ensure our projects are relevant and influential. We work within the framework of the Sustainable Development Goals – this project contributes to a number, but particularly "Industry, Innovation and Infrastructure". Monitoring and evaluation: Active monitoring is fully integrated into our development cycle (Figures 2 and 3). The first iteration of HealthyR Notebooks will be tested and refined in the UK. Further development and feedback will be sought in Estonia, followed by in-country testing and finalisation in Ghana. Estonia and Ghana were selected based on our deep-rooted long-term collaboration with several academic clinicians in these countries. A preliminary needs-analysis performed with PhD students in Estonia has helped shape this project and identify appropriate methods for evaluation. Ghana is an official international hub within our NIHR Unit on Global Surgery. We have published several data driven research papers with our Ghanaian collaborators, we are confident that with their input we can democratise open and reproducible data analysis in resource-poor environments. We work closely with a motivated Applied Statistician in Ghana who will be involved in development. Our initial evaluation has two main success indicators. Firstly, feedback from the two international cohorts (Estonia and Ghana), including specific questions on overall usefulness and relevance of the Notebook methodology and training materials; the ease of set-up, installation, and on-going use; and whether individuals achieve independence in using the platform for their own data analysis. In the longer term, collaborators will have access to our cloud-based service. This has built in analytics that will allow us to track its use by collaborators providing a clear measure of the impact of the project. Further evaluation will include the publication of academic papers and success in achieving grant funding. [1], [2], and Figures 1-3 can be found in the Additional Information. [3] https://rmarkdown.rstudio.com/r_notebooks

Decision

Funded

Comment on decision from Wellcome

This was an excellent proposal from a strong team. This proposal has the potential to impact health research, internationally, across multiple fields.

Title

STASH - Stimulating direct archiving and sharing of research data

Lead Applicant

Dr Zoltan Kekecs

Details of proposal

i. Rationale, aims, and activities Clinical data is an invaluable resource in enabling the advancement of healthcare, which is collected at great costs and at risk of patients. To ensure the sustainability of research, the stakeholders and participants of clinical trials have to be able to trust that data collected in the studies will be put to good use. The Replication Crisis led to a growing skepticism toward the credibility of biomedical research findings, and it is widely recognized that transparency is of key importance in regaining the trust in the field. Our vision is that in the near future the authenticity and trustworthiness of research data will be indisputable and that the data collection process will be transparent and reproducible for all clinical trials. We believe that direct data deposition can lead to achieving this goal. Direct data deposition means recording data directly to a version-controlled trusted third party repository instead of managing data locally. Thus, in this project, our specific aims are to: Aim 1: Develop tools for direct data deposition Aim 2: Make these tools easily accessible for researchers and developers Aim 3: Popularize the use of direct data deposition among clinical psychology researchers In accord with these aims, our activities will be organized under three project branches: In Branch 1, we will develop open source computer code allowing for setting up a direct data deposition service, and will demonstrate its use on a demo server. We will support different data collection methods (such as OpenSesame, Qualtrics, MTurk, and Google Forms) and repositories (such as Open Science Framework and GitHub) commonly used in clinical psychology research. In order to facilitate open research best practices, we will support saving data according to data structure specification standards (such as FORCE11's FAIR principles, or Brain Imaging Data Specification used in neuroimaging research), and the attachment of data code books. In Branch 2 we will produce a website of the project; a community support forum, where developers and end users can provide support for each other; and online learning resources including short video guides and step-by-step tutorials for end-users and developers. We will include demonstrations using different input formats and target repositories. Learning resources will also cover how to make data open immediately as it is being collected (born-open data), and the creation of automatized research reports that grant stakeholders continuous insight into the progress of the project (live, dynamic research reports). The main deliverables of Branch 3 will be a campaign to popularize the use of direct data deposition; registered users of the community forum; and official endorsements of direct data deposition by research organizations, professional societies, and/or funding bodies involved in clinical research. The campaign will target early career researchers, champions of open science in psychological science, and research organizations, professional societies, and funding bodies, because our experience shows that they are the biggest drivers of change facilitating open science innovations. We are in contact with the Center of Open Science and the Psychological Science Accelerator, both of which expressed its support of our goals, and its openness to potential collaboration. To maximize the impact on healthcare, we will focus our campaign on researchers and organizations involved in clinical research. The campaign will include: social media and podcast appearances conference talks workshops teaching the use of our tools press releases targeted promotional emails personal calls and emails within our networks. ii. Influence of the proposal on open research practices Our project will lead to a greater utilization of direct data deposition and an increased awareness of its benefits in clinical psychology. This will have a profound influence on open data practices on the field, since this solution enables: complete transparency about the history and completeness of the data real-time data sharing among collaborators and research sites seamless transition from data collection to open data easy audit and monitoring of clinical trial data sharing data of incomplete or failed research trials Some commercial clinical trial management software already have capabilities enabling real-time online version controlled data storage. However, the use of these software is largely limited to industrially sponsored research. By implementing free open source solutions that are compatible with free and commonly accessible data collection tools and repositories, our project will set the stage for large-scale dissemination of direct data deposition, and innovations that build on this technology. Compatibility with different data structure specifications and codebooks will also make the use of these open research best practices easier. Ultimately, our project will facilitate the buildup of quickly accessible, highly credible open data, that is machine readable and is easy to interpret and re-use. Furthermore, born open data and live, dynamic reports made possible by our solutions will empower researchers to give insight into their research in a previously unprecedented manner. This new level of transparency can revolutionize scientific collaboration and the involvement of the public in research. iii. Evaluation of success The success of the project will be monitored through the completion of its deliverables. See Table 1 for the detailed assessment plan (attached as additional information).

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting and innovative proposal, with the potential for impact. However, the application would have benefitted from more information about what the final product would look like and how it would integrate with existing researcher workflows, and how sensitive data would be handled.

Title

Octopus: Making scientific publishing serve science, rather than the other way round

Lead Applicant

Dr Alexandra Freeman

Details of proposal

(i) There are major issues in science (including widespread questionable research & communication practices, publication bias, lack of replication, poor methodology, inequalities in access, data hoarding, slow progress, wasted resources) and whilst there are many moves to tackle each of these, good practices are still not recognised and hence not valued by the current incentive structure. Publishing is the only way that work can be quality-assessed, so what is needed is a universal, user-friendly publishing system designed to facilitate and assess the true qualities of a research and their work (against agreed criteria of what makes a good research and good research). This information can then be clear to hiring/promotion/award/funding committees. The aim of Octopus is to produce a one-stop-shop system that works both to encourage and reward good scientific practice. It should provide the anchor for all the other Open Science technologies to attach to, since it will be non-commercial, 'unowned' by any institution or funder, self-governing, and built in such a way that allows developers to add their work to it and enhance the overall offerings for the scientific community (in the way that wikipedia works). Technologically, this platform is not too difficult to achieve - and once it exists, and works well for users, it will finally offer a 'better alternative' for both researchers and for those who control the incentive structures around them (funders and institutions). Then the work to encourage adoption and switching by both will be much easier. My aim for this 1-year project is to hire suitable technically-qualified people to help build such a platform - built around encouraging the qualities of 'good science' and 'good research practices', and around the needs of everyday, non-technical, diverse researchers. The key audiences that I think will initially embrace such a new platform are: - Early career researchers who can publish, and get credit for, work that they would struggle to get published in a traditional journal (such as a hypothesis, small data set etc) - Eminent researchers who are no longer concerned with their publication record and wish to encourage a better scientific culture - Open Science aficionados - Researchers in countries (or institutions) where access to traditional journals is limited and English is not their first language I also feel that the Octopus will also attractive to funders and academic institutions because: - the individual author pages, with their detailed metrics, will be useful as part of their staff selection and assessment processes (driving use of the platform by researchers in response) - the normalisation of publishing protocols before they are carried out (and their open review post publication) will be an efficient means for funders to assess projects. I plan to work with all these groups during development. However, I plan to work equally with researchers who are the 'squeezed middle' of the research chain - who do feel that Open Science is a 'luxury' unaffordable to them because they are time-pressed, do not have a choice of how to work given their need to stay on the publication treadmill, and who work in fields where competition is hardest. What is most critical is that the platform works well even for these users. The main stages of the project will therefore be: - Sitting down with those individuals and organisations working in Open Science whose work will directly or indirectly inform the design of the platform (discussed in Participants: Team) - a half-day workshop with technical advisors to spec the software and the skills necessary to build it. - hiring a software engineering team (probably 6 months of one person, and shorter period of others as required) - building a prototype - user-testing with a selection of 5 different groups of researchers, plus funders/institutions - iterative work to refine the platform - final evaluation to assess how well the platform works for users and encourages open practices (ii) This proposed platform is designed to produce a transparent assessment of the quality an individual's achievements, encouraging good and Open scientific practices and streamlining the research process hugely, as already described. It will be built to suit a wide range of users, making working in this way easy and rewarding, as well as encouraging the building of a new incentive structure (through quantification of the qualities that define 'good research' and 'good researchers'). Once built, it can then provide a basis for future work in establishing a platform based on this work as a bedrock for a new research environment. (iii) I start with the principles of what we in science want to encourage, as outlined, as well as thoughts of how to do that. During the design and development these will be constantly monitored through the user-centred design process - collecting qualitative and quantitative data throughout the project to ensure that it meets these aims. At the end of the project I will carry out a formal evaluation to compare the platform against traditional publishing in how well it meets the needs of researchers, how it encouraged good practice, and how well those practices are measured and valued by the platform.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a clearly written proposal, which sought to introduce a culture shift with a new publishing system. However, the level of innovation proposed was considered limited, and the proposal would have benefited from more detail on how user uptake would be ensured.

Title

Simplifying the use of computational workflows across different clouds through Galaxy and Kubernetes container orchestrator

Lead Applicant

Dr Irene Papatheodorou

Details of proposal

We propose to use state of the art cloud technologies and standards, coupled with open-source tools and workflows, to cater for the need in the life sciences for reproducible and computationally scalable data analysis, all through user-friendly workflow interfaces. The irruption of cloud computing brings opportunities in terms of scalability and flexibility, but also technical challenges tied to migrating cluster/HPC-based analysis to the cloud. With conventional University/Institutional computational resources, researchers find pre-configured hardware/software infrastructure – e.g., cluster, database servers, etc. – that the local IT and/or Bioinformatics services provide (based on institutional overheads). On the other hand, when working on the cloud, especially under an Infrastructure-as-a-Service (IaaS) model, researchers have to deal directly with low-level resources (CPU, RAM, disk, network throughput) as no pre-configured clusters or infrastructure are directly available. This new expertise gap, normally filled in industry by DevOps, is highly sought after and many times too expensive for academic institutions to afford. The work proposed through this funding aims to close that gap for researchers interested in using Galaxy workflows in the cloud by 1) automating the process of deploying a working Galaxy instance on top of the major IaaS providers, and 2) making the process configurable so that the deployed Galaxy platform can be customized for the various -omics application domains. The proposed setup uses analysis tools encapsulated as Docker containers running on the Kubernetes container orchestrator (https://kubernetes.io). This approach avoids the tool installation problem, often at a cost for researcher time because of software dependencies, versioning problems or unfamiliar platforms. Installation tasks are thus reduced to a single command invoking the creation of the container. This complexity is further reduced by leveraging efforts like Biocontainers, currently producing containers for hundreds of Bioinformatics tools and making them available for free to all the community. Thanks to past work in the community and the contributions of key collaborators in this proposal, such containers can be used out-of-the-box in Galaxy inside Kubernetes. All the proposed work will be open-source and freely accessible, maintained, documented and disseminated. This will enable scientists with varying levels of bioinformatics support to diversify their portfolio of computational solutions and to have easy access to a truly elastic solution, applicable from proof-of-principle analysis on their laptops to large-scale analysis on the cloud. In brief, the main aims of this proposed work are to: A1. Modularize the existing integration between the Kubernetes (https://kubernetes.io/) and Galaxy to improve its sustainability and promote its adoption by the Galaxy development community. A2. Generalize the Kubernetes-Galaxy integration to simplify the creation of custom Galaxy flavours, based on existing tool containers. A3. Maximize the scalability and performance of Galaxy when running on Kubernetes. A4. Introduce alternatives for data handling: better shared file-system, remote file-system support and object-store support. A5. Engage with collaborators and train bioinformatics communities (e.g. single cell RNA-Seq analysis, Galaxy workshops) to disseminate the use of the solution. More detailed milestones and their relation to aims can be seen in Table 1 of the additional materials. The main audience of the proposal are bioinformaticians, computational scientists or life scientists that:

need to easily recreate or share their workflow environment, to reproduce previous work; need to leverage cloud computing resources to scale their analysis to large quantities of data. The main activities to fulfill these aims and reach these audiences are: Software development, testing, continuous integration and documentation for the integration between Galaxy and Kubernetes and its deployment on cloud providers. Active collaboration between a group of software developers that

have as common interest to use this technologies integration on their own projects and in the community. Besides online interaction, this will materialize through a hackathon of 2 working days where most of these developers will be able to come together. Dissemination of the solution through proper documentation, examples and the delivery of talks/workshops.
Our proposal will positively influence the acquisition of open research tools through Galaxy across different omics fields. Within RNA-Seq, the Papatheodorou Group within EMBL-EBI is developing a Galaxy flavour for analysis tools for single cell RNA-Seq, based on the proposed setup. In Metabolomics, the PhenoMeNal H2020 Galaxy flavour currently relies in the ability of Galaxy to work inside Kubernetes to run on Google GCP, Amazon AWS, Microsoft Azure and various OpenStack installations across Europe. Our work will contribute towards the CloudMan Galaxy launching scheme (https://galaxyproject.org/cloudman/) to migrate to the fully containerised setup used on the Galaxy-Kubernetes integration. This will facilitate the deployment of CloudMan across more cloud providers and expand the set of tools that can be included in their installations, increasing the impact of our solution. We expect to monitor progress through the completion of a number software development, documentation and dissemination milestones, made explicit in the additional materials and in the outputs management section. These milestones are aligned with the aims proposed before. The major success indicator will be: Number of supported cloud providers Number of initiatives using this setup Number of unique sessions on documentation Number of deployed instances (provided the user allows the automatic anonymous recording of this).

Decision
Shortlisted, not funded

Comment on decision from Wellcome
This application was from a strong team, proposing to generate an innovative tool. It had good potential to impact health research across multiple fields, globally. However, the proposal would have benefited from more detail on how the impact of the work would be evaluated.

Title

Kickstarting Open Science Practices in Positron Emission Tomography: Open Data, Tools and an end-to-end Reproducible Analysis Pipeline from Image to Outcomes

Lead Applicant

Mr Granville Matheson

Details of proposal

Background PET neuroimaging yields in-vivo measures of protein concentrations and biochemical processes, and is used extensively for clinical diagnosis, pharmaceutical evaluation, and for brain research more broadly. Studies are often performed on limited samples due to high costs (single measurements can cost over $10,000) and due to exposure of participants to harmful radioactivity. Storage and analysis of PET data are highly idiosyncratic between, and even within, research groups, thereby limiting generalisability10. It is therefore of ethical and scientific importance that PET measurements are utilised to their full potential: that they are openly shared to allow re-use and enlarge sample sizes, and that they are processed in a manner which is transparent, generalisable and optimal. This is not currently a reality: to our knowledge, 4D PET measurements from three individuals have ever been openly shared without requiring application for permission for use, and there exist only a handful of PET papers which are fully reproducible (i.e. include all analysis code). The Brain Imaging Data Structure (BIDS) defines a standard for arranging and storing neuroimaging data11, and thereby allows for the development of generalisable processing software which can read and process this data. The standard was recently extended to include PET, and it was unanimously decided by a vote (at the NeuroReceptor Mapping Conference 2018) that a field-wide consensus paper would be drafted specifying, among other things, that the BIDS data structure should be used for the sharing of PET data. Aims We seek funding to be able to prepare and openly release a large set of PET test-retest data arranged in BIDS structure, as well as to create and openly release tools for processing of PET data arranged according to this standard such that analyses can be reported transparently in reproducible analysis notebooks. Our vision is thereby to make it substantially easier for researchers working with PET data to adopt open research practices by allowing them access to open data and code specifically for this purpose. We believe this will lead to improved research quality, quality assessment and accelerated progress in this field. Our target audience is therefore any researchers working with, or intending to work with, PET data. Aim 1. Preparation and release of open dynamic PET dataset. Our specific activities will include the following:
Developing tools for conversion of existing data into BIDS structure
Perform quality control of the resulting data Process this data using our tools to confirm we can obtain similar outcomes from the this data as we did from the data in its original format Release this data on OpenNeuro.org12
Write a data paper to describe this output, outlining our technical checks.
Upload the resulting paper to bioRxiv, and submit it to Nature Scientific Data, and release all code used to perform preparation and conversions on GitHub and Zenodo. Aim 2. Open-source dynamic PET analysis software development and pipeline integration. Our specific activities will include the following: Develop capabilities currently missing from the kinfitr R package4 such that it can perform blood preprocessing as well as voxel- and vertex-wise parametric quantification Write a Python wrapper for this package
Incorporate the kinfitr Python wrapper into a NiPype module13, such that it can be run within NiPype pipelines (a Python-based set of tools which wrap around numerous neuroimaging tools such that they can all be flexibly applied within a consistent framework)
Create a NiPype PET analysis pipeline which incorporates all of these tools, and which can process the shared data in an end-to-end fashion. Write an article describing these outputs, and demonstrating the advantages of reproducible reporting.
Upload the resulting paper to bioRxiv, and release all code used to perform preparation and conversions on GitHub and Zenodo. It should be noted that by operating on

BIDS data, this means that this pipeline should be applicable to almost any PET BIDS dataset, and the variety of tools included in the kinfitr R package (currently 14 different models) means that the correct kinetic model can be selected for the vast majority of common use cases. We believe that this software will be of greater utility than other non-commercial PET pipelines: MIAKAT14, PVElab15 and MAGIA16 require a MATLAB licence and are not BIDS-compatible. APPIAN17, when it is released, will be BIDS-compatible, but does not include pharmacokinetic models for tracers which require blood sampling (i.e. it is unable to process much of the data we will release). Success Indicators The major success indicators for both aims of this project are the release of the data and code, and the release of the preprints describing them. Secondary success indicators will be measured by the online repositories to which all materials will be uploaded: GitHub, OpenNeuro.org and Zenodo track download statistics, and with these numbers, we will be able to assess that the software and data are being accessed. Neither aim is dependent on the other, and any limitations of the BIDS format can be addressed and fixed in collaboration with co-applicant Melanie Ganz, who is the moderator for the BIDS PET extension.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal has the potential to create a valuable and impactful resource. However, there were concerns over the feasibility of the approach described and the relatively small number of datasets being used to populate the resource. The evaluation plan would have benefited from more detail.

