Title

LABIRE - Latin American BIoinformatics REsource

Lead Applicant

Dr Adrian Turjanski

Details of proposal – team members and collaborators

Adrian Turjanski: Buenos Aires University. Argentinian Genomics data System (SNDG): Group Coordinator. Previous experience with similar systems

César Rodríguez Sánchez: University of Costa Rica: Microbiology Expert. Key User

Edgar Carvalho : Universidad Fedetal da Bahia: Negleted Diseases Exptert

Guy Cochrane: EBI Team Leader – Compare Project initiative. Bioinformatics infrastructure Expert

Key Collaborator:

Cath Brooksbank: EBI – CABANA Project coodinator. Training and course management Expert

Details of proposal – vision, aims and influence on open research

Our vision is to provide a bioinformatics resource that allows storage and analysis of locally isolated pathogen genomes to fight communicable diseases. We want to create a Latin American hub for collaboration between health institutions to foster translational bioinformatics.

We aim to create a Latin America database for communicable disease analysis and surveillance. This resource will provide standard based, ready to use, bioinformatics pipelines to analyze pathogen genomes. Such a system will benefit those tackling individual problems at the frontline, clinicians, as well as policy-makers. It will provide a web resource to identify and analyze outbreaks, resistant strains, compare genomes, allow to add metadata, and share the information to international resources. We will catalog Latin America data, tools, researchers and groups to allow collaborations between health institutions and bioinformaticians.

Latin America needs to find solutions to fight communicable diseases and pathogens such as chagas, leishmaniasis, malaria, tuberculosis and dengue. Understanding the genomic makeup of local pathogen strains and sequence variations can have a major impact on clinical research, surveillance and diagnostic. However, international databases lack Latin America data both because is not well annotated and health institutions do not have bioinformaticians and personnel trained to analyze it. On the other hand, there is a new generation of bioinformaticians working in latinamerican research institutions that are developing tools to analyze genomic data. Finally, local governments lack a resource that allows them to know what data is generated and how resources need to be allocated to foster research in communicable diseases.

The idea of connecting local bioinformaticians with health researchers is to have a sustainable development that can not be achieved by international collaboration alone. By developing a user-friendly web resource that allows sharing and processing locally obtained data through well-established pipelines will enhance distribution of not previously available local datasets.

For the first, we will reuse the current National Genomic System currently function in Argentina, so can be extended to the rest of the region. In this way, much of the effort has to be done in the consolidation of the Latam network rather than a development from the scratch initiative. By making local datasets open and easy explorable, our project will encourage open research practices in Latinoamerica.

The following activities will take place in the scope of the first year:

1) Software development: Currently SNDG and TargetPathogen supports genome, gene and structure searching and visualization. The engine behind those platforms will be extended to support user/group/country management, pan genome analysis (using panX), outbreak surveillance, genotypic resistance profile and comparative genomics (with GMOD or equivalent tools). Also standard pipelines must be ready to use to motivate local datasets upload and analysis. For example, process a group of samples to get sequence variability and resistance profiles. This would be carried out by a workflow management system such as Galaxy. As pilot project we will upload and process 200 Clostridium genomes from Costa Rica with the developed platform.

2) Meeting: An expert meeting formed by the team members, key collaborators and researchers of other recruited institutions would be held within the first three months of the project to define relevant pipeline guidelines and good practices for the platform use. In this meeting we will also discuss and establish how to disseminate the initiative.
3) Workshop: A training workshop would be held in November 2020, with attendees from the team, key collaborators and other recruited institutions. There we plan to debut the first release of our planned software framework to the community so that we can solicit feedback and encourage contributions.

Details of proposal – evaluation plan

The project progress will be evaluated by the success of the following goals according to schedule. Implementation of the platform with all the proposed features. The success of this goal will be tested by a pilot project performed by the available local Clostridium genomes provided by César Rodríguez Sánchez.
(A) Adoption of the platform by the seed group to upload and analyse their own local datasets. This goal will be achieved if the institutions represented by the seed group actively upload and share datasets with the platform environment. (B) Recruitment of a seed group of at least 10 researchers of diverse Latinoamercican countries. This goal will be tested by the success of a first meeting with our team, key collaborators and new recruited researchers for the initiative. Platform adoption beyond the seed group. The two key success indicators of this goal will be the adoption of the developed platform by research and health care organizations of Latinoamerica to upload and analyse local datasets and a solid community engagement to broaden its scope beyond the lifetime of the grant.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a potentially impactful proposal with a good evaluation plan. However the level of innovation proposed was limited.

Title **Using Natural Language Processing on all existing Open Access scientific publications to form the framework of the Octopus publishing platform**
Lead Applicant **Dr Alexandra Freeman**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title **Piloting the use of transformative agreements in the NHS to accelerate the transition to immediate Open Access**
Lead Applicant **Dr Alicia Wise**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

Increasing access to open source spatial demographic data using worldpopR and QGIS plug-ins

Lead Applicant

Dr Andrew Tatem

Details of proposal – team members and collaborators

Andrew Tatem (WorldPop Director) supervise the project development and delivery, integrate outputs with GRID3 capacity strengthening efforts

Natalia Tejedor Garavito (WorldPop Geospatial Data Technical Lead) Coordinate workshops and assist project delivery

Maksym Bondarenko (WorldPop Spatial Data Infrastructure Lead) Deliver the R/Python packages and QGIS plug-in

Alessandra Carioli (WorldPop R and Demographer) Provide expertise on the building of the R package and data visualization from a demographic perspective

Pulane Tlebere (UNFPA) Implementation partner to convene health ministry representatives to take part in workshops to evaluate the outputs

Details of proposal – vision, aims and influence on open research

Vision: Enabling non-technical experts to use WorldPop spatial demographic data and integrate it into workflows to support health research and decision-making.

Background: WorldPop (www.worldpop.org) provides a wide array of open access spatial demographic data, which are widely used by scientists and decision makers across the globe, particularly in low- and middle-income settings. These include geospatial datasets with estimates of population distributions, demographics and dynamics at 1x1km grid squares or finer, with the mapping of women of childbearing age, pregnancies and live births recently funded by the Wellcome Trust. These geospatial datasets have proven valuable to researchers and implementers in health and social research (>10,000 citations, >500,000 downloads), particularly in the field of health metrics, spatial epidemiology and in tracking progress towards development goals, where the use subnational data is increasingly emphasised.

However, often the software required to analyse and visualise such geospatial data, such as ArcGIS, are often not open to researchers and practitioners, requiring expensive licenses and significantly limiting the range of data applications, particularly for low- and middle-income countries with insufficient computational resources. This makes it difficult for researchers and health practitioners to use the data, in addition to a general lack of expertise in many cases in ArcGIS and/or coding. In this context, R-Cran, Python and QGIS are free software and currently widely used among health data scientists, for their flexibility in data management, efficiency, reproducibility of outputs and high quality of data visualization. Therefore, our aim is to make WorldPop spatial demographic data easy and ready to use for a much greater range of health scientists and practitioners than at present, through a wider range of data formats and being more accessible via QGIS plug-ins, and R and Python packages. Being able to easily acquire open access demographic data with relevance to policy planning, health metrics and epidemiology, would help researchers as well as practitioners to access and share knowledge, positively impacting scientific outputs and beyond. In addition, our proposed tools will allow the analysis and visualization of any linked datasets in a convenient way without prior knowledge of the software or specific training in spatial analysis.

Methods:

1. We will provide a vignette on how to format and standardize input data that can be easily acquired by a QGIS plug-in to make WorldPop data easily accessible. This would make it possible for any user to employ WorldPop data, integrate them with their own datasets as well as to access repositories from different providers (e.g. World Bank, Humanitarian Data Exchange, International Public Use Microdata Series, Demographic and Health Surveys).
2. We will create an R and Python package (worldpopR) that can easily access existing datasets. Users will be able to access a query builder to analyse and visualize the datasets in different

formats, relevant to their objectives. The analysis tools will include descriptive and zonal statistics and, geospatial analysis that will help explore datasets. Moreover, the packages will provide functions for plotting and mapping geospatial demographic at the national or subnational level. For instance, this will help identify the total number of population/births/pregnancies/unvaccinated children at the district level and/or producing heat maps to classify areas of high and low density for multiple years and, multiple countries at once. These packages will be stored in open repositories (such as GitHub and on the WorldPop website) for easy access and updatability. The packages will be demonstrated in a shiny GUI app that will be designed in collaboration with end-users, and shared broadly, together with the code. Everything will be published on the WorldPop website and in GitHub to be used by other researchers, and to allow their integration into other platforms.