Title

Building and Management of Open Access (Visualization) Platforms for Re-purposing Research

Lead Applicant

Mr Sridhar Hariharaputran

Details of proposal

Our vision is Building and Management of Open Access (Visualization) Platforms for Re-purposing Research - to serve the research community who would like to access, share the thoughts, outputs etc. By building the platform we expect better access and representation of data that will be useful re-purposing research ideas instead of creating their own. We will be building in phases to evaluate and discuss the proposal and build the platform.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to create an open access platform for visualising data. There were not enough details provided about the proposed activities, or how the team would monitor, evaluate and disseminate the resource.

Title

REPEAT: A platform for open and REProducible data Extraction AT scale for electronic health record studies

Lead Applicant

Mr Mohammad Al Sallakh

Details of proposal

Vision We aim to improve the transparency, reproducibility, validity, and efficiency of studies that use electronic health records (EHR). These studies have been enormously proliferating and increasingly used to inform evidence-based practice and health policy evaluation around the world. In the United Kingdom, major investments are being made, as in Health Data Research UK, to further catalyse this endeavour through capitalising on the country's considerable EHR data assets. However, there are ongoing concerns about the validity and reproducibility of EHR-based studies. Data extraction methods, including clinical codesets and algorithms to measure health outcomes from databases, are key to these studies but are often poorly reported and therefore their quality cannot be judged [1, 2]. Mechanisms that effectively facilitate the reporting, reproducibility, and reuse of data extraction methods are lacking. Researchers often develop their own methods on an ad-hoc basis, leading to unnecessary variations in the definitions of the same health outcomes [2]. Manually written database queries are common but are laborious and error-prone. These issues compromise the validity and comparability of studies, drawing significant clinical and health policy implications. To address these issues, we will develop REPEAT, a platform for open and REProducible data Extraction AT scale for EHR studies. REPEAT will bring data extraction methods into an open platform and automate and standardise their implementation. It will organically integrate with the study workflow, support collaboration between researchers, and minimise the barriers to share methods with the research community (see the attachment). Audience Researchers involved in undertaking EHR-based studies. The health research community. Approach We have developed a prototype data extraction tool within the Secure Anonymised Information Linkage (SAIL) Databank and successfully used it in several studies (see the attachment). It allows creating and sharing Read codesets and data extraction specifications from primary care data and it automatically generates database queries. We will further develop this prototype into REPEAT and make it available for the research community on a public website with the following features: Easily browsable clinical code dictionaries. A codeset library. A data extraction planner: Allows designing complex algorithms to extract data from predefined and user-defined data sources. Version control of codesets and algorithms. Ability to reuse, share, publish, and export methods into human-readable and executable formats. Automatic generation and execution of database queries. Automatic generation of descriptions of data extraction methods. REPEAT will support data extraction requirements of a wide range of EHR-based studies, from facilitating clinical codeset management [3] to simplified construction of complex algorithms to measure health variables. We will identify requirements by surveying a sample of UK-based researchers and data scientists involved in data extraction. We will invite those people to focus groups, semi-structured interviews and electronic questionnaires to enquire about their data extraction practices, challenges, requirements, and attitudes towards the proposed platform. We will seek ethical approval from Swansea University Medical School Research Ethics Sub-Committee and will handle the interviewees' data in accordance with the General Data Protection Regulations (GDPR). Project progression will be monitored against the milestones shown in the attachment. The main success indicator will be delivering a viable first version for wider use, as measured by successful testing on several databases and user acceptance. Impact REPEAT aligns with the vision of Wellcome and Health Data Research UK towards improving health research transparency, reproducibility, and efficiency [4]. It will facilitate the compliance with reporting guidelines, particularly the RECORD Statement [5]. REPEAT will: promote open research practices by bringing data extraction methods into to a public platform where they can be shared, cited in publications, and scrutinised by the research community;   facilitate research reproducibility by simplifying the sharing of methods that reflects exactly how study variables are measured from a database so they can be reproduced with confidence and minimal changes on other databases;   improve quality of studies, by replacing error-prone, manually written database queries with collaboratively reviewed methods and standardised, automatically generated queries, allowing researchers to focus on methods than query syntax;   save time and effort by facilitating rapid, automated and scalable data extraction, and by allowing researchers to reuse and repurpose methods; allow researchers with no programming skills to collaborate in specifying and reviewing data extraction methods; References 1. Hemkens LG et al. The reporting of studies using routinely collected health data was often insufficient. J Clin Epidemiol (2016) 79:104-111. 2. Al Sallakh MA et al. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. Eur Respir J (2017) 49.6:1700204. 3. Williams R et al. Clinical code set engineering for reusing EHR data for research: a review. J Biomed Inform (2017) 70:1-13. 4. Reproducibility and reliability of biomedical research: improving research practice. Symposium report by the Academy of Medical Sciences, BBSRC, MRC and Wellcome Trust (2015). 5. Benchimol, EI et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med (2015) 12(10):e1001885.

Decision

Not shortlisted

Comment on decision from Wellcome

The applicant opted not to share this information

Title **Blockchain-enabled application for cognitive assessment in dementia**
Lead Applicant **Dr Jon Brock**
Details of proposal (i) Vision At Frankl our mission is to make open science easier and more rewarding for scientists. We recognise that there are significant practical barriers and disincentives to the adoption of open science practices. Our approach is to develop tools that have "open" as the default setting but - crucially - are desirable for researchers to use even without these open features. Our initial focus is cognitive assessment - an area of expertise within the Frankl team. There is scope to substantially improve the user experience for researchers, clinicians, and educators, as well as the individuals completing the assessments. There are also clear market opportunities. This is important for the long term sustainability of the Frankl project. Aims The aim of this proposal is to develop an application for assessing memory abilities in patients with diagnosed Alzheimer's disease dementia or other forms of dementia, people with Mild Cognitive Impairment (who have higher risk of developing dementia), and cognitively normal elderly people with subjective memory concerns. The assessment will take the form of a simple paired associate learning task that measures the ability to form new memories. Participants are presented with pairs of words to memorise. After a short delay, they are given one word from each pair and required to select its partner. Variants of this test have proven sensitive to memory impairments in patients with early stage dementia. The test is simple and quick, meaning that it can be conducted at multiple intervals in longitudinal research to reliably track changes in memory over time and ultimately find utility as a convenient and accurate screening tool in primary care settings. The application will act as a vehicle for the features described above, including: integrated data management that pushes data to a secure repository; cryptocurrency micropayment that can be refunded when data are shared; blockchain record to establish the existence and provenance of data and increase findability. Target audience The application will be of immediate benefit to researchers and clinicians working with individuals with actual or suspected memory impairments. The data management will allow secure sharing of data within collaborations or between relevant clinicians, patients, and caregivers. The application will also be of interest to researchers across all scientific domains, serving as a working demonstration of blockchain-enabled research. Activities The project will incorporate four major activities: Development of a minimal viable product (MVP) Pilot phase in which the MVP is provided to clinicians and researchers for feedback Refinement of the application Wider release of the application (ii) Influence on open research practices The application developed within this proposal will facilitate and incentivise open research practices in the following ways: Making data sharing easier: The application builds data management into data collection by pushing the data to a secure repository as it is collected. Ensuring data are FAIR: Data sent from the app will be fully annotated. Consistent data protocols across apps will allow machine readability. Blockchain record of metadata will ensure findability. Incentivising data sharing: The Frankl cryptocurrency token provides a mechanism by which researchers and clinicians can be rewarded for sharing their data. Addressing the file drawer problem: Writing metadata to blockchain provides a permanent public record of the existence of all the data collected using this test. Creating a scientific supply chain: Hashed data on the blockchain provides a means of establishing the provenance of scientific data. Enhancing reproducibility: In the longer term, Frankl will provide a platform for other researchers to add their own tests to the platform and be rewarded with tokens when other researchers re-use their tests. This will enhance reproducibility of data collection methods. (iii) monitoring and evaluation Software development aspects of this project will be delivered in accordance with an Agile methodology. Progress will therefore be monitored in terms of the number of tasks delivered, the number of commits to a production branch of code, and conversely the number of

blocked or backlogged issues. These will be measured closely throughout the build, and especially at the commencement and finalisation of each 2 week ‘sprint’. Success indicators will include the existence of the MVP and final prototype, positive feedback from users including researchers, clinicians, and patients (or individuals completing the test). Indirect success indicators include engagement with the cognitive and clinical neuropsychology fields regarding open science practices, FAIR principles, and the development of standards for data and metadata. In the medium term (i.e., beyond the timeframe of this project), success will be demonstrated by the adoption of the application in research and clinical practice, and the re-use of the code we have developed. Our longer term vision is for Frankl to become a marketplace for open science software applications in which researchers and clinicians can share applications that they have developed and receive Frankl cryptocurrency tokens when they are re-used. The application proposed here represents the first step towards this goal and the development of such an ecosystem is the ultimate indicator of its success.

Decision
Funded

Comment on decision from Wellcome

This was an interesting application, focused on blockchain-based incentivisation, from a strong team. It has the potential to impact health research, through openness, across multiple fields.

Title

Open Data for Health Impacts of Air Pollution

Lead Applicant

Dr Marija Risteska

Details of proposal

(i) Project idea – vision and aim Air pollution is a major environmental health problem in Macedonia. Three biggest cities in the country (the capital Skopje, Bitola and Tetovo, with over 50% of the total population) were ranked among the top ten most-polluted in Europe in 2017. Harmful particle concentrations (PM10 and PM2.5) in Skopje (reaching as high as 1219 mug/m3 in the winter months at certain periods of the days) significantly exceed legally mandated thresholds (average daily concentration of 50 mug/m3). Rates of respiratory disease are soaring, amounting to 4.5% of total death causes. Since 2011,17 air quality monitoring stations have been put throughout Macedonia. Each station houses a maximum of six sensors that measure the following pollutants: PM 10 (particulate matter 10 micrometers or less in diameter), PM 2.5 (2.5 micrometers or less in diameter), CO (carbon monoxide), SO2 (sulfur dioxide), NO2 (nitrogen dioxide) and O3 (ozone). Using datasets from the Ministry for environment in 2015 an application was published called MyAir (MojVozduh) https://mojvozduh.eu/web which provided the public with a clear visualizations of pollution levels so that everyone could understand them. SkopjePuls https://skopjepulse.mk/ and AirQuality http://skopjefinki.ekoinformatika.mk/ are two other online apps that help citizens understand the air quality in real time. Macedonia has been ranked between the first 15 countries in Europe for having one of the most sophisticated e-health systems in Europe. The DRG system that facilitates pay-for-performance reform and My Appointment introduced in 2010 improved the scheduling of clinical appointments and reduced long waiting times. It is used by more than 5000 healthcare providers, integrating over 1000 applications and systems, including secure e-health records, pharmacy prescriptions, a performance-based pay module, automated provider credentialing, specialist referrals, ambulance service management, public booking interface for health interventions and medical equipment, etc. The cloud-based system is designed to be scalable by using modular programmes and solutions that can be integrated with other health care applications. It combines the HIS, and the services such as registering for organ transplantation, shared decision-making on health policy, text messaging and a live dashboard showing requests, referrals, most frequent diagnoses and prescriptions in real-time. With the expansion, it will integrate curative and preventive services, screening outcomes and risk factors, and will be used for health resource planning and management. Scientific understanding of air pollution and its health effects is incomplete in ways that are important for air pollution management and public health. This project aims to fill in this gap by establishing and measuring the correlation between air pollution and people's health in Macedonia. The main result will be the development of a software tool to allow for quantification of the health effects of exposure to air pollution, combining the data sets in the two systems for e-health and air pollution measurement. Specific activities include: Mapping of all data sets available within the MyAppointment and the DRG (Diagnosis-related group) system (Month 1) Mapping of all measurement sets for air pollution (Month 1) Developing hypothesis using the data sets (ex: consumption of (certain types of) medication during months with increase in air pollution; reported patients with acute respiratory problems throughout the year, etc.) (Month 2) Developing software and testing of computing, modeling and visualization of different datasets (Month 3-7) Developing evidence based health policy to decrease the health risks of air pollution and inform environment protection policy development (month 8-12) Sharing, presenting outputs and capacity building for stakeholders to use the resources (Month 9 to 12). 2) Influence on open research practices The project will create original data sets combining the data from the two sources (the air pollution monitoring stations and e-health system) presenting them in open data format. It will in addition generate visualizations for wider audiences, media and journalists to use for reporting, civic activism and further research. The fact the project will clearly present methodology, documentation, and open data will help the academic community in general to easily access resources. This will in turn allow for reproductibility of similar studies in other parts of the world, provide for follow-up work, build an epistemic community in this field of study and help its members to find new projects and collaborators, because it is easy for peers to understand what we are doing and how our research is embedded into the community. Open data on the correlations between health status, use of pharmaceuticals and delivery of health services; together with the air pollution will have an awareness raising effect and can instigate civic activism for better air quality. Furthermore, providing an open-access tool to enable research, debate, analysis and policy development will help develop policies and investments to support health reform, cleaner transport, energy-efficient homes, power generation, industry and better municipal waste management in order to reduce key sources of outdoor air pollution. 3) Monitoring and evaluation A detailed M&E plan will be elaborated in the beginning of the project, based on the following indicators: Number of users of the software New research being undertaken using the software; number of papers written, presented, published; number of research projects commenced (Social) Media mentions Civic actions undertaken Policy initiatives undertaken based on information from the software

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to build a data integration and sharing platform, which would be of value for assessing the effect of air pollution on health in Macedonia. However, the applicability of the system to other countries was not well described and so the potential impact of this proposal to transform health research through openness was unclear.

Title

The development of a novel interrogable open resource on work and health

Lead Applicant

Dr Melanie Carder

Details of proposal

Commonly reported work-related ill-health (WRIH) issues such as stress, back pain, dermatitis and asthma are estimated to cost the UK approximately 15 billion annually1. Central to the prevention of WRIH as well as the promotion of health at work and the maintenance of 'work ability' is the availability of good quality WRIH data regarding burden and risks. Such data are central to preventing WRIH as enshrined in regulatory guidance. The only medically verified, UK-wide source of WRIH data is that reported to The Health and Occupation Research (THOR) network, hosted by the University of Manchester and partially funded by the Health and Safety Executive (HSE) in the UK and the Health and Safety Authority (HSA) in Ireland. Established in 1989, THOR comprises a number of schemes to enable different groups of physicians to report incident cases of WRIH2. The database contains approximately 112,000 WRIH case reports (from both existing and legacy schemes) with an average of 150 new reports added monthly (currently from chest physicians, dermatologists, occupational physicians and general practitioners), and includes comprehensive information relating to diagnoses, occupations, industries, and exposures as well as sickness absence, reasons for referral and symptom onset. The importance of THOR as an existing WRIH resource is established, with the database generating numerous outputs which help inform HSE and others to determine their priorities and work programmes3. It also acts as a sentinel scheme to identify any new causes of WRIH. However, THOR as a resource is currently under-utilised. Access to (bespoke) WRIH information is currently only possible by request via the THOR researchers (Figure 1). This limits both the accessibility and the timeliness of the response (approximately 30 data requests were completed in 2017 with an average response time of one week). It also prohibits the type of hypothesis generating and testing possible when open access to data and research methods is enabled. Increased accessibility to THOR data (for example, during clinic when an unusual case is encountered) was highlighted as a priority by physicians participating in a recent workshop aimed at exploring health surveillance opportunities /innovations. Given that physician participation in THOR is voluntary, improving access to THOR data (and the benefits of participating for physicians), is key to the network's sustainability. The aim of this proposal is to develop the first directly accessible and interrogable UK and Irish resource on (medically verified) WRIH. The main beneficiary would be the UK and Irish working population through improved knowledge and awareness of WRIH issues and better targeting of resources to tackle identified problems. Specific objectives include: To improve the utility of THOR by developing an online, interrogable resource on the determinants of WRIH, To provide increased access to WRIH information including to anonymised data, To provide tools to analyse the data (e.g. trends over time or across sectors), identify sentinel cases and generate outputs and infographics that can be used for results dissemination. The target audience, key benefits and methods of evaluation/success indicators are shown in Table 1 (attached). The proposed resource would act as a portal to allow users to view, visualise and download data. The database would be made open by displaying and visualising the underlying data on the website which would connect to a number of tables located in a Microsoft SQL Server database and would be informed by similar systems within the National Drug and Evidence Centre (NDEC) https://www.ndtms.net/. The database would consist of different features with public and/or password protected access available. At the 'top' level would be controlled access to individual level data via a registration and endorsement system. Data would be anonymised (data reported to THOR are already pseudo-anonymised) to ensure the confidentiality of the participating physicians, employers and employees and to ensure compliance with the General Data Protection Regulation (GDPR). We will also seek advice as to whether an amendment to the current favourable ethical approval will be required. Other features would be partly informed by the needs of the target audience but would likely include searchable lists of reported asthmagens or skin allergens/irritants, reported diagnosis – occupation – agent triads, summary infographics (including user created), and resources relating to operational matters and research methods. A number of specific activities will be carried out. Representatives of the target audience will be surveyed to establish their key requirements from the proposed resource. Additionally, the 1000+ previous 'data requests' will be reviewed to establish the most frequently sought information (to help inform the type of summary data/infographics required). How to incorporate existing and proposed technological advances (for example, the educational tool EELAB), will also be considered. A pilot version (based on a subset of the data) would then be created and trialled amongst the target audience with the full version implemented following evaluation and feedback. Key to the resource would be its dynamism with new data, resources added at regular intervals. References 1 Health and Safety Executive. Available at: http://www.hse.gov.uk/sTATIstics/ 2 Carder M et al. The Health and Occupation Research Network (THOR) - an evolving surveillance system. SHAW 2017;8(3):231-236 3 Money A et al. The utility of information collected by occupational disease monitoring systems. Occup Med 2015;65(8):626-631

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to create a resource which had good potential to impact health research in occupational health research. However the level of innovation proposed was considered limited, and the evaluation plan would have benefited from more detail, for example identifying targets that would indicate success.