3. Additionally, a QGIS plug-in will provide a menu of mapping tools, where users can link datasets from online repositories and any other geospatial datasets. Users will then be able to visualize datasets, apply multiple classification techniques and have the option to export them into a readable format to perform further downstream analysis.

4. We will pilot and evaluate our R/python packages and QGIS plug-in by carrying out at least one dedicated workshop in a low/middle income country, where we have contacts through our UNFPA partners with health ministers/practitioners. These partners are working to address targets for universal health coverage including sexual and reproductive health and rights, as well as improving the health of women, children and adolescents. These newly created tools will substantially simplify and improve efforts to track programs and develop health metrics at subnational scales. Moreover, the testing of the tools in multiple other workshops will be feasible through the GRID3 program (grid3.org), where geospatial data support and capacity training is given to low-income nation governments and universities. We will also participate in at least two international conferences (e.g. ASTMH), where we will put together a symposium to present the applications as well as the outputs and receive feedback from scientists.

Details of proposal – evaluation plan

We will be able to track the number of people downloading our datasets and applications through GitHub and the WorldPop website, where we currently have over 5000 dataset downloads per day. We will create a section within worldpop.org to provide links, training materials and information regarding the packages and plug-in. There we will include a feedback form to collect suggestions anonymously, and also conduct dedicated user surveys actively.

Additionally, we will also get feedback from the workshop and symposium participants to improve the package and incorporate any further improvements from both academic and policy perspectives. Furthermore, dissemination of the products will be alongside geospatial capacity strengthening activities carried out within WorldPop in countries where we collaborate for existing projects such as GRiD3 (grid3.org). This will enable us to obtain further feedback from National Statistics Offices, other ministries, scientists and UN country offices in low- and middle-income countries.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

afrimapr : facilitating the use of spatial data in African public health operations and policy with reusable R software building blocks.

Lead Applicant

Dr Andy South

Details of proposal – team members and collaborators

Andy South, Liverpool School of Tropical Medicine. Coordinator, coding, training. Andy works on geospatial training for operational vector control staff and develops widely used R mapping packages.

Paula Moraga, Lancaster University. Coding, documentation, book writing. Paula develops spatial methods for disease surveillance including an R spatial epidemiology package, and is writing a book on geospatial health data in R.

Anelda van der Walt, Talarify, South Africa. Engagement with African data communities. Between 2014 - 2019 Anelda led the African Carpentries initiative through which thousands of learners from 13 African countries were introduced to contemporary open science and data-science skills.

Julie-Anne Tangena, LSTM. Needs assessment for operational staff, trialling resources in Malawi. Julie-Anne is a public health entomologist with experience in both research and mosquito control operations. She will be working in Malawi to improve the use of evidence in mosquito control.

Robin Lovelace, University of Leeds. Coding, book writing. Robin develops R spatial packages and is lead author on the 2018 book 'Geocomputation with R', the most comprehensive resource for geographic data methods in R.

Margareth Gfrerer, Education Strategy Center, Ethiopia. Planning training in Ethiopia.
Margareth has grown the Data Carpentry community in Ethiopia working with the Ministry of Education.

Details of proposal – vision, aims and influence on open research

Vision : Researchers and operational staff able easily to make useful maps and applications for Africa from spatial scientific data using open-source software.

A wealth of health relevant spatial data are available for Africa including disease model outputs, human population estimates and vector abundances. Unfortunately, in-country health programs and researchers rarely benefit from them. This project will develop software building-blocks to facilitate the use of such spatial data, including the creation of web applications. We will use R - a top data-science language, ubiquitous within research and becoming more popular for operational programs. There is a growing data-science community in Africa with high potential to develop software tools to address local issues. The aim of this project is to support this group (and others) by developing these easily useable software building-blocks, which can be used to create tools relevant to local circumstances. We recognise that researchers are likely to be more able to use these resources in the first instance and can provide a route to use in operations and policy.