Title **One Medicine – Delivering Improvement in Osteoarthritis Treatment by Combining Knowledge From Human and Animal Medicine.**
Lead Applicant **Dr Amy Drahota**
Details of proposal (i) Our vision is to enable practices in the repurposing of research in veterinary and human medicine, through a pioneering consensus-derived tool. Our tool will use the paradigm of osteoarthritis treatment as an exemplar. There is evidence that consensus tools can change open research practice (e.g. they have enhanced research reporting/reproducibility; http://www.equator-network.org/). The practice of 'One Medicine'; reciprocal learning from animal and human medicine, aims to improve healthcare and the speed of treatment advances, and reduce and replace the need for animal experiments. Veterinary practice has much to offer human medicine and vice-versa (see supplementary information), however research is not being accessed across disciplines due to lack of knowledge, understanding, and standards. Specifically, we aim to produce a tool with the necessary credibility and utility to change perceptions, understanding, and practices towards the reciprocal application of research conducted in naturally diseased animals and humans. Osteoarthritis is a long-term age-related condition, considerably impacting quality of life, with impaired mobility and substantial pain in many cases. Treatment options span lifestyle changes, medication (topical, oral, injection), supportive treatments and surgery. Research in osteoarthritis in both humans and animals is aimed at primary and secondary prevention, novel and advanced treatment methods (e.g. regenerative medicine and manufacture of biomaterials). For example, osteoarthritis treatment studies in animals and now in human trials are investigating the use of stem cells, and this is being reciprocated in veterinary practice. Information between these studies is not exchanged. With the necessary thoroughness, veterinary practice can provide information that should be utilised and would be advantageous as veterinary patients are, like humans, clinically diverse and the disease state is not induced. This would replace the need for a significant number of animal experiments; potentially resulting in a reduction of animals used under the Animals (Scientific Procedures) Act. We will complete a standard Delphi study to underpin the tool, which will establish the circumstances in which research practices can be informed by each discipline, including the 'who' (species), 'what' (treatment options, site of condition), 'when' (stage of research pathway or intervention development) and 'how', with illustrative examples and promotion of relevant open research resources (e.g. The Osteoarthritis Initiative). We anticipate approaching approximately 25 panel members, with 3-4 rounds of electronic anonymous surveys interspersed by controlled feedback to reach consensus. A scoping review (informed by an interdisciplinary workshop of stakeholders independent of the Delphi panel) will be undertaken to create a preliminary framework of items for tool development. See roadmap for a detailed plan. We will utilise our networks to recruit international experts to this Delphi study, including academic scientists, bio-engineers, orthopaedic, veterinary, rheumatology, and patient/public informants. Consensus will be sought to demonstrate the similarities in treatment of osteoarthritis across species and how research from both disciplines can inform the other. Our target audience includes current and future generations of practitioners and scientists, and recipients of healthcare, in order to drive a step-change in practice. Our dissemination activities, as detailed in the 'proposal summary', are a key project component. (ii) The tool and associated publicity is expected to encourage and enable researchers to access and utilise research evidence from across disciplines; helping researchers understand the circumstances (who, what, when, how) under which evidence from their alternative discipline can inform their work. Our ambition is not only to inform osteoarthritis research and practice, but showcase how veterinary and human research can be combined so everyone can benefit sooner from advances in care in many conditions by reducing the critical pathway to translation and partly circumventing the so-called 'valley of death' associated with

technology readiness levels 6-7. This tool will also encourage researchers to consider new impact pathways for their research, leading to an ongoing promotion and open cross-disciplinary exchange of research. (iii) We (co-applicants), along with public representatives (people with osteoarthritis/owners of osteoarthritic pets), and Senior Research Associate, will form an interdisciplinary Steering Committee to monitor/advise on the project. The Committee will meet quarterly, but communications will continue between meetings as required. We will host two workshops (approximately 50 people each) at the beginning and end of the project. These workshops will be open to the public, researchers, clinicians, and veterinarians. The first workshop will help establish the problem-space in which the project is operating, and document perceptions/attitudes to cross-disciplinary research exchange. The second will seek feedback on the tool, assess changes in perceptions/attitudes, and consider next steps. Members of the General Medical Council, Royal College of Veterinary Surgeons, and Arthritis Research UK will be invited to this last workshop. Our evaluation of this project's impact will, by necessity, extend beyond the funding period. We will evaluate the Delphi study's impact by tracking panelists' perception change with pre- and post- questionnaires. Success indicators include positive changes in perceptions, understanding, and increased openness to cross-disciplinary exchange of research. We will monitor access statistics and citations of our results (identifying those implementing our outputs), reach to delegate numbers at conferences and webinar (to include pre-and-post feedback questionnaire), and viewer statistics for the television programme. We will analyse social media activity in response to the television show and news reports, to gauge wider reach and public response.

Decision

Not shortlisted

Comment on decision from Wellcome

The vision behind this proposal was strong and the resource could be potentially impactful. However, there was a lack of detail on the proposed activities and the methodology was not clearly described.

Title

FAIR Enough? An Open Data Commons for Spinal Cord Injuries

Lead Applicant

Dr Fiona Murphy

Details of proposal

We intend to study the evolution of the Open Data Commons for Spinal Cord Injury Research (ODC-SCI), a data portal for sharing and publishing pre-clinical spinal cord injury data. In response to the difficulties encountered in formal attempts to reproduce the findings of major spinal cord injury research findings and the lack of meaningful translation (Steward et al., 2012), a group of leaders in SCI have come together, led by Dr. Adam Ferguson from the University of California, San Francisco, and Dr. Karim Fouad of the University of British Columbia, to establish the standards, infrastructure and governance for sharing of primary research data. This work is predicated on demonstrations by Dr. Ferguson and his colleagues that aggregating individual data sets shared by individual researchers (i.e., long tail data), produces a complex, multidimensional picture of spinal cord injury they call the Syndromic space. Analysis of the syndromic space led to better predictive modeling of the evolution of SCI, and more robust cross-species biomarkers (Ferguson et al., 2014). The concept of the syndromic space and the SCI example are important drivers for establishing publishing paradigms in pre-clinical research, because they suggest that reproducibility and robust translation lie not in individual experiments, but in the piecing together of these data sets in order sample the full variability of the syndrome. If true, then for individual investigators, there is demonstrated and critical value for the field in publishing FAIR, open data, and the community needs to establish the standards and best practices for doing so, and the citation system for ensuring that individual contributions are recognized and credited in subsequent analyses. It is important, therefore, to understand how the governance and technical infrastructure can aid or impede the production of FAIR data and their open sharing. With some seed funding from the Craig H. Nielson Foundation, Dr. Ferguson in collaboration with the Neuroscience Information Framework (led by Dr. Martone) were able to launch a beta version of the ODC-SCI (https://scicrunch.org/odc-sci). The design of the portal was heavily influenced by the work on the scholarly commons by FORCE11, and also reflects concerns expressed by the researchers regarding their professional and personal vulnerability, e.g., from animal rights activists. Thus, the portal was designed around 3 spaces (Fig 1): a private space where individual laboratories can upload their data and only make limited metadata available to others; a semi-private space where vetted members of ODC-SCI share data with each other, and a public space where data are made available via a CC-BY license. The researchers have agreed to make their data FAIR, and they are coming together to define what FAIR means for their community (e.g., what are a plurality of relevant attributes? What are community standards for SCI? The focus to date has been on defining common data elements (CDE's). We are already seeing the challenges of having to manage conflicting and evolving requirements for CDE's, and managing the process of establishing what is FAIR enough at any point in time. We are particularly interested in how much the infrastructure helps or hinders the process of data sharing. We often hear it said that the problem with data sharing is not technological, but sociological. However, our experience is that technology can very much get in the way of the willing, particularly when trying to implement FAIR fully while asking researchers to change the way they do their research. The ODC-SCI, as with many research infrastructures, was designed after an iterative process of requirements gathering and design. It is our experience that researchers will give many requirements for what an infrastructure needs to do, but when the full complexity of a workflow is implemented, and enough resources aren't available for things like user interface design, the infrastructure ends up too complex to use. We are finding that this is the case with the current ODC-SCI. The goal of this research will be to examine the current barriers to use and help to implement policies and procedures that balance the needs of both data producers and data consumers for FAIR SCI data. As part of this work, we will produce a set of machine-readable decision trees that document the policies and procedures currently in use within the community (Sweeny et al., 2015) The SCI community is a unique community in which to test and evaluate open research practices because unlike in many data sharing efforts, the value of primary data sharing has been established by Dr. Ferguson and colleagues and the leaders of the field have recognized the potential value and agreed to share. Because the data sharing portal is in early stages, we can evaluate what effect different policies and requirements have on the ability and willingness of researchers to make their data available and what are the minimum requirements for doing so to enable reuse. Success will be measured directly by the number of data sets that are uploaded to the private portal by the individual laboratories, how many employ the agreed upon standards, and if and how quickly they migrate through the different spaces to the public portal. References provided in additional information

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was from a strong team proposing to study an interesting resource. However, the methodology was not well described, and the potential impact of this proposal to transform health research was not clear.

Title **BRAINS (Benefit-Risk Assessment INteractive Software): an open source interactive tool for drug benefit-risk assessment**
Lead Applicant **Ms Gaelle Saint-Hilary**
Details of proposal (i) Vision, aims, target audience, activities Context: A drug benefit-risk assessment consists of balancing its favourable therapeutic effects versus adverse reactions it may induce. The benefit-risk balance is a strong predictor of the therapy's long-term viability and a key element for decision-making during the drug's development. Quantitative methodologies have been recently proposed to make a benefit-risk assessment more comprehensive and consistent. Rationale: The quantitative evaluation of a drug's benefit-risk balance is recognized to be a complex process [1, see attachment]. Difficulties arise due to the lack of software implementing these evaluation tools. Although there are at least two freely available R packages (MCDA and SMAA) that can potentially accommodate the benefit-risk assessment, they require an experience in the decision theory and in programming. These packages were not designed for medical research. On the other hand, the most popular software for the benefit-risk assessment (using MCDA only), Hiview, requires a licence to be used. Therefore, there is a crucial need for simple, open, interactive tools to support the discussions within interdisciplinary teams, and to communicate the results. Aim: The aim of the project is to develop an accessible tool supporting benefit-risk assessment which can be used by users with various backgrounds. The distinguishing feature of the software is multiple levels of customised interface. The tool will accommodate: various types of benefit and risk outcomes (e.g. binary, categorical, continuous) various tools to elicit the user's benefit-risk trade-off (e.g. swing-weighting, discrete choice experiment method) various types of aggregation methods (e.g. linear and multi-linear utility score) planning of a benefit-risk analysis (comprehensive simulation studies) implementation of a benefit-risk analysis using real data Target audience: pharmaceutical industry, health authorities and academic teams involved in drug benefit-risk assessment and healthcare decision-making. Activities (12 months): 1. Identify an advising committee: 5 experts working in benefit-risk analysis (before start) 2. Conduct interviews with members of the advising committee and online survey to collect experts' needs and preferences (month 1) 3. Develop Beta versions of the R Shiny App and R-package (months 2-8) 4. Users Acceptance Tests of the R Shiny App by the advising committee, collect feedbacks (month 9) 5. Finalising the app and its launch (months 10-11) 6. Dissemination meeting which will include an extensive workshop on the App application, the demonstration of its tools, and showcase of App application using real clinical data (month 12) During the course of the project, all members will have a bi-weekly video-conference discussing the progress of the project. On-going communication: In order to inform the scientific community, we plan to present the project and its progress at a webinar (month 9) and 3 conferences: PSI 2019 (targeting pharmaceutical companies and health authorities) ISCB 2019 (targeting academic institutions) ISPOR 2019 (targeting health economists and outcomes researchers) The interactive tool and R package will also be described in two publications. (ii) How do we influence open research practices? Open research practice is recognized to improve the quality and the transparency of healthcare information systems. The superiority of open source licensing promotes safer, more effective health care information systems [2]. The R Shiny App for Benefit-Risk Assessment will be: Free of use Therefore, all stakeholders could use the same tool to perform the analyses, with a consistent presentation of the results. These results could be used as a basis to facilitate the discussions between stakeholders, e.g. between the pharmaceutical industry and health authorities, and to ensure a transparent communication of the decision-making process. Moreover, all the analyses could be reproduced independently by the different parts, along with sensitivity analyses. Open source This will permit the users to influence the quality and potential upgrades of the tool. It will also permit to tailor it according to their individual needs. All

these aspects will be widely emphasized during our communication process, with specific examples and success stories (when available), in order to encourage the scientific community to adopt open research. Generating standardised reports of the analysis This will start the systematic work on the report templates providing complete information about the benefit-risk analysis and allowing stakeholders to make a conscious decision based on this analysis. (iii) Monitoring On-going success indicators: The survey of experts is completed on time and the programming is started on month 2. The "working" version of the App is available on month 6. The Beta version of the App is completed on month 8. The positive evaluation of the Beta version. R package is available on CRAN. Final success indicators: The positive evaluation of the App by experts at the final dissemination meeting (through an assessment survey). 40 users in 12 months post release. Two papers submitted 2 months after the software release. Risk assessment. The main challenge that might arise during the project is that the software takes longer to be developed due to many customised functions to be included. To make sure that the project is not affected by a delay, the team will use "onion coding", i.e. implementing some of the functions for one statistical method, but structure it in a way that methods can be easily added after the release by the team or by the users.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to create an open source tool and R package, which could contribute to transparency in benefit risk assessment for new medicines. However, the level of demand and buy-in from the target audience was not well described and so the potential impact of this proposal to transform health research through openness was unclear.

Title

ORIGIN: an international network to establish open research standards in genome-wide association study

Lead Applicant

Dr Jie Zheng

Details of proposal

Since 2007 more than 3471 GWAS have been published (https://www.ebi.ac.uk/gwas/). The results of these studies are valuable for genetic epidemiology research and show massive potential in supporting drug development and developing health policy. With complete GWAS summary statistics, we will be able to line up associations to different variants in the same region and have the ability to assess overlaps between traits where often the variant has not yet reached significance. However, to date, most GWAS publications only shared the top genetic association signals, which limits the translational potential of downstream research. In some areas of health science, reporting standards have been established. For example, CONSORT standardized the reporting of randomized controlled trial results, which greatly improves the reusability of these results (www.consort-statement.org). However, in published GWAS, researchers still report results in heterogeneous format. Over the last 3 years the MRC Integrative Epidemiology Unit at the University of Bristol (MRC IEU) has been developing a large database of full GWAS result datasets. MRC IEU further linked these to open analytical platforms: MR-Base (www.mrbase.org) and, in collaboration with Broad Institute of MIT and Harvard, LD Hub (http://ldsc.broadinstitute.org/ldhub/). Our database contains standardised full GWAS summary results for more than 4000 complex human diseases and traits, which greatly extends the accessibility and reusability of these genetic association data. In LD Hub, we also experimentally developed a GWAS sharing platform to enable researchers to make their GWAS results findable, accessible, interoperable and reusable. MR-Base and LD Hub highlight the value of this "open science" approach to integrating high quality data and cutting-edge analytical tools, achieving 3000 visits per month by users from more than 100 countries (data from Google Analytics) and citations by more than 190 research publications within the last 2 years (data from Google Scholar). In parallel, the European Bioinformatics Institute has continued to develop their GWAS Catalog of "top hits" and meta-data from published GWAS originated by the National Human Genome Research Institute (NHGRI) (https://www.ebi.ac.uk/gwas/). The GWAS Catalog now hosts 255 summary statistics datasets in heterogeneous formats, as published by authors and is developing a database to store these data in a standard manner. They will also implement a platform for prospective collection of GWAS results submitted by researchers. There is, therefore, enormous potential to establish data and meta-data standards to improve interoperability between these (and other) GWAS databases, improving data accessibility, reusability and value to the research community. Until now the proportion of GWAS studies sharing their full summary results has been limited to approximately 5~10%. For the results of GWAS studies to fully realise their potential an open science model needs to be adopted that fulfils the following principles: 1) GWAS summary results should be openly accessible.; 2) GWAS results should be published in a standardised format to maximise reusability of these data; 3) A single open database platform should be developed to share GWAS results, reducing duplication of effort through different organisations building their own database. With support from the Wellcome Open Research Fund, we plan to set up a committee – Open Research In Genome-wide association study International Network (ORIGIN) – to achieve the following Aims: Establish a reporting standard for GWAS summary statistics, which includes: Establishing the standard for study level meta-data of GWAS, which should include key information such as PubMed ID, publication year, trait information, ontology mapping and so on. Discussing the standard for SNP information, which includes genome build, strand, alleles and so on. Defining a standard format for GWAS summary results, which includes SNP identifier, effect sizes, p values and so on. Establish a work plan for integration of the GWAS databases between UoB, EMBL-EBI and Broad. Developing a GWAS data harmonisation approach Creating database schemas for the GWAS data and study level meta-data. Engage the community – engage more people to share GWAS data with the community by: Communicating with funders to discuss data sharing standards for future grant application. Contacting journal editors to discuss data sharing policy for future GWAS publications Principal investigators of major cohort studies and GWAS consortia. Encouraging researchers to make the GWAS results openly accessible. Engaging with the pharmaceutical and genomics industries to share GWAS results. To achieve these aims we plan the following Activities: Establishing the ORIGIN core committee, comprising GWAS database leads from MRC IEU, EMBL-EBI, and Broad Institute. Organising a two-days' workshop to bring key stakeholders (GWAS researchers, funders, journal editors) together to discuss and establish a draft standard for GWAS summary statistics. Developing a final GWAS summary statistics standard document (the ORIGIN statement 2019) agreed by the ORIGIN Committee. Establish a workplan for implementation of the standards within the EMBL-EBI GWAS database and integration of the UoB, EMBL-EBI and Broad GWAS databases. Publish outcomes in top genetics journal and present at American Society of Human Genetics (ASHG) Conference. Timetable

0-3 months

3-6 months

6-9 months

9-12 months

Activity 1

Setup the ORIGIN core committee

<p style="

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal sought to develop a standardised format for reporting GWAS resutls, which would increase the interoperability and reusability of GWAS data. However, there was no evaluation plan provided, for example to identify targets that would indicate success of the work.

Title

Promoting adoption of transparency policies by journals

Lead Applicant

Dr Brian Nosek

Details of proposal

Researchers endorse the principle of transparency of research (Anderson et al., 2007) but the present culture does not provide good incentives for openness of research content or process. As a consequence, rates of open research practices are low (Alsheikh-Ali et al., 2011; Vines et al., 2013; Stodden et al., 2018). The TOP Guidelines (TOP, https://cos.io/top) were created by a group of journal editors, publishers, and research funders (Nosek et al., 2015) with a goal to provide consistent policy specifications for publishers, journals, and funding agencies. TOP has eight policies for promoting more transparent scientific practice, including: Data citation
Data transparency Materials transparency
Code transparency Design and analysis transparency
Preregistration of studies Preregistration of analysis plans
Replications Each of the eight policies can be implemented in one of three levels of increasing rigor: 1. Disclosure of whether or not an action occurred or data are available 2. Requirement for transparency when ethically and logistically possible 3. Verified third party verification of transparency. For each of the eight policies, a journal receives a score ranging from 0 (no level of compliance) to 3 (verified third party verification), resulting in an overall journal score of 0 (no TOP policy compliance) to 24 (level 3 compliance on all eight TOP policies) More than 850 journals have implemented TOP standards. However, adoption has stemmed largely from idealism--publishers or editors that believe open research is important make proactive steps to improve openness in their journal(s). Editors do not have strong incentives to adopt TOP generally or increase their rigor and have few opportunities to compare their editorial policies with those of their peers. However, if an editor observes that other journals have adopted openness policies, they may feel some pressure to adopt similar policies. This behavior change has cascading influence. Each additional editor that changes policy increases the normative pressure for recalcitrant editors to follow suit. Changing the entire research culture will require scaling beyond these early adopters. In March, 2018 COS piloted a TOP adoption strategy among 33 social-personality psychology journals. We measured the editorial policies of these 33 journals against TOP Guidelines and created a Google spreadsheet (http://bit.ly/TOPjournals) which was sent via email to the journal editors March 1st. As of March 1st, 20 of the 33 journals (60%) had TOP scores of 0-3 (none to minimal transparency standards) and 6 (18%) had scores of 11-20 (strong to very strong standards). When exposed to information that included that of their peers, ten journals (30%) reported updating their policies following this single email intervention. As of August 1, 13 journals (39%) had TOP scores of 0-3, and 13 (39%) had scores of 11-20 (see Figure 1). Building from this pilot project, COS proposes to define and test a reproducible workflow, including identifying clusters of discipline specific journals, scoring these journals, and communicating with editors to increase TOP implementation. Additionally, we will seek to increase the inter-rater reliability of our scoring procedures to provide a highly repeatable set of scoring guidelines. COS will engage six clusters of journals (two in phase 1; four in phase 2), with the first round informing the second round. See Figure 2 in Project Methodology for a visualisation of this strategy. The initial cluster of journals represents subdisciplines with at least ? of the journals being TOP adopters, including generalist journals (e.g., Science) and relevant subdiscipline journals. In each round, project staff will 1) determine the journal groups, identifying 15-30 journals within a discipline; 2) score the policies using the prototyped visualization tool. The visualization tool will be shared with the editors of these journals to make them aware of the emerging norms and motivate them to adopt open research policies; 3) share the scores with the journal editors, including how each journal compares to the aggregate score; 4) help journal editors to assess existing policies, and draft and implement more open,

transparent policies; 5) track progress over time; and 6) use successful interventions to seed the next cluster of journals. The visualization tool and workflow will be updated as journals update their editorial policies and as the proposed project progresses. Success indicators for this proposal will include: Successful creation and execution of a efficient and maintainable workflow and prototype visualization tool Dataset of TOP scored journals that will be openly available for use by all Accelerated TOP adoption among targeted journals. Not only will this project help to increase the number of journals that are implementing TOP policies, but it will create a reusable workflow and a visualization tool that will be valuable far beyond the scope of this project. TOP Guidelines visualization prototype will benefit journal editors, research funders and scholars, who could use this tool to determine which journals share their values. With future funding, COS would transition the prototype into an open and reusable tool for community use. Additionally, the open dataset that will result from the scoring of journals' policies will be a valuable resource for the entire research community. Researchers may use both the data and workflow materials (all available open source) for future scholarly pursuits.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an application from a strong team, and the commitment to advancing openness was clear throughout this proposal. However, the level of innovation, as well as the potential impact of this proposal to transform health research through openness was limited.

Title

An Open, Patient-centric Information Commons for Tuberculosis

Lead Applicant

Dr Keith Elliston

Details of proposal

(i) Vision: The development of an Open Patient-Centric Information Commons to enable Tuberculosis drug discovery and development through clinical and translational research. Aims: (1)To pilot the development of global Patient-Centric Information commons for TB (2) to load up to 12 distinct, high dimensional data sets to enable TB clinical and translational research (3) To encourage data sharing in the global TB community by providing a shareable resource. (4) to encourage the Federation of TB Patient data using the PIC-SURE API and i2b2/tranSMART. Background The i2b2 platform has been developed over the past 12 years, substantially funded by NIH, and is the dominant platform in use for clinical research. The i2b2-tranSMART Foundation was formed specifically to enable the integration of tranSMART and i2b2, along with PIC-SURE, Hail , Fractalis, etc., to form i2b2/tranSMART. This is the first deployment of this platform for a single disease (TB). Target Audiences: Our target audiences are: (1) Scientists and clinicians at OSPF doing TB drug discovery and development. (2) other TB related non-profit entities (Such as the TB Alliance). (3) TB Patient registries, (4) Students and scientists at Academic Medical Centers and Research Hospitals, (5) Pharmaceutical and Biotech companies doing research and development in TB. Activities: During the course of this grant, we will: (1) Analyze and Curate up to 12 High Dimensional datasets from various Tuberculosis research studies stored in GEO. The selected datasets include but are not limited to the following:

GSE107995: A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection 414 samples 262 patients some at multiple timepoints

GSE19491: Blood Transcriptional Profiles in Human Active and Latent Tuberculosis 498 samples 498 patients

GSE83456: The transcriptional signature of active tuberculosis reflects symptom status in extra-pulmonary and pulmonary tuberculosis 202 samples 202 patients

GSE42834: Human whole blood microarray study to compare patients with tuberculosis, sarcoidosis, pneumonia, and lung cancer 356 samples 356 patients

GSE70478: Epigenetics and Proteomics Join Transcriptomics in the Quest for Tuberculosis Biomarkers 102 samples 38 patients 3 datatypes 6 only methylation

GSE62147: Gene expression in M. tuberculosis and M. africanum infected tuberculosis patients prior to and following treatment 52 samples 12+14 patients 2 timepoints

GSE103147: Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease 1650 samples 119 patients 5 timepoints 2 celltypes 4 sampletreatments

GSE89403: A blood RNA signature for predicting the treatment outcome in the Tuberculosis Treatment Response Cohort 914 rows 453 samples 142 patients 4 timepoints 2 duplicates

GSE94438: A blood RNA signature for tuberculosis disease risk in household contact study - GC6 cohort. 434 rows 418 samples 335 patients

(2) Load these curated data into i2b2/tranSMART so that they can be analyzed alone or in concert with other studies (3) host regular conference calls to manage the project and all shared documents, data and software (4) enable the Federation of TB data through the PIC-SURE API, enabling the query and analysis of

data across implementations. (5) provide hosting and support, to ensure its widest possible adoption in the TB community. Hosting and support will be provided by Axiomedix, Inc, a commercial provider of hosting, support and training for i2b2/tranSMART. (6) ensure the sustainability of this platform by supporting and maintaining this implementation using OSPF resources. (ii) Influence on the Field: The development of an i2b2/tranSMART Instance for Tuberculosis research and development will not only provide an enabling resource for TB R&D, but will also provide an organizing resource for the integration of TB patient registries and TB clinical studies into a single, Federated data commons. In particular, other research efforts can take the same approach that we are, and integrate their data into an instance of i2b2/tranSMART. We can then collaboratively develop a Federated network by connecting these instances through the PIC-SURE API. Thus, we can provide a means for querying distributed TB patient data across the internet, linking distributed TB patient data on a global basis. This project will be the first of its kind in the TB field, and will provide the basis for further open and collaborative work in this field. (iii) Monitoring and Evaluation This project is very straightforward to monitor and evaluate, both during development, deployment and adoption. In the development phase, success will be measured by the number of data sets and patients represented in the platform. We are targeting a minimum of 6 datasets encompassing at least 1,000 patients. If time and resources allow, we will extend this pilot up to 12 datasets and up to 2000 patients - based upon the available data from GEO. During the deployment phase, success will be measured by the ability to effectively access and utilize the loaded patient data over distributed connections within India, the United States and Europe. Success will be measured by achieving suitable performance on selected analytical tools to ensure usability. Finally, during the Adoption Phase, success will be measured by acquring new users for the platform - facilitated through directed training (interactive and video training) for the platform, and adoption by end users. Adoption by 50 scientists in the first year will be considered success, growing this to over 200 scientists in 2020.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to collate and curate datasets into a central resource, which would be of value to the TB research community. However, the wider impact of this proposal on open practices was limited.

Title

Evaluation of a new reporting standard for interventions in surgical systematic reviews

Lead Applicant

Dr Matt Vassar

Details of proposal

(i) Here, we propose to test a new reporting standard designed to improve the reporting of interventions in systematic reviews. More transparently and completely reported interventions would result in greater research reproducibility and improved understanding of an intervention for physicians to base clinical decisions. We briefly describe our proposal below. Complete reporting of healthcare interventions is essential for interpretation of research findings and reproducibility of results. A completely reported intervention must include a number of key features such as duration, dose or intensity, mode of delivery, essential processes, and monitoring. However, authors may incompletely report these features, compromising the effect of the published literature and potentially contributing to research waste. To address the poor reporting quality of interventions, an expert committee convened to create the Template for Intervention Description and Replication (TIDieR) checklist and guide. TIDieR was designed to be broadly applicable to all healthcare interventions. After publication of TIDieR, an adaptation was created for systematic reviews. Since systematic reviews may be considered among the highest levels of evidence and are able to resolve discrepancies between primary studies, the consequences of incompletely reported systematic review interventions are more severe. For that reason, our aim is to apply the TIDieR checklist to systematic reviews of surgical interventions published in high-impact medical and surgical journals. We will also compare the completeness of intervention reporting of journal-published systematic reviews with Cochrane systematic reviews. (ii) Reporting standards should maximize transparency about the research process and minimize potential for vague or incomplete reporting a study's methodology. Therefore, we expect our proposal to influence open research practices by highlighting the current deficiencies in intervention reporting in surgical systematic reviews; more completely reported interventions will improve the usability of systematic reviews to patient care. Furthermore, successful completion of this project will result in a new transnational network (U.S., Australia, U.K.) to deliver impactful secondary research in surgery. (iii) We will use a series of safeguards to ensure that our study is completed in a successful manner. First, we will use a systematic review platform called Rayyan that closely tracks our study's progress. We will establish a timeline with key milestones for completion of our database searches to retrieve systematic reviews, title and abstract screening, data extraction, resolving disagreements between extractors, data analysis, and manuscript writing. We use a project management platform called Freedcamp to track performance for these key milestones. We will also hold mandatory weekly meetings to ensure that we are meeting or exceeding our weekly goals. Our success indicators will be the prompt completion of these milestones.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to test a new reporting standard for systematic reviews. The methodology was not clearly described, and the potential impact of this proposal to transform health research through openness was limited. The proposal would have benefited from details on how the impact of the proposed activities would be evaluated.

Title
Trustless Third Parties

Lead Applicant
Dr James Cunningham

Details of proposal

Vision Over 8 months this project seeks to implement a system of 'trustless third parties' based on distributed ledger technology. We define a trustless third party as an open, distributed algorithmic mechanism that reproduces the functionality and behaviour of the trusted third parties that form a crucial part of the data-oriented medical research pipeline. Further the project will explore means by which token-based solutions can be employed to incentivise more active participation and engagement in the research process by the ultimate providers of the data, the patients themselves. A trusted third party is an organisation that provides a means of linking data between data sources in a secure and legally compliant manner. We see the potential to replace these entities with a system based on the cryptographic primitives inherent in blockchain technology, giving us a more open and secure means of data linkage. The nascent field of crypto-economics is beginning to explore ways in which token mechanisms can incentivise behaviour in bespoke environments; the trustless third party system will give us a platform to apply this thinking in the area of medical research. Aims, target audience and activities The primary aim of this project will be the production of a series of Ethereum-based smart contracts that explicitly replicate the properties of Trust Third Party entities in their role in the medical research pipeline. Particularly the emphasis will be on functionalities concerning the pseudonymisation of patient identifiers in a cryptographically secure manner; the linkage between identifiers originating from disparate data sources; and the coordination of gathering, transmitting and updating of patient consent for the use of data in a manner that conforms to current ethico-legal guidelines. This implementation will serve as a reference implementation for a potential standard for on-chain data linkage mechanisms. Further, the project will produce an initial implementation and outline of a system for using token-based mechanisms for incentivising the sharing and use of anonymised data-extracts for research use. This portion of the project will include an analysis of potential ways in which blockchain based solutions for encapsulating incentive as an exchangeable token can be applied in the area of data sharing and reuse. The target audience for this research falls into three categories. Firstly data custodians, particularly within the NHS, who are the primary first-stage users of trusted third party solutions and who would benefit from the increase in security, the decrease in the need for explicitly trusting external organisations, and the (internal) economic benefits of increased use of their research data and the increased participation of patients in the research process. Secondly, the community of research practitioners who will benefit from any system that streamlines the process of acquiring and linking data, and additionally from the analysis of means by which data providers (both individuals and data custodians) can be incentivised to increase use of their data. Finally, the use and application of blockchain technology applies directly to researchers in computer science, where the practical application of this rapidly developing area will feed directly back into theoretical research. The activities this project will undertake are summarised as follows: Analysis of the functional properties of existing TTP solutions Development of smart-contract implementations matched against functional properties Development of testing and security analysis framework Crypto-economic analysis of incentivisation mechanisms for research participation Token-based incentivisation framework implementation Standards development and results dissemination Influence on open research practices In itself the use of blockchain technology at the core of this project represents an attempt to move the pipeline of medical data-oriented research on to a more open footing. Distributed ledger technology is inherently open, operating on consensus mechanisms and resulting in immutable, auditable, trails of data use. This translates directly into an application of FAIR principles. By allowing the trusted third party mechanisms to operate in the open on a blockchain, meta-information about how data is used, what it is used for and why it was requested can be put in the open thus encouraging the use and analysis of this information by the wider research community. This will further give an underlying mechanism for enabling the reproducibility of research results by making publishable the portions of the research pipeline that pertain to the initial sourcing and linkage of data. An explicit output of this project is an analysis of incentivisation mechanisms for consent and use of data, further fitting in with open research principles. It will be made explicit in the plan for disseminating outputs for this project that the implementations of the smart contracts and token mechanisms that are developed will be released and disseminated under an open-source license. Evaluation / Success The initial output of this project, a functional analysis of existing properties of trusted third parties, will serve as the basis for evaluating and gauging the success of the smart contract implementations. The primary evaluation of the project will use this analysis to produce a framework for evaluating the implementation. Success will be measured as the degree to which the implementation conforms to this framework. For the cryptoeconomic analysis, evaluation will be in the form of a qualitative evaluation of the potential impact of proposed solutions.