Activities :

1. Develop R package(s) in the open on Github based on existing team experience.
2. Establish collaborations with online communities and invite feedback.
3. Submit package(s) to rOpenSci and Journal of Open Source Software to receive constructive peer review. Submit to the Comprehensive R Archive Network to make accessible to R users worldwide.
4. Develop training and train-the-trainer resources.
5. Trial software and training resources in UK, Malawi and Ethiopia.
6. Write an open book targeted at entry-level R users outlining how these resources can be used to address local use-cases.

At the heart of the technical solution will be a modular R package with re-usable components targeted at entry-level users. Designing for ease-of-use will be a priority. The key functionality will be to improve access and use of African administrative boundary data, so that scientific datasets can be converted easily into maps, tables and applications that are useful for local decision-

making. For example, code like “afrimapr(country=’mali’, adminlevel=2, detail=’simple’)” would plot and return a simplified map that could be displayed in a web application. A template ‘shiny’ web application will be included that can load any spatial surface and calculate summary statistics for pre-loaded or other administrative boundaries. Documentation will include instructions on how to copy and modify this template according to user needs. Early in the design process we will consider other functionality such as access to openstreetmap, health facilities, population estimates and satellite imagery. We will keep in mind the tension between usability and flexibility, and whether additional functionality should be located in other packages. The package components would make it relatively straightforward to create an application like the runner-up created by the lead applicant for the Wellcome 2019 Malaria data re-use prize.

Our software and resources will catalyse the use of spatial research data in Africa. The software will also provide data producers more efficient and standardized means to increase the reach of their data. To achieve a large and positive impact we see two main challenges that this team is well placed to address. Firstly developing useful and usable software resources and secondly achieving a high adoption of those resources. Whilst high adoption is helped by a good product it is not guaranteed. Firstly, we have a depth of experience in creating software that is well used. Andy created rworldmap, an R package for mapping country data, that has 282k downloads,161 citations and continues to be used (including by the winning entry in the 2019 Wellcome antimicrobial resistance data re-use prize). More recently Robin created ‘stplanr’ for transport research (39k downloads) and Paula ‘SpatialEpiApp’ for disease risk detection (6k downloads). Secondly, the team has a depth of experience and contacts necessary to promote adoption within Africa and beyond. Anelda helped coordinate 120 Data Carpentry workshops for African researchers including 10 train-the-trainer events yielding 50 qualified instructors and many more in the pipeline. Since 2017 Margareth has initiated training, hackdays and competitions reaching 500 researchers in Ethiopia. Julie-Anne has extensive public health experience in Africa and Asia, leading multi-disciplinary collaborations. She will be well placed in Malawi to investigate the challenges of software-use by public health workers and researchers.

Details of proposal – evaluation plan

We will consider our project successful if the software components are adopted and used across Africa and elsewhere in operations and research. Adoption takes time, we expect clear indications by year-end but true success will only be apparent when these components or their successors are in wide use in five years and beyond. For monitoring and evaluation, the project can be split into three foci :

A. Code

Development will be in the open on Github allowing others to contribute. We will monitor community feedback.

Early versions will be tested on researchers within LSTM through a user questionnaire.

Submission for collaborative open peer review at rOpenSci (an initiative fostering an ecosystem of open-source tools for open science - the team has experience of submitting and reviewing code for rOpenSci).

Submission to CRAN (Comprehensive R Archive Network).

Submission to JOSS (Journal of Open Source Software) - an open access journal for research software allowing for the software to be cited and for us to track it.

B. Training resources

Training resources will be trialled in LSTM on research staff and students.

Informal training trials will be conducted in Malawi and Ethiopia.

Formal train-the-trainer sessions will be conducted in UK, Malawi and Ethiopia

C. Adoption

Good choice of an R package name will allow internet searches to track code use. (e.g. rworldmap gave no hits on google when Andy created the package 10 years ago, now it gives 60,000). afrimapr currently gives 20 hits and could be a good package name.