Decision

Not shortlisted

Comment on decision from Wellcome

The proposal was an innovative and novel application of blockchain technology. However, the approach for securing wider uptake and community buy-in was unclear.

Title **AI-curated Knowledgebase and Knowledge Graph for Nipah Virus and Tuberculosis.**
Lead Applicant **Dr Keith Elliston**

Details of proposal

(i) Vision Our vision for this project is to build a living knowledgebase focused on the specific content needed to enable drug discovery and development for TB and Nipah Virus. We will leverage advances in artificial intelligence (AI) to constantly scan the literature, databases, blogs, podcasts, videos, books and more, to find those sources relevant to our audience of Scientists, Clinicians and Patients. This will help us to engage our Open Source Pharma community, and further our mission of the development of new therapies and diagnostics for the neglected diseases of the third world. (a) Aims 1. We will build a Tuberculosis Pathobiology knowledgebase using the Charisma AI curation platform. This system will be trained initially using and existing corpus of scientific articles curated by the OSPF. The platform will be configured for up to 50 sources of content (including Pubmed, PubChem, Patents, Clinical trials, blogs, podcasts, videos, abstracts, etc.). Subject matter experts will help to train the system to recognize content of interest over the course of a 6 week training and normalization period. The resulting platform will be an automatic intelligent curation system that scans its sources for relevant content on a daily basis.   2. We will repeat the KB process for Drug Resistant TB and Adjunct therapies, and for Nipah Virus Pathobiology.   3. We will use the Ingentium NLP and Semantic predication platform to read the new content added to each knowledgebase, and this knowledge content will be used to populate a disease-focused knowledge graph (one for TB, one for Nipah Virus). This content will be incorporated into the Ingentium Standardized Knowledge Graph framework, that unifies the content into a single semantic knowledge space. These knowledge graphs will be embodied initially in Neo4j - but can be ported to other graph databases.   4. We will develop a set of knowledge applications, that will make this content avialable and useful for our target audiences. These will include: RSS feeds, News magazines, twitter and facebook feeds, and feeds to flipboard and apple news. These applications will bring users back to the OSPF knowledge environment, where they can engage with the OSPF Community.   (b) Target Audiences 1. We will tailor our content to direct content specifically to (1) Scientists (scientific references, patents, books, datasets, etc.), (2) Clinicians (clinical trials, new drugs, patents, blogs, etc.) and (3) Patients (videos, blogs, podcasts, bookmarks (websites) etc.).   2. We will also target three specific audiences by disease area: (1) Tuberculosis pathobiology, (2) TB drug resistance and adjunct therapy, and (3) Nipah Virus pathobiology.   (c) Activities 1. We will initially create the knowledgebase and curation engine for Tuberculosis Pathobiology. This will be trained over the course of 6 weeks until it reaches equilibrium.   2. We will create the knowledgebase and curation engine for Nipah VIrus Pathobiology. This will also be trained to equilibrium. 3. We will create the KB an curation engine for TB drug resistance and adjunct therapy, which will be trained to equilibrium. 4. We will modify the standardized Ingentium Knowledge graph (containing 5M nodes and 20M edges) for TB and Nipah Virus respectively. This will be accomplished through the addition of domain specific ontologies and databases.   5. We will tune the NLP and Semantic Predication platform for each knowledgebase, and will process the entire kb and load into the relevant knowledge graphs (one for TB, one for Nipah Virus). This engine will be set up to process new knowledge content on a daily basis.   6. We will develop a set of social media feeds specific to our target audiences (Scientists, Clinicians and Patients - for TB and for Nipah Virus) - and a news magazine that brings them back to the OSPF environment for enhanced engagement.   (ii) How our proposal will influence open research practices This project will help the OSPF to engage the research, clinical and patient community using very specific and relevant content. It will further the OSPF mission of using open source approaches to drug discovery and development for neglected third world diseases (such as TB and Nipah Virus), and this content will enable scientists and clinicians to better participate in our research and development efforts. (iii) how we will monitor and evaluate our proposal, including success factors 1. The progress of the project will be readily measured, including (i) the deployment of the 3 initial kbs', (2) the deployment of two knowledge graphs (TB and Nipah VIrus), and the deployment of the news and social feeds. 2. Uptake of this content will be measured by social media standards - numbers of followers, numbers of likes, numbers of reposts - for each content type (Scientist, Clinician, Patient) and Topic area (TB, TB drug resistance, and Nipah Virus). 3. Our ultimate goal is to increase the engagement of the scientific community with the OSPF, and to enable new drug discovery efforts in the areas of Tuberculosis and Nipah Virus treatment. Our success will be measured by subscribers to our social media sites, email lists and visits to our websites, and the initiation of new programs for the development of treatments for TB and Nipah Virus.

Decision

Funded

Comment on decision from Wellcome

This application has the potential to impact health research in the areas of TB and Nipah virus. The application would have benefited from more information about how the commercial platform would be used and funded past the end of the project.

Title

Enhancing youth participation in community-based health research

Lead Applicant

Dr A Doyle

Details of proposal

Vision: To develop an innovative and scalable method to engage and inspire young people in the co-creation of health research priorities Aims: To co-create a youth platform to solicit young people's health research priorities in the UK. Co-creation involves collaborative knowledge generation by academics, community members and other stakeholders (Greenhalgh et al, Milbank Quarterly, 2016). To test the feasibility and use of the platform amongst a diverse group of secondary school-age children in Bradford, UK. To identify youth research priorities and to disseminate to policy and research decision makers. To build a collaboration between LSHTM and Born in Bradford which is focused on youth engagement and open research. Target Audiences: We will develop a platform to collect inputs on research priorities from secondary school age students (11-19 years) living in deprived and multi-ethnic areas within Bradford, UK using crowdsourcing. Crowdsourcing is the process of having a large group, including experts and non-experts, solve a problem and then share the solution with the public (WHO crowdsourcing guide, 2018) . We will disseminate our findings (youth platform, research priorities) to young people, researchers, teachers, health professionals, policy makers, and commissioners. Lessons learnt in terms of youth engagement and open research practice will be shared with researchers and programmers. Activities: Form a Community Steering Committee in Bradford (with 50% youth) to develop all aspects of the research project in partnership with the study team. Youth platform development informed by two co-creation workshops with young people in Bradford. Feasibility study: Evaluate the platform in a crowdsourcing contest with 4 secondary schools in Bradford.; process evaluation including semi-structured interviews with young people, the study team, and other stakeholders; review of volume, diversity and quality of material collected Refine research priorities with local community: workshop with young people and stakeholders to refine and prioritise areas for research, and design dissemination activities Platform and guidebook refinement: co-creation workshop to refine youth participation guidebook; proposal development for larger scale contest scale contest: Dissemination of identified research priorities and feasibility study findings; open sharing of the platform and guidebook. (ii) how your proposal will influence open research practices in your field or more broadly; There is increasing emphasis on the importance of health research priority setting to identify the issues that matter the most to public, patients and health professionals (Mador et al, Health Res Policy Sys, 2016). However, there is limited guidance on how best to involve young people in health research priority setting (Hildebrand et al, Reprod Health Matters, 2013). This proposal aims to develop a scalable open access platform that can be used to crowdsource young people's health research priorities. We will work in partnership with young people and other stakeholders to co-create our methods and tools. We will share identified research priorities, and learning from the development of our youth platform widely with communities, academic institutions, the NHS, NGOs and others interested in adolescent health. We will do this through established BiB dissemination networks (including young ambassador and parent governor advisory groups, newsletters, social media and regular community events in community and school settings), the participating schools, and through online forums. (iii) how you will monitor and evaluate your proposal, including success indicators. The first monitoring milestone will be the formation of the community steering committee; with whom we will agree on additional milestones. Achievement of milestones will be monitored by the BiB executive. Potential milestones include: Creation of the prototype youth platform in ODK Drafting of the youth engagement guidebook and contest call for entries (including details on judging panel, criteria, how to choose topic for contest etc.) Completion of the contest. Dissemination of research priorities/insights Revision of the youth engagement guidebook including the terms of reference for the committee Preparation of proposal for further activities The success of the contest will be evaluated as follows: Contest metrics: Number of young people who participate in the contest. The contest will run in 4 secondary schools in Bradford, with an estimated student population of 3000. We would expect a successful contest to receive 100-200 entries over the two-month test period. Diversity of people contributing to the contest. A successful contest would attract entries from young people who are representative of the age, gender, ethnicity, and socio-demographics of the included school populations. Diversity and originality of contributions: as observed by the judging panel Short survey online and/or at school events: Satisfaction with the contest, suggestions for improvement, would recommend to a friend etc. Analysis of social media to estimate community engagement by measuring hits on contest website and dissemination materials Qualitative interviews (n=12) with youth researchers, contest participants, committee members, families/community stakeholders, and researchers to explore the appropriateness and feasibility of the youth engagement model (advisory committee, co-creation workshops, crowdsourcing), any beneficial and/or negative impact of participation, usefulness of the contest contributions/ideas, and success of the contest and dissemination activities in reaching the more marginalised and vulnerable young people and families.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal benefited from a strong team and an excellent evaluation plan. However, the level of innovation proposed was considered limited, and it was not clear which parts of the proposal advanced open research.

Title

Citizen Health Search

Lead Applicant

Dr Peter Murray-Rust

Details of proposal

Vision To put the citizen at the heart of published health research by making it more discoverable, in a shorter time with greater accuracy and relevance to produce faster, better evidence-based medicine focussing on obesity and dementia. And to create a citizen health search toolkit. The current medical literature overwhelms citizens: EuropePMC gives 450,000 papers for "obesity", How does a policy maker find the relevant ones? Or 200,000 papers for "dementia". Where does a patient group start? Most of these papers are irrelevant to the specific citizen. Through the platform we are developing, and bespoke dictionaries, users can rapidly restrict the hits to their specific area of interest (eg, school children? Refugees?). See attached additional information for details. We want to make health research output FAIR by developing, piloting and evaluating an Open Platform that will achieve the following: speed up the way of finding, sharing and re-using health research through the use of Text and Data Mining (TDM) technology make it easier and rewarding for citizens to:

find Open publications relevant to their interests

read those publications easily and with understanding

bring their own knowledge and resources to improving the process (this is particularly through contributing dictionaries) Wikidata is a game-changer; it's become the primary open semantic infrastructure for science and medicine. We've been funded by Wikimedia for dictionaries (WikiFactMine) and semantic medical search (ScienceSource). New developments in open data (OD), information science and crowdsourcing will revolutionize the evidence curation and synthesis process. Systematic reviews (SRs) are the gold standard for evidence synthesis and form the backbone of evidence-based medicine. However, they take too long to conduct (typically 2 – 3 years), and they require intense time input from highly trained and expensive experts. Making the process more efficient is vital; we already lack SRs on key clinical questions, and those SRs which do exist are increasingly outdated. Aim and target audience Our system will create an open index of millions of papers and metadata. This is designed for a better clinical trial systematic review framework, but also produces a completely general open health search platform. We'll make a dashboard where users can visualise, sort and annotate papers based on scores from dictionaries (drugs, conditions, organisations, countries, etc.). Our proposal is primarily aimed at reviewers, decision-makers and patient groups but it will equally be useful to other groups who want access to the medical literature (including charity volunteers, students, Wikimedians, etc.). Activities Our main activity will be the development of the system in response to community requirements and requests. The key principle for the software will be "don’t complete; interoperate". ContentMine are therefore building a system designed to work with tools by UCL (Klasifiki), Cochrane, Wikimedia and repositories such as EuropePMC and preprints (*rXiv). We expand the current reviewing process to automatically create FAIR metadata and (where allowed) semantic versions of the literature:

Semantification. PDFs are converted to XML using per-journal pubstyles.

Example: a Crowd reviewer using CCUHealth processing "PMCID:PMC5496305 Dietary diversity is related to socioeconomic status among adult Saharawi refugees" will automatically generate a semantically enhanced paper ("sectioned, annotated, XHTML").

Extraction of information from tables and diagrams. Dictionaries. Essentially a list of terms, with links to Wikidata. Many of these will be created by citizens. Reviewer reading the Saharawi paper creates a new term: "term="MDD-W" description="Minimum Dietary Diversity – Women (MDD-W)" description="Global Dietary Diversity Indicator for Women 2014" link="http://www.fsnnetwork.org/sites/default/files/minimum_dietary_diversity_-_women_mdd-w_sept_2014.pdf". An open index of key facets/metadata (dictionaries) applied to all papers read. Multidisciplinarity. The software allows information from any field to be mined (chemistry, genes, organizations, etc.). Multilinguality is supported by Wikidata. Influence over open access practices The result of the proposed work will enable reviewers, decision-makers and patients to identify research of interest far more reliably than can currently be achieved. Users will consume open access papers (or open abstracts from Microsoft Academic) Users will be generating dictionaries for others to see and contribute to Tool development and the work of the community champions will be recorded on Open Notebook Science Users will be incentivised to publish in open access format in order for their work to be included in this analysis system The result is a mixed community of professionals, volunteers and patients actively consuming medical literature far more efficiently and valuably than at present, overseen by a community manager from Cochrane Crowd. Monitoring and evaluation We will be monitoring the following milestones: Months 1-3. Code development and delivery to UCL and Cochrane for alpha-testing. Months 4-6. Interoperation of CCUHealth with Klasifiki and Cochrane and development of RESTful APIs Months 6-9. Code enhancement and feedback. Delivery to beta testers (UCL and Cochrane). Months 10-12. User queries and early adopter community Evaluation of the project’s success will track several KPIs: We aim for 1% of the Crowd to be involved (100 ppl) People from 20 countries to get involved in the project Dictionaries creation and size (both in-house (20) and contributed by the community (20)) Direct output (Structured Annotated XHTMLs): >>100,000 on dementia and >>100,000 on obesity Features in the index (100) Terms in dictionaries linking to Wikidata (50,000)

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an ambitious application was from a strong team. However, it was unclear if the project scale and scope could be achieved with the resources available. The application would have been strengthened by more detailed information about the links with Cochrane.

Title

Collaborative data collection platform

Lead Applicant

Mr Patrick Short

Details of proposal

Vision and aims Heterogeneous will build a technology platform in the form of a web application that researchers will be able to use to recruit and sequence eligible individuals for studies. Researchers would create profiles and fill out the details of their study on our system such as: Details of the study aims, description and design. Eligibility conditions of case and control participants that they are looking recruit. Sequencing technology they plan to use Details of the study aims, description and study design Once multiple researchers have inputted information regarding the details of their planned studies and eligibility criteria, Heterogeneous Connect will calculate the most cost efficient pairings of concurrent studies, allowing the researchers to coordinate and obtain the data they require for a fraction of the cost. Furthermore, Heterogeneous Connect would store all sequence data and metadata from the participants securely in the cloud so that it available for future research groups. Study participants would own their data, and could choose at any time to enrol or withdraw participation in one or all studies. Our platform will operate a dynamic consent model, allowing individuals to decide exactly how their data is used and to access relevant insights from studies they participate in. Importantly, each additional research project will add significant value to the research community by either adding sequence data for new individuals on the platform or enriching existing profiles with new data points. Target audience We aim to target research groups conducting medical and genetics research both in academia and industry. These research groups would be recruiting participants for human genetics studies and would be involved in one of the following activities: Obtaining the genotypes of large cohorts of individuals. For example, researchers conducting research into complex traits such as autism or Crohn's disease. Obtaining Whole Genome Sequence (WGS) reads from small cohorts of rare disease patients. For example, pharmaceutical companies testing the efficacy of therapeutics for rare diseases in clinical trials. Impact We see the value proposition of a platform such as this in the following three ways: Allowing concurrent research groups to share the sequencing costs on their study participants, saving as much as $250,000 for complex trait studies and up $500,000 for clinical trial recruitment. Allowing researchers from around the world who are working on the same disease to pool costs and share data between them in order to increase the power of their studies. Allowing additional lines of scientific inquiry to be investigated that would otherwise have been prohibitively expensive or example, by adding data points to existing profiles with whole genome data instead of sequencing new participants from scratch. The value of the platform is best seen when considering the following scenario in which multiple research groups are concurrently running medical research studies. Here we assume that conservative costs to recruit rare disease patients, complex trait patients, and control patients are £5000, £50, and £10 respectively. We assume that the cost to genotype an individual is £50, and the cost to obtain WGS is £1000. For researchers running complex trait genome wide association studies the cost to acquire and sequence individuals is relatively cheap, but the large sample sizes (n=5,000-10,000) result in large project costs. In a scenario where we have three Research Groups (A, B, and C) respectively investigating autism, depression and cardiovascular diseases, each group might want to recruit up to 5,000 patients and 5,000 healthy controls. Under the cost assumptions listed above, each study based on genotyping would cost approximately £800,000. Assuming a worst case scenario in which no participant with autism, depression or cardiovascular disease has been sequenced on the platform (and in which no other research groups have expressed interest in splitting the costs for these participants), each research group would save £200,000 by sharing the sequencing cost for the control group, thus a combined cost saving of £600,000. Similar cost savings could be applied to two research groups studying different aspects of a specific disease (e.g. diabetes) and looking to recruit diabetic participants. All three research group would then benefit from exclusive access to the data for a negotiated period of time. After this probationary period, all the data generated as a result of these three studies would then be tied to the participant profile and she/he would be deciding whether to share their personal data for subsequent studies. We believe placing the decision-making in the hands of the participant is not only a good way to remove conflicts of interests and mistrust between groups/institutions but also the right way to obtain consent and invigorate public engagement. This accumulation of data would then enable future studies to build upon existing findings, for example by studying relationships between autism and depression, or by complementing genotype data with DNA methylation arrays. Success metrics We will evaluate our project by calculating: The amount (in £) saved in sequencing costs to answer a given set of scientific hypotheses across multiple research groups. The number of active participants with genetic data enrolled on the platform and available to participate in studies. The number of scientific publications referencing data collected on the platform.

<u>Decision</u>
Not shortlisted

<u>Comment on decision from Wellcome</u>

This application proposed an interesting and innovative resource. However, it was not clear which parts of the platform would be open access, and so the potential impact of this proposal to transform health research through openness was difficult to determine.

Title

Open, complete, disambiguated database of authorship metadata in biomedicine

Lead Applicant

Dr Heather Piwowar

Details of proposal

Aims The aim of this proposal is to create an open, complete, disambiguated database of authorship metadata in biomedicine. We believe this data source will be rapidly integrated into open science toolchains, facilitating innovations not otherwise possible. Target audiences The target audience of this proposal is software developers in scholarly communication (both nonprofit and commercial), librarians, bibliometricians, and ultimately biomedical researchers themselves. Current ORCID approaches will solve the author disambiguation problem eventually, but it will take many years for all funders/journals to require ORCIDs for all authors. Even then, the back catalogue of papers will not be disambiguated. The current proposal addresses this problem by disambiguating all authors of papers in biomedicine and making this data wide open for integration and reuse. Activities We propose to disambiguate the authors of articles in PubMed, keep this data up-to-date as new works are published, and make this data openly available. This task has been completed before, but the results have not been kept up-to-date or made open for commercial use. Our approach will be to use methods from the literature that have proven successful for author disambiguation (work by Torvik, Smalheiser, Liu et al, Lerchenmueller, Sorenson), leveraging metadata available in Crossref, PubMed, and ORCID. We will assign an author cluster (named by an ORCID whenever possible) to each author position for each paper in PubMed. This means that for a given author we have a list of everything they've published that has been indexed by PubMed. Institutional information is a useful attribute in author disambiguation. As part of the disambiguation process we will model employment history (institutional affiliations, with year ranges) for each cluster. This information will be part of the released dataset. We will have a mechanism that makes it easy for authors or their delegates to correct mistakes made by the automated algorithms. Influence of this work on open research practices We expect the long term influence of this work on open research practices to be profound: we believe it will unlock a new wave of open research tools. Right now, all authoritative sources of author metadata are proprietary (Google Scholar, Scopus, Web of Science), not readily available for commercial use (Author-ity, Microsoft Academic Graph), or too incomplete to be usable (ORCID). With an Open source of disambiguated author information, developers will have the missing link to create an Open Google Scholar, an Open SciVal, an Open ResearchGate. In addition, this dataset will enable the evaluation of institutional Open Access policies using Open tools. Bibliometricians will be able to track the changes in openness behaviour, providing an evidence base for future open research policies, using open source and auditable data sources. Removing the need for an expensive subscription to Scopus or Web of Science allows a broader base of participation (globally, and outside academic institutions), saves money, and produces a more nimble, active, competitive set of alternatives. This will complement the work being done in Open Citations (i40c) and Open Altmetrics (Crossref Event Data), leverage the standards that have emerged for Open identifiers for people (ORCID), institutions (GRID), and funders (Crossref Funder Registry), and build on publicly available metadata for publications (PubMed and Crossref). This proposal is limited to biomedicine: the rich Pubmed metadata will help us get started. If the work in this grant is successful, in the future we will expand the author disambiguation service beyond Pubmed to all authors of all academic works. Monitoring and Evaluation The proposed work will be considered successful if it achieves high adoption. Within six months of release, we expect the community to have initiated multiple integrations, including the following: a search and link ORCID importer to allow authors to populate their ORCID profiles using this data software applications rolling out a type in your name and we will pull in your publications wizard a Europe PMC Open Author Profiles integration an R wrapper for analysts and bibliometricians several published analyses combining this information with other data sources, such as Unpaywall data to study Open Access adoption by region, institution, career stage, and other variables use by libraries to further populate their institutional repositories integration with wikidata multiple blog posts or preprints with proposed algorithm improvements (which we will then integrate into our code whenever possible!) We will be monitoring to see if these integrations and uses take place. If they don't, there are two main possibilities: The dissemination methods are too inconvenient The resulting author metadata is not high enough quality We should be able to determine which of these issues is the problem by conversations with potential integrators. If the former, we will refine our dissemination methods to better meet the needs of the community (different API endpoints, bulk releases of subsets of the data, ResourceSync, etc). If the data is not high enough quality, we will first understand what kind of errors are causing problems in integrations and then focus on a subset of the data, work on its quality until it is sufficient, and expand from there.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The application was from a good team with a strong track record, proposing to create an important database. However, the potential impact of this proposal to directly transform health research through openness was considered limited.

Title

Feasibiity sutudy to create a National Health Research Knowledge Management database in Rwanda

Lead Applicant

Mr Abidan NAMBAJIMANA

Details of proposal

Several health research activities in RBC are being undertaken by different divisions within RBC. The absence of a central Coordination requires individuals and divisions to archive and share data if any need arise, these data are usually retained by the divisions or individuals that produced them. Consequently, the status and quality of the research archives available for future use remain uncertain. Recently, MRC has initiatives of advocating the routine Health Research archiving and administration within RBC. The rationale for this is both scientific and economic. Research administration facilitates reinforces the collaborative and cumulative processes involved in creating scientific knowledge, it can also promote new research and enable the testing of new or alternative hypotheses. In addition, Research archive and administration can increase the transparency and accountability of research and bolster its reliability and authority by enabling other investigators to repeat or extend analyses. Addition to this, available divisional health research information is not efficiently monitored which makes secondary use impossible. This proposed National research database will ensure Systematic and secured data to value resources available for answering future public health questions. Currently, Rwandan health sector can only access Rwandan historical data from non-Rwandan platforms such as web of science and others. Therefore, The proposed National Health Research database will improve scientific knowledge platform at the level of Ministry of Health to enhance familiarity and ownership towards the Health research data, as well as facilitate the information sharing on research joint activities. Research data base to the National registry and archive in Health Sector as one of the most powerful tool for the knowledge management. This will become a national research registry to guide the development and implementation of National Health Research Agenda. The rational of this will be to ensure accountability in approval processes of protocols and monitoring of research conduct, systematic archiving and retrieval of research materials such as datasets, biological and clinical data (sequences or bio-images), avail Information for priority setting and research environment enhancement. who developed this platform to the RBC ICT staff who will be maintaining it at day today activities as well as initiating the issue of RBC knowledge management systematic review of available publication to inform decision makers on policy reformulation based on the evidence based information. The following activities are expected to be provided during the establishment of the proposed research database Working on publications so that they looking like APA referencing style. Embed referencing tools in the software. Include the number of visitors of the website. Add areas of interests of the PI among other details. Related reports to interventions Add guidance on how to conduct research on the public port Generate excel reports on database. Add another page which has the following features of the MRC (home, about MRC, team, services, and the database)

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was for a national-level resource that could be of value. However, the methodology was not clearly described and there was no evaluation plan provided. The proposal would have benefited from more detail on which data would be included in the resource and where it would be collected from.

Title

Distribution of neuronal types in the primate cerebral cortex: an online resource

Lead Applicant

Dr Piotr Majka

Details of proposal

The contributions of different neurones to perception, action and cognition are determined by their connectivity and synaptic actions. Understanding the organizational principles of this vast network is a core objective of contemporary neuroscience. For the first time we have the computational power to generate, store, and analyse large datasets, which enable increasingly detailed computational models and simulations. This approach has already resulted in fundamental insights into brain function, and new hypotheses about the origins of dysfunction. The impact of "big data" in understanding the mammalian brain has been felt at two extremes of a scale. In one hand, mechanistic insights at the cellular and molecular levels have resulted primarily from studies of the mouse brain, including major initiatives from the Allen Institute (alleninstitute.org) and the Human Brain Project (www.humanbrainproject.eu/). On the other hand, studies of functional and structural brain mapping, together with low-resolution molecular and connectional data, have been enabled by neuroimaging studies of the human brain (e.g. Human Connectome Project, humanconnectomeproject.org). There is presently a large capability gap in integrating these two streams of investigation. Open access resources providing access to cellular and molecular-level data on the brains of non-human primates are necessary to bridge this gap. Among the animal models suitable for such resources, marmosets (Callithrix jacchus) have emerged as the choice of several projects. Marmoset brains are relatively small (which facilitates obtaining and sharing comprehensive datasets, as well as computational analyses). Yet, they have all the specialised structural and functional characteristics of the primate brain, such as developed frontal and temporal cortical association areas, complex visual and auditory cortices, and frontoparietal networks which enable sophisticated planning and execution of actions (Solomon and Rosa 2014; Miller et al. 2016). The present applicants have implemented the first open-access online resource for sharing, and analysing cellular-level connectivity patterns in the cortex of this species (www.marmosetbrain.org; Figs. 1, 2). This has required the development of a computational pipeline for this purpose (Majka et al. 2016), which can be adapted for the present project, and has enabled us to conduct the first quantitative census of neuronal distribution in the primate cortex (Atapour, Majka et al. 2018, https://doi.org/10.1101/385971). Other groups have embarked on gene distribution mapping (gene-atlas.brainminds.riken.jp) and high-resolution MRI mapping of white matter pathways (www.nitrc.org/projects/nih_marmoset/). We propose to establish a publicly available, first repository of high-resolution images of the distribution of neuronal subtypes across the marmoset brain. We will build a resource which will allow interactive visualisation of excitatory and inhibitory neurons at different levels (from sections to high power views of cells, Fig. 3), and attribution of these cells to different brain areas (Fig. 1a). Further, we will incorporate online tools for quantification, based on purpose-developed deep learning-based object detection algorithms (Fig. 4). Like all data in www.marmosetbrain.org, the materials will be released immediately following processing, under a Creative Commons Attribution-ShareAlike 4.0 license. Funding will allow us to build a first data release, comprising 10 neuronal types (Table 1). Taking advantage of the already developed computational backbone, we will focus initially on the cerebral cortex. However, visualisation and quantification tools will be applicable to the entire brain sections, enabling follow-up work targeted to subcortical structures. During this period we will process and release materials obtained from young adult (24-36-month-old) marmosets, which have been the focus of our studies of connections in the brain of this species (Fig. 5). This will allow integration with other types of data for animals of the same age range, while reducing the investment needed for future releases that address developmental changes. The intended audience of this project is neuroscientists interested in developing realistic models and simulations of cortical function in the primate brain. In addition to data on connections (which we are building up in www.marmosetbrain.org) these require data on numbers of different cell types, and their allocations to cortical layers (the fundamental knowledge gap this project will begin to fill). Neuroscientists have enthusiastically embraced open-access resources such as the Allen Brain Institute portal, as demonstrated by the increasing number of papers using those datasets. Although work on primate brains will necessarily always involve smaller numbers of animals, the datasets are larger, and more complex. This creates a new set of challenges, which cannot be addressed until appropriate datasets are released. Thus, the present project will have an impact on triggering further basic research (for example, using transgenic marmoset models of neurological and psychiatric disease; Sasaki et al. 2009). The availability of this resource will also open opportunities in neuroscience education, being used as a "virtual microscope" for exploring brain structure. We will vigorously pursue dissemination of the outcomes of this project, not only through rapid outreach means (e.g. presentations at major conferences and focused symposia, bulletin/discussion boards) but also open-access publications. In the short term, measures of success will include access statistics to the web site, and contact from scientists interested in collaborations. Long term impact will be reflected in the usage of data in other groups' publications, hopefully in the forms of functional models that reveal new insights on the operation of the cortex, and evolution of the brain (by comparison with the mouse).

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal had the potential to create a valuable resource that targeted an existing gap in the neuroscience field. However, there were concerns over the appropriateness and feasibility of the approach proposed.

Title

Data Linkage and Model Sharing for Predicting Avian Influenza Outbreaks

Lead Applicant

Dr Kathleen Steinhofel

Details of proposal

Avian influenza outbreaks impact public health, food security and the economy, with a high impact risk of strains mutating and emerging to become transmissible in humans. The vision for our group of collaborators is to build a cohesive and integrated set of tools and data management procedures that make important modelling work on zoonotic diseases such as avian influenza easily and openly accessible to decision makers, academics, and other audiences from a wide variety of backgrounds. Sharing data and models to enable open research in this field is important because of the global public health challenges, and the impact on LMI countries. Moreover, recent technology enhancements make more data available which leads to an increased complexity when it comes to gaining deeper insights for more efficient protection and intervention against avian influenza outbreaks. To tackle these challenges it becomes necessary to take an interdisciplinary approach. Significant effort is required to ensure important developments from the various disciplines are communicated effectively. Furthermore, decision makers as well as scientists will benefit from gaining different views and comparing models derived from varying assumptions. Therefore, a tool for communicating work in the field is so important. We have existing work on three aspects of our open research vision. Firstly, as part of an EPSRC Global Challenges seed grant at KCL, the team has developed a geospatial database system to provide a single interface for accessing multiple datasets with different formats, geospatial resolutions and so forth. This was developed together with a prototype for an online geospatial machine learning tool. Secondly, UNSW has collated epidemiological datasets as a resource for influenza research, and included information on case species (human, domestic poultry or wild bird), country, date of illness onset, and if available, place of poultry infection (market or poultry farm). Currently, funded by a seed grant, the team are exploring enhancement of these datasets, and its linkage with avian influenza gene sequence data. And thirdly, a phylogeography tool, ZooPhy, has been developed by Matthew Scotch at ASU. The goal of ZooPhy is to enable phylogeography and virus migration to be studied by public health professionals. With that theme, it is designed for both advanced and beginner users such as public health epidemiologists who do not have training in bioinformatics. Whilst we have a track record of collaboration, the results exist mostly in separate systems. The aim for this grant is to create a unified system combining these individual elements. Specifically, this means combining systems and models such that we create an open online tool that combines machine-learning based risk computation, geospatial analysis, phylogeographic computations, epidemiological data and molecular risk factors. For the nature of our system, we have four aims. Firstly, extensibility through modules, i.e. the system must be made to easily grow and develop over time. Secondly, it needs to focus on communicating inferences and key outputs of models to a degree such that it can be understood by a broad user group. Thirdly, it must include a comprehensive help sub-system to be accessible to a larger audience. Finally, it will support open formats via various data-download options to enable export of analysis results to which further widely used tools might be applied. See Figure 1 in the additional information for a high-level overview of the system. We aim to stimulate open research practices in both an active and a passive way. The passive method is achieved naturally, simply by making the experimental results on zoonotic disease modelling available through this tool, by enabling easy reproducibility of our work, and by providing export functions to allowing further analysis by academics in the field – in our discipline or otherwise. We believe that the convenience and confidence this brings, will naturally lead to better dissemination of advances made – both in terms of accessibility, and inspiring academic works citing or building on it. In the future, one of the options open for this tool is to allow other academics to build third-party modules for it and integrate into this tool as well. In terms of actively pursuing open research practices in our field, in conference, workshops

and papers, we will point out the benefits of making work open, reproducible and easily available, and we will use this tool as an example of how it could be done and what the effectiveness of it is. We also hope to inspire industry to engage more with academia by sharing data and models. Progress during the proposal is monitored in two areas, namely, the technical build of the tool, and its dissemination. For the technical side, we will have a continuous integration system in place with weekly milestones, and some leeway for unexpected problems. Throughout, we will regularly consult with the WHO Collaborating Centre for Reference and Research in Influenza via John McCauley. A complete prototype for evaluation is expected in month 10, see Figure 2 of the additional information. Regarding dissemination and receiving feedback on what aspects are particularly beneficial for various user groups, we intend to communicate with key organisations such as WHO, FAO and OIE; and interested parties through meetings and conferences.
Decision **Not shortlisted**
Comment on decision from Wellcome *The applicant opted not to share this information*

Title

Avicenna - An open data platform to support the early detection of complex clinical conditions

Lead Applicant

Dr Kevin Koidl

Details of proposal

This proposal aims to support the early detection of symptoms related to complex medical conditions based on capturing the cognitive processes of a healthcare professional when they are evaluating a clinical condition from unstructured textual data from sources such as the web. To achieve this a Machine Learning and Natural Language Processing Model is to be created that can be applied in a wide range of applications to support early symptom detection for healthcare professionals. We envisage that these models in the near future will be embedded in applications including call and helplines, chat interfaces and other text or speech based application. This will enable the support for decision making by healthcare experts in detecting complex clinical conditions. The overall objectives are to decrease condition detection time, misinterpretation of symptoms and increase cost efficiency. This first data model, which will target Sepsis, serves as a trial model to evaluate the overall approach with the long-term goal to apply the same for the creation of other complex condition early detection models, such as dementia and depression. The main aims are: 1) Creation of a Machine Learning-based Natural Language Processing Model trained by one of the leading Sepsis Consultant in Ireland. 2) Creation of a web-based application to facilitate the training of the model with the goal to apply it for the creation of more models in the near future. 3) Evaluation and comprehensive testing of the created model with international experts. For this specific Sepsis-related model the target audience is mainly experts and professional staff that rely on the early detection of complex and often combined symptoms that combined created a clinical condition. In the first instance, the target audience is the Sepsis expert who helped the team to train the model. After which the target audience is to be extended to an international expert panel in the area of Sepsis. The final target group are the professional staff of hospitals that have to detect symptoms quick such as in a helpline. The main activities are as follows: 1) Detection of text and speech based positive examples of Sepsis symptoms for the creation of an initial sample training set (Manual creation of initial dataset together with Sepsis expert). 2) Extension of the initial dataset by deploying Natural Language Tools (among others FREME) that automatically detects similar text samples that describe Sepsis symptoms that and which are similar to the data used for the manually created dataset. (Automatic extension of the initial dataset). 3) Development of web-based interface to annotate data samples by experts to create a complete data model. 4) Evaluate data model with new text examples. 5) Creation of a deep learning network (RNN - Recurrent Neural Network) based model for continuous classification (training) of data samples with the goal to reduce the overall error rate. 6) Testing the resulting model with an expert panel and professional hospital staff. The resulting approach and data models will provide openly accessible data that enrich the state of the art in Machine Learning approaches (specifically Natural Language Processing) within the healthcare space. This approach, therefore, will extend the mostly flat and keyword-based approaches that are unable to detect connected keywords throughout the text sample. The evaluation will be two-fold. The first evaluation is quantitative and based on data sample splits in which an initial sample is used together with a test sample to enable system based and statistical (error rate reduction e.g. false positives etc.) evaluations. The second evaluation is qualitatively and based on user trials which include feedback loops that help the re-training of the underlying model.

Decision

Not shortlisted

Comment on decision from Wellcome

In this application a potentially useful resource was proposed. However, it was not clear which parts of the proposal advanced open research, and the proposal would have benefited from a more detailed evaluation plan, for example identifying targets that would indicate success.

Title

MouseBytes: Cognitive Data Integration and Sharing Platform

Lead Applicant

Prof Timothy Bussey

Details of proposal

Vision: Neuroscience is undergoing an open-access revolution. Databases in areas such as neuroimaging and genomics have provided a paradigm shift in the way we are now able to analyze, share, and re-analyze vast amounts of data from multiple laboratories. Unfortunately, one area that is very far behind in this respect is behavioural neuroscience. This is a problem, as assessment of behaviour in animal models is a critical component in our understanding of normal brain function, as well as understanding and treating brain dysfunction in neurodegenerative and neuropsychiatric diseases. However, conventional behavioural testing does not lend itself to open-access sharing because the methodology is inconsistent, and the data are not generated in digital formats that lend themselves to deposition in databases. However, we are at a watershed moment in that one automated and highly reproducible behavioural method that is ideal for this application – touchscreen testing – has reached a critical mass of users, around 300, such that behaviour using this method can now join the open-access revolution. To capitalize on this opportunity, we have developed MouseBytes: a web application that can access and organize touchscreen cognitive data deposited in an open access database allowing for data storage and validation. We envision that MouseBytes will be the premier platform to allow cognitive data to become open access, searchable and to be reused and re-analyzed. MouseBytes will build on the fully automated and standardized touchscreen technology we have developed to allow the community to take full advantage of data generated with this technology. Aims: Touchscreen technology is being used by close to 300 independent laboratories for deep phenotyping of cognitive function in mouse models, leading to generation of large amount of cognitive data. We have designed MouseBytes to check the content of uploaded data files against potential errors; hence, making data suitable for comparison and validation. We propose to enhance the functionality of MouseBytes to attract users to build a comprehensive database of cognitive function in mouse models. For this overarching goal, we have 3 specific aims: Generate graphic data visualization in MouseBytes Attract 20-50 independent laboratories to become MouseBytes beta-testers Introduce MouseBytes to Scientific Journal Editors We anticipate that these 3 aims will help in supporting the goal of having all touchscreen cognitive data as open access and fully available for the community. Target Audience: Scientists interested in conducting cognitive analysis and investigating brain function and the underlying causes of brain disorders. Editors of scientific journals, publishers, scientific societies, grant agencies and research institutions interested in reproducible and open access science. Activities: MouseBytes is a web-based application allowing its use without any software installation. Users can download the data as CSV files and process as required for reuse under Creative Commons (CC0) licensing. We propose to create an add-on application for MouseBytes for easy graphic data visualization to increase the value of MouseBytes for the community. Data from experiments downloaded from MouseBytes will be automatically used for generation of graphics. MouseBytes allows for links between published manuscripts and the data collected by using Digital Object Identifiers (DOI) to link raw data to a published manuscript. However, MouseBytes can also be used to upload non-published data which can be optionally private and so not shared with others until publication. We encourage authors to share their data once an experiment is finished. However, we recognize that many scientists will prefer to share their original and raw data only after publication. Hence, we will maintain the option for keeping data private until publication of manuscripts. We will use our extensive scientific network to introduce MouseBytes to laboratories using touchscreen technology. We will design interactive videos, presentations, and data visualization to facilitate MouseBytes use for authors to deposit their data. We will use our extensive network of scientific contacts with Scientific Journals, editors and publishers (Editors or Associate Editors of 4 premier

Neuroscience Journals) to introduce MouseBytes to the community. Our ultimate goal is that similar to genomics and imaging, datasets for cognitive assessment in mice will become fully available for the community and will be required to be deposited in an open access mode to accompany publications. This service will be offered free of charge for authors, editors and other users. Our University has committed to maintaining the open access MouseBytes using a server for 6 years. Proposal influence in open research practices: By allowing storage of non-processed data, but also providing tools for quality control and data visualization, MouseBytes can significantly reduce the redundancy of data production, increase data availability, reproducibility of cognitive research, and collaboration, and ultimately accelerate the translation of research findings to clinical practice. Authors will retain copyright of their data under CC0 licensing, and the data downloaded from MouseBytes will be available for users by citing the original source of the data. These efforts will increase the transparency and data availability for the community. Success Indicators: We anticipate that within this year data visualization will be implemented, 50 laboratories will be using MouseBytes to upload their data and editors will become aware of this new platform to increase data validation and reproducibility when using touchscreen for cognitive assessment.
Decision **Not shortlisted**
Comment on decision from Wellcome *The applicant opted not to share this information*

Title

Open Africa

Lead Applicant

Dr Kevin Tyler

Details of proposal

OVERVIEW In the summer of 2018, UEA financed a meeting of 25 life scientists in Arusha, Tanzania, (Fig 1.) to explore how African scientists could make better use of their native plants, and their metabolites, with observed/potential medicinal properties. We concluded that a multi-disciplinary team could provide support and advice on scientific, economic, business and legal changes to help create an environment where these compounds could be produced and sold, safely and commercially, with health benefits. To achieve this we will need to develop a website, a wiki site and a phone app so that people can photograph and geotag the many medicinal plants already used in Africa for health reasons. (i) VISION Africa's ethnobotanical pharmacopeia offers a remarkable indigenous resource that can be exploited to help address the Neglected Tropical Diseases (NTDs) with greatest impact in this region. African scientists are eager to provide local solutions to the NTDs which affect their communities. However, many lack reagants, trained scientific help, and multi-disciplinary advice, especially concerning patenting or intellectual property as it applies to medicines. Many also lack the necessary more advanced equipment. AIMS: This project aims to establish an open research platform OpenLabAfrica enabling African researchers (and collaborators) to record and share their research resources/discoveries/outputs in a manner that will further empower them to maximise collectively their sustainable research activities. The focus of action is sub-Saharan Africa, where the economic cost of NTDs is most acute. The communications technology on which the Open Science strategy (LabTrove) is based, was pioneered by Frey (Southampton) in 2013 [http://www.labtrove.com]. LabTrove will be further developed for 'Open Lab Africa', and integrated with other social media tools. Target Audiences and Participants Students, tutors, scientists, pharmacists, business leaders from the health sector, traditional healers, legal advisers and policy makers. (ii) RESEARCH PRACTICE To co-ordinate more open research practice amongst our partners we will develop a website, a wiki site and a phone app for photodocumenting, geotagging and annotation of the many medicinal plants already used in Africa. The communications technology on which the Open Science strategy (LabTrove) is based, was pioneered by Frey who is a Co-I on this project. It supports the Open (Science) Source Malaria project http://opensourcemalaria.org/ to address malaria in India. Building on these firm foundations, LabTrove will be further developed for 'Open Lab Africa', and integrated with other social media tools. We will host a 2-week LabTrove training workshop for three representatives of the African network (plus 8-10 UEA members of the Open Lab Africa consortium), coordinated by Frey at UEA. "Training of Trainers" activities will be provided to the visiting African Partners (Omolo, Bathelemy, Becker). The workshops will also involve team members planning and developing prototype templates that can be used to store different experimental procedure and data formats (Natural product isolation, structure elucidation, screening, chemical synthesis & optimisation). Whilst all users of LabTrove can create their own templates for data storage, provision of some standardised formats is expected to increase uptake by members of the OpenLabAfrica research community as it grows. UK team members will also be trained in and utilise the Labtrove-Africa platform to upload relevant recent (as well as all future) research procedures/results. The African team members will return to Tanzania, Cameroon & South Africa where, they will coordinate the propagation of Labtrove training activities within their own research-groups/institutes/countries as well as (within the timeframe of this project) travelling to institutes in other OpenLabAfrica partner countries (Kenya, Ghana, Democratic Republic of the Congo, Uganda) (Fig. 1) to provide training to a broader field of researchers. It is noteworthy that Becker is the coordinator of the South African Biochemistry and Informatics for Natural Products (SABINA) network (www.sabina-africa.org/sabina/) which includes Natural Products research groups within 8 South African universities as well as others in Malawi, Namibia and Tanzania. Propagating Labtrove uptake across the SABINA network will enable our Open Lab Africa network to connect and extend its Open Source Research activities into new areas. A freely available App (OpenAppAfrica) will be developed enabling the general public to photograph and publically archive images of native medicinal plants that can be further curated by experts (Fig.2) A website www.openlabafrica.org will be designed through which all these resources can be accessed and applied. (iii) MONITORING AND SUCCESS INDICATORS We have already purchased the web domain www.openlabafrica.org, which is where we will develop the website linking all the open lab resources and activities. This will host links to our bespoke LabTrove domain (and its accompanying wiki) as well as a repository for information created through user activities with the OpenAppAfrica. (Fig.2) As administrators of the website, labtrove and app databases, we can monitor user metrics (eg. number of clicks per week/month and their geographical distribution). Members of the network will be registered for use labtrove as they are trained and so successful uptake and amount of usage will monitored. Likewise, use of the natural products database, number of species captured and annotated and all access will be logged. An early success indicator we'd like to see is a sustained increase in website/database activity during the 1-3 year following their setup.

Decision

Not shortlisted

Comment on decision from Wellcome

The proposal was to develop an innovative resource and benefited from an international team. However, it was not clear how content added to the platform would be curated, and there was no consideration of which open licence the data would released under.

Title

REDER - Repository of Epidemiological, Data cleaning, and Economic R code

Lead Applicant

Dr Laura Webber

Details of proposal

We accompany this brief overview of the project with two attachments: 1. providing an example R script we have produced; 2. a schematic to illustrate the user interface and some functionality that it will contain. Aims Review and collate existing epidemiological, data cleaning/extraction and health economic code into a single public GitHub repository Create a user-friendly interactive web-interface for accessing the repository and support both expert and novice users in data cleaning, manipulation, extraction, and analysis Use agile development practices through surveys to understand user needs Target audiences We envisage students, researchers and analysts who are involved in the analysis of epidemiological health and economic data will use the repository. Some scripts will be applicable to a wider range of datasets with minor modifications. R is chosen because it is free and highly advantageous for structured data manipulation, visualisation and statistical analysis. Though the tool is still useful to non-R users or novice users since the user friendly interface will enable code to be run without prior knowledge of R. Activities 1. Google search of existing open source code for manipulating and cleaning epidemiological and economic data 2. Carry out an initial consultation to establish user needs and understand repeated practices used across the epidemiology and health economic community on established datasets. Users will be drawn from our existing public health, data science, and health economic networks 3. Development of an interactive web-interface in R shiny incorporating feedback from users (see attachments) 4. Extract code identified and store in a single repository. Any code which has been produced in another coding language will be translated into R 5. Write a user guide including worked examples 6. Monitoring, evaluation and dissemination Influencing open research practices This project will influence open research practices by improving research reproducibility and developing open platforms for collating data analysis tools. We will develop a "Repository of Epidemiological, Data cleaning, and Economic R code" (REDER) to prevent researchers having to redo essential cleaning and analysis tasks on well used datasets. REDER opens up research because data cleaning processes are published, archived and searchable. This will build on existing generic repositories for data cleaning (e.g. https://github.com/leriomaggio/getting-and-cleaning-data) by being specific to cleaning and manipulation of epidemiological and health economic datasets. This repository will contain code that UKHF and others have developed for their recoding of the datasets. An example might include the difficulty with coding ‘smokers’ in a longitudinal dataset such as Whitehall II. In one wave of the data a person might identify as a smoker, but in the next as a never smoker (not possible if one has already been a smoker), and the subsequent years as a smoker again. Making assumptions about how to treat such a person in the analysis of epidemiological trends is important. Sharing this thinking and these assumptions enhances reproducibility of analysis and also enriches how we think about analysis of complex datasets. In addition to data cleaning and reproducibility, the UKHF has already developed an array of existing epidemiological tools for managing missing data and calculating epidemiological and economic metrics. Examples include calculating incidence from prevalence, calculating survival from incidence and mortality, relative risks using longitudinal data, or Incremental Cost-Effectiveness Ratios (ICERS). Our previous work has mainly focused on developing mathematical models, and complementary tools using C++. These tools will be translated to R scripts and included in the repository. The repository will be a living entity such that researchers can implement adjustments to the existing R scripts and new R scripts related to specific datasets via the UKHF who will monitor and check its validity. The code will be annotated by the researchers who post it ensuring user-friendliness and transparency. The web-interface will be hosted on the UKHF website and the repository will be available for those wishing to use the R Scripts without the web-interface. Monitoring and evaluation Monitoring

1. Agile development practices will be employed to ensure tool development is aligned with user needs. Users will be recruited via email using snowballing. An initial survey monkey consultation on the ideas, concepts, and scope of the tool will be circulated via email through our existing epidemiology and data science networks. Within this email we will request that recipients pass on the consultation link to other interested parties for completion. Success indicator: 25+ responses 2. We will develop an analysis plan with associated timeline to monitor progress. Success indicator: meeting deadlines 3. The UKHF holds an up-to-date risk register which will ensure that potential risks are mitigated. One potential risk is changes to staffing. However, all staff are currently in place and a wider team is available (including consultants) to mitigate this risk. Evaluation An evaluation survey will be included in the circulation of the final tool to the users consulted in stage 1, and a feedback survey embedded within the interface for ongoing user feedback. We will also carry out an internal evaluation of what worked well and didn't work well in the project to carry learnings forward to new projects. Success indicators: 1. Number of hits on the website 2. Citations of the REDER tool in peer-reviewed publications

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The application was from a strong team, proposing to generate a useful resource. However, the potential impact of this resource to transform health research through openness was limited. The application would have benefited from more information about how the team would monitor, evaluate and disseminate the resource.

Title

BrainBox Badges: Online certification for collaborative annotation of brain imaging data.

Lead Applicant

Miss Katja Heuer

Details of proposal

i Vision   Our vision is to create an interactive space where academic and citizen scientists can work together to study public neuroimaging data. Instead of multiple isolated studies, we want to enable a worldwide collaborative effort to curate, analyse and understand the data to which all of us have access. This space should be such that all researchers feel welcome to contribute with their different levels of expertise, and can learn and participate with their creativity towards understanding our brains and minds. The fruits of this common effort should be open, easy to find, access, and re-use.   BrainBox is our Web application implementing this radical open science vision. BrainBox allows users to view, curate and annotate any neuroimaging dataset available on the Web, currently >13,000, and provides real-time interactive tools for manually segmenting and editing brain regions (3 min video about BrainBox: http://tiny.cc/brainbox).   Aim. For neuroimaging research to embrace our vision we need a way to assess and develop contributor's expertise. Researchers leading a BrainBox project need confidence about the quality of the results, and users need a clear path to start contributing to a project.   We propose to develop training and certification modules for BrainBox: BrainBox Badges. Using the training module, researchers will be able to explain the tasks that are required. Users will be proposed examples to practice, and then the evaluation module will test them on new data. Users will earn a badge for their achievement, which will be added to their profile.   Through BrainBox Badges researchers will be able to recruit collaborators, weight their contributions based on expertise, improving their confidence on the results. Users will have a clear path to start contributing to projects. Badges will increase their reputation, motivating them to keep learning.   Activities. We will design and code BrainBox Badges based on the Open Badges standard (https://openbadges.org). We will build 3 sample training and certification modules: data quality annotation, manual segmentation, and automatic segmentation correction.   We will host 2 sprints at Institut Pasteur aimed at academic researchers. We will also organise 2 "Brain Challenges" online during Hacktoberfest (https://hacktoberfest.digitalocean.com) and as a Savanturiers workshop (https://les-savanturiers.cri-paris.org) aimed at a more general public. They will help us detect and fix bugs, test our tool's useability and record material for tutorials.   Target audience. Our main audience is the neuroimaging community: a very diverse group of researchers with backgrounds from cognitive sciences, to medicine and physics. Our tool should have the precision, performance and reliability they require. We believe, however, that many tasks are accessible to a broad range of citizen scientists, patients, caregivers, or school students.   ii Influence on open research practices   Broadly, we hope that our project will have an effect on the research community similar to that of the Wikipedia on society: we want to facilitate and encourage open collaboration in our field. More precisely, we want to: Increase trust in open collaboration by increasing trust on the contributions made to one's own BrainBox project, and trust on the results made available by other projects.   Encourage researchers to use public data by making it easier to find and use it. Imagine having to download the Wikipedia just to read it or edit it – that is what our community currently does, and what we want to change.   Improve transparency, by making data annotation open, fostering consensus and making different points of view explicit. The creation of Wikipedia has made it harder for isolated websites to publish unreferenced information, we aim at having a similar influence on our community.   Allow our community to tackle much larger projects than before, through an incremental, less wasteful process. Currently, when public data is downloaded, it is analysed redundantly by every research team in isolation. This is at best like the endless threads of Word documents shared by e-mail of the time before Google Docs. The clumsiness of our collaboration methods leads to wastefulness: unable to quality control and fix the results of automatic algorithms, we end up discarding up to 30% of the data.

Inspire the creation of a new infrastructure for collaboration, covering other fields of biomedical research. We are ourselves participating in the creation of tools such as Brainspell for the collaborative annotation of the neuroimaging literature, and more recently MicroDraw for the collaborative visualisation and annotation of high-resolution cell-scale data. iii Evaluation and success indicators We will measure the impact of BrainBox on user engagement: we aim at attracting at least 1000 per month, and reach 300 users opening an account. We aim to certify at least 200 users, gather 50 in-person participants during the sprints and 100 online at each Brain Challenge. Engagement indicators will be collected using Google Analytics and the BrainBox API. We will conduct online surveys on the perceived benefits of BrainBox. We will measure the number of projects created, and their activity, and aim at tripling the current 15 active projects. We also consider an indicator of success the publication of articles using BrainBox.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to introduce a certification element to an existing crowdsourced platform, which had good potential to impact the field of neuroimaging. However, the level of innovation proposed was considered limited, and the proposal would have benefited from more detail on how the team would publicise the badge system, sprints and challenges.

Title

Open cloud computation tools for image segmentation

Lead Applicant

Dr Matthias Haberl

Details of proposal

With the below outlined new features for CDeep3M we expect to close the gap between the state-of-the-art deep learning-based image segmentation and the more tedious and less precise methods now still widely used by the biomedical community, with no expertise in application of deep learning-based approaches. We further expect those improvements and additions to reach an even wider target audience and further improve performance by integrating novel algorithms, as they are developed in the field. In aim 1 we will integrate CDeep3M as an ImageJ plugin, which will be able to execute all required standard functions, while still running on the cloud. The plugin functionality will include launching the cloud formation (which brings up an instance where CDeep3M is automatically installed and operational, a feature already implemented on our GitHub repository), copying training data and image data to the cloud, launching training, performing validation and plotting loss and accuracy, launching prediction, copying results back to the local computer and deletion of the cloud formation stack. ImageJ is likely the most widely used tool in the biomedical imaging community. However, our proposed concept is different from most ImageJ plugins, since the image processing is not performed locally, which would require the end user to work on a high performance computer. The reasons instead, why we intend to implement these basic functions into ImageJ are, first that ImageJ is already so widely used, most users will not need to install any software (or e.g. deal with a complex online portal to copy data and execute functions) and are already familiar with its design and functions, and second that ImageJ plugins are functional across-platforms / operating systems, which is a tremendous advantage over designing a standalone program. Altogether, we expect a broader target audience will be reached through the ImageJ implementation. In aim 2 we will implement additional network architectures, to ultimately improve the performance of the image segmentation algorithms and therefore reduce required post-processing. We will benchmark the different algorithms against one another (training time, prediction time and accuracy) and provide end-users with the choice of several algorithms. Importantly, we expect by our continued development of CDeep3M, to set a new standard in terms of openly providing functional code, even in cases where high performance compute resources are required. Deployment through cloud templates implies an end to software bugs and crashes, since developers are more directly accountable if those occur (please see https://doi.org/10.1371/journal.pcbi.1005994 for more details on the numerous advantages of cloud computing). To date CDeep3M is to our knowledge the only cloud implementation for biomedical deep learning based image segmentation but we hope our continued efforts will encourage other groups to release their code in the future together with cloud formation templates, to circumvent complicated scenarios and failed attempts of installing and testing code. In aim 3 we will port our cloud formation template (currently only available for Amazon cloud) to alternative cloud providers (Microsoft Azure and Google Cloud), as well as the cloud-based computational reproducibility platform 'Code Ocean'. We expect that by enabling alternative cloud providers, we can give more flexibility for end-users and reach a broader target audience. We aim to accomplish aim 3 in two steps, in the first step we will port our code into a container in CodeOcean, which provides free cloud access to all researchers for a limited amount of hours per month. This serves the purpose of reproducibility, as well as allowing users to quickly evaluate if the code is suitable for their research question. If users are convinced it is useful for their research purpose and find that they need to process large amounts of data, they can use one of the commercial cloud providers, which provide access to high-end compute resources, starting at 0.9$/hour. With recent advances in our development TB sized image volumes can already be processed below 200$ (on cloud compute hardware which would cost ~$100000 to buy). However, in the interest of reducing the dependency on a single cloud provider, we wish to port our code also to Google Cloud and Microsoft Azure, giving users more flexibility. We will monitor the success of our project in several ways, for one traffic on the GitHub repository is an indication for the overall interest. We will insert a simple counter for each cloud instance, which is launched from our GitHub repo. An increased number of users will be a positive indicator for our success. We will further periodically conduct short surveys to receive feedback on newly implemented features and the layout of the ImageJ plugin.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a good proposal which had the potential to add value by building on an existing resource to increase its potential usability and reach. However, the potential impact of this proposal to transform health research through openness was limited. The application would have benefited from more information about the long term sustainability of the tool.

Title

Investigating the value of analytical studies of Peer Review methods (including Liquid Peer Review) using a computational, agent-based modelling approach.

Lead Applicant

Dr Arty Ruseckiy

Details of proposal

Knowledge production, as Michael Polanyi pointed out[6], cannot be subjected to centralised control without loss of 'requisite variety'[7]- that capacity of a system to respond adequately to relevant attributes of its ground reality. Since the ground of science is base reality, variety is highly valuable. It is thus desirable that knowledge output will have high variability. This brings a secondary problem - filtering; we require means to direct attention to knowledge of high quality and clear salience. As the quantity of published output rises, this becomes increasingly important and problematic. Peer review has come to reference a range of techniques for filtering, but has been the subject of a range of criticisms[8], in addition to being problematically implicated in the economics of science publishing[9]. Peer review studies are scarce. This proposal has been designed to develop novel techniques for investigation and to lay groundwork for further study. The proposal is to investigate and develop models for several interlocking issues, as follows;

An empirical model of paper and review attributes as seen by researchers.

A model of the relation between paper and review attributes and their distribution,

A model of author characteristics as a basis for simulation of populations of researchers. A model of how real researchers behave through a 'Liquid Peer Review' process. A model for relative assessment of peer-review quality under different models of production. These models will be used to simulate a variety of approaches for analysis. Method A sample of papers and associated reviews will be sourced from open peer-review journals. Participant researchers with relevant knowledge will assess these to produce data for use in modelling. They will also participate in a Liquid Peer Review exercise. Participants will be split into groups 'A' & 'B'. 'A' will study a subset of papers, identifying key attributes that reviewers respond to. These will be tabulated against published metrics. 'B' will study reviews of the same papers, identifying key attributes reviews have responded to, alongside quality metrics. Multi-dimensional models of paper- and review- attributes will be developed. 'B' will score the remaining papers against the paper-attributes model. These will be tabulated alongside metrics for paper and lead author. A distribution of scores will be derived for each paper-attribute, correlated against author metrics, to give a parametric distribution model of paper attributes related to author metrics. 'A' will score reviews of the same papers against the review-attributes model. A distribution model of reviews will be derived, correlating review attributes against paper attributes. Finally, a distribution model of author metrics will be developed from published data. Agent-based model A sample of simulated 'researchers' will be created using Monte-Carlo techniques across the author metric distribution model. Each of these 'researchers' will be assigned a 'reviewer profile' obtained through Monte-Carlo techniques within the 'review-attributes' distribution. Thus we derive a population of simulated 'researchers' with particular reputations and characteristics they bring to bear when reviewing papers. These constitute the 'agent' population used in simulations. Liquid Peer Review Under this approach, researchers freely assign their reputation to other researchers for review purposes. This process can happen recursively, so that a researcher having been delegated the reputation of other researchers can delegate another, and so on. Crypto-economic platforms offer the prospect of 'earned' reputation tokens. In this model, each reviewer carries the 'weight' of all the reputations assigned them. Review 'value' is varied with the 'weight' of the reviewer. Study participants' public science output will be anonymised and each will 'score' a number of others for status. A number of liquid peer review delegation rounds will be carried out with these anonymised personas, and a model derived. Simulations These data, models and agents will be combined to produce four Scenarios of papers and reviews as follows: Real papers / simulated reviews. Simulated papers / simulated reviews. Real papers / simulated liquid reviews. Simulated papers / simulated liquid reviews. In each scenario, simulated reviews (represented by combinations of attributes) will be generated on the basis of the model relationships between paper attributes and reviewer profile, and according to the model of liquid review relations. Simulated papers will be produced, having attributes assigned by Monte-Carlo techniques according to the distribution, with respect to author metrics. Each project stage will be reviewed before proceeding - design adaptations may be needed. Two are key: adequacy of paper/review and participant pool sample sets as a basis for study, correlation of Scenario 1 outcome against real reviews. The project does not set out to prove any hypothesis, but to develop robust modelling techniques for a key aspect of knowledge filtering and look for any difference in review quality arising from the 'Liquid' approach. A successful outcome will provide; Well structured data-set repository suitable for use by others. Framework for modelling paper and review attributes that delivers clear characterisation. Published models that can be developed by others. Scenario outcome data susceptible of robust analysis. Three to five publishable papers on aspects of the study Further, it is hoped that the idea of building computational models of science processes can provoke a debate as to the viability of 'evidence-based' self-governance within the world of science.

Decision
Not shortlisted

Comment on decision from Wellcome
This was an interesting proposal which sought to study peer review using blockchain technology. However, the methodology was not clearly described, and the evaluation plan would have benefited from more detail, for example identifying targets that would indicate success, or seeking user feedback.

Title

Development of a high-quality, low-cost, open-source hemodialysis machine. Phase I: Device design & prototype development

Lead Applicant

Dr Tarek Loubani

Details of proposal

Vision The goal of our medical research project is to develop, validate, certify and disseminate high-quality, low-cost, open-access medical equipment. The project relies on two pillars: the use of 3D printers and other rapid prototyping technology; and leveraging Open Access and Open Source principles and devices to decrease development costs and disseminate results to stakeholders. This model has been proved with two simple medical devices (the stethoscope and the tourniquet), with two more complex devices nearing completion (pulse oximeter and electrocardiogram). The main question of our research is: Can the successful model that developed, validated and deployed other low-cost devices also be used to develop much more complex devices such as a hemodialysis machine? Over a 12 month period the Glia team and their collaborators will complete the first phase of our research, device design and prototype creation. Our aims for this project are: Investigate what off-patent, open-source work has already been achieved that can be built upon Acquire access to appropriate ISO documentation and resource the international standards Consider the different types of end-users for the device and make a plan for their needs, such as those in low-resource and low-income communities Consider end-user experience in design, such as; using materials and building equipment that are easy to source in low resource settings and creating a device that resembles and operates as existing devices do today Consider the most appropriate device design for communities that have low-access to device disposables Construct a device that functions mechanically and electronically Ensure this device is low-cost, easy to use, and widely-accessible Meet regulatory requirements for Health Canada Engineer a usable prototype ready for clinical trials Subsequent to the terms of this proposal, we will conduct validation and clinical trials to gain regulatory approval through Health Canada as a Class III or Class IV device. This will result in a final version of the device ready for distribution. Influence Glia's stethoscope and tourniquet have already made a significant impact. Replicating this success with a low-cost hemodialysis machine will provide kidney patients around the world with access to one of the most essential medical devices. The availability of specifications for generic manufacturers to manufacture devices and the subsequent downward price pressure on premium brand manufacturers will increase the standard of care even in areas where our device is not directly available. It will also allow low-income communities to save costs while maintaining equivalent quality of care. In the developing world, availability of a low-cost hemodialysis machine will allow ministries of health and hospitals to forgo rationing of devices and provide them to hospitals and clinics, multiplying the availability dramatically. In addition, a hemodialysis machine that is universally compatible with branded disposables, meaning that low-resource communities can begin using the materials they have on hand, regardless of the brand of machine they currently have access to. In broader terms, providing communities with open-access, low-cost medical device designs fosters a culture of self-reliance and sustainability. If low-resource communities can access the equipment they need via an open-access model, they feel more encouraged to troubleshoot problems, customize designs to meet their needs and share their findings with others. The ability to share successes in an open-access market allows medical and technical communities to work together, avoiding duplication of work and long feedback cycles. Evaluation In evaluating the success of the project, the team will review the following questions at the end of the 12 month period: Are the materials used easy to source? Is the machine equipped to accept the most common types of disposables? Does the device function as well as comparable premium brand devices on the market today? Is the device user-friendly? Is

the device ready for clinical and validation trials? Can another person or group use our work to replicate their own. In other words: is our work open enough in a practical way?
Decision **Not shortlisted**
Comment on decision from Wellcome This proposal aimed to develop a piece of open source hardware. There were concerns raised about the feasibility of this project, and it was not clear which parts of the proposal advanced open research.

Title

Creating and sharing reproducible research code the workflowr way

Lead Applicant

Dr John Blischak

Details of proposal

We work in an area where computational analyses of large genomic data sets have become an integral aspect of all scientific output. Our aim is to help biomedical researchers develop computational analyses that are more reproducible—for example, to make it easier for new graduate students to organize the code and data files of their first project; for postdoctoral scholars to reproduce their figures when revising a manuscript; and for principal investigators to review the methods in a manuscript. A realization driving this effort is that a few simple practices using widely available software tools can greatly facilitate the development of more reproducible computational analyses. However, the cognitive burden imposed on researchers to combine these simple practices into their daily programming routines can greatly impede their progress. Motivated by this, we developed a software tool called workflowr. It combines widely available tools into a common, easy-to-use interface, thereby lowering the barrier to development of reproducible scientific code resources. We have designed workflowr to be used within the R interactive programming environment—workflowr is short for workflow in R because it facilitates a better code development workflow in R. R is the predominant programming language in biomedical research, as well as many other scientific research areas, and it is one of the most important programming tools for data analysis in science and industry. The workflowr package is built on mature open source technologies: (1) the R programming language for statistical computing and graphics, (2) the R Markdown literate programming framework for generating reports directly from code, and (3) the Git version control software for tracking code development. In workflowr, researchers develop their analysis by writing code in R Markdown files, then run workflowr functions to reproducibly generate the webpages containing the source code and results to be shared (see Figure 2 in Additional Information). The workflowr commands time-stamp both the source code and webpages without the researcher needing to know how to use Git, which can be daunting for less experienced programmers. Additionally, workflowr embeds links to past versions of the research website so that any researcher can explore the development history of the project without having to download additional files or software. The final product, generated with simple commands in R, is a website containing time-stamped, versioned, and documented results, which researchers can host online using free services (e.g., GitHub Pages). The workflowr package is already being actively used by dozens of our trainees and collaborators in biomedical disciplines at the University of Chicago. It has also been discovered by a wide range of enthusiastic computational research scientists and data scientists. We released workflowr on the Comprehensive R Archive Network (CRAN) in April 2018, and it was recognized as one of the top 40 new packages released to CRAN that month by RStudio R Views. Our aim in developing this Open Research Fund proposal is to expand the influence of workflowr beyond our immediate research network and the most enthusiastic data scientists to the larger biomedical research community, particularly to researchers that do not necessarily have strong computational backgrounds, yet need to develop code as part of their work. This requires a dedicated effort in outreach, education, user support, and continued development of the workflowr package to address issues and allow for new and unanticipated uses as the user-base grows. To achieve our aims, a summary of our proposed efforts are as follows: 1. Organize at least two workshops to teach the workflowr framework to biomedical researchers. 2. Improve the documentation and develop tutorial materials, including a series of online video tutorials on reproducible research with workflowr. 3. Create a centralized website for curating workflowr websites and highlighting success stories in which workflowr-supported results were published in scientific articles. We will monitor and evaluate our proposal using multiple quantitative metrics. For example, we will measure: (1) the number of workflowr projects hosted on public sites such as

GitHub and GitLab; (2) the number of downloads of the workflowr R package from CRAN; (3) the number of unique visitors to our newly created workflowr curation website, and (4) the number of published research articles that include a URL to a workflowr website (success stories). Please see the Additional information section for examples. By next year, our aim is to have over 100 active workflowr projects (which we define as having been updated in the previous 3 months) shared on our new centralized website, and 15 published articles that include a workflowr website URL. In summary, with funding from the Open Research Fund, we will encourage biomedical researchers to create reproducible analysis code by providing an easy-to-use software accompanied by extensive documentation, and by fostering a vibrant online community for sharing reproducible research results.

Decision

Not shortlisted

Comment on decision from Wellcome

The proposal concerned the Workflowr platform, which was a novel and potentially very valuable resource. However, the level of innovation proposed, as well as the potential impact of this proposal to transform health research through openness were limited