CRAN downloads will be tracked. JOSS citations (although may not appear within the timescale of the project). Finally the project will be evaluated at the end for likely sustainability by how much it has attracted contributors and further funding.
Decision Funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

PROM: Platform for Reusable Open Models for Predicting Antimicrobial Resistance

Lead Applicant

Dr Anshu Bhardwaj

Details of proposal – team members and collaborators

Roberto Toro is researcher at the Institut Pasteur in Paris, and research fellow at CRI, with expertise in neuroinformatics, bioinformatics and the development of platforms for scientific collaboration.Marc Santolini, Research fellow at CRI, Paris, visiting researcher at the Barabasi Lab (Network Science Institute at Northeastern University, Boston) and research collaborator at Harvard Medical School, has expertise in AI, open science and network biology, will provide expertise on customization of the Galaxy platform with respect to metadata definitions and data integration.Jonathan Grizou, Research Fellow, CRI, Paris. Research Affiliate, University of Glasgow. He has expertise in combining AI algorithms and roboticsinto the physical and life sciences. Previously co-founded a robotics start-up. Hewill provide expertise on integrating machine learning models on our platform.The project PI is a member of the Galaxy Community. Started in 2005, the Galaxy Project (https://galaxyproject.org/) is a web-based scientific analysis platform used by scientists across the world to analyze large biomedical datasets with focus on making analyses accessible, reproducible, and simple so that they can be reused and extended (PMID: 29790989). Members of the this open source community will be engaged from time to time in developing the state-of-the-art platform.

Details of proposal – vision, aims and influence on open research

(i) Aims: WHO released a list of priority pathogens in 2017 against which new drugs are sorely needed. Given the increasing scourge of antimicrobial resistance (AMR) it is imperative that antibiotics are prescribed cautiously based on the drug sensitivity profile (DST) of pathogens. Culture-based assays are currently the gold standard for drug susceptibility testing. However, recent studies have evaluated the combined approach of whole genome sequencing (WGS) followed by analytical tools as a preferred method to reduce time and cost of molecular diagnosis (PMID: 30886350; PMID: 29653190). A large number of these methods have been published and used to infer quantitative relationships between genomic variation and drug resistance phenotypes (PMID: 31047860, PMID: 31182025, PMID: 30550564,PMID: 31106066, PMID: 30689732, PMID: 30333483, PMID: 31174603, PMID: 30514867). Nearly 30 prediction methods utilising WGS for predicting drug sensitivity profile are published every year. However, it is difficult to share these models on a common platform as there are no standards for depositing them. These models are currently available from code repositories, as services or as supplementary materials of publications. Lack of a standardised framework for sharing AMR models hampers their optimal use, limits their application on new datasets and makes it difficult to reuse them for composite models. Moreover, it requires expertise in data handling and machine learning to use these models. Efforts have been made to benchmark existing algorithms for predicting DST and have lead to observations of non-trivial errors in reporting the phenotypes (PMID: 30736750). It is also important to mention that these models are not available for all the priority pathogens and for the ones they are available, the models need revision with higher resolution of drug sensitivity data obtained from discovery of new drug resistant determinants. Given the pace at which genome sequencing is performed for various clinical isolates, it is imperative that a unified framework is designed to deposit and execute AMR models and datasets in a workflow system that makes it easier for the research community to reuse/reproduce the same. Therefore, the current proposal, aims at establishing an open workflow system that:

I) Is easy to use,not limited to machine learning / data handling experts

II) Overcomes challenges in customisation of input data formats and pre-processing both for reference and pan-genome datasets

III) Allows for hosting and community curation (crowdsourcing) of data on drug resistant determinants of priority pathogens based on FAIR principles
IV) Has provision of creating computational workflows and sharing them
V) Defines standards for building new models for interoperability and ease of reproducibility of findings with clearly defined metadata structure
VI) Offers a plug-n-play flexible environmentfor the end users to develop workflow systems for predicting AMR.
Galaxy Workflow system is ideal for implementing this project as it allows for customization of all the above-mentioned features. Target audiences: Researchers working in the field of AMR canuse existing models on their datasets, contribute new models and datasets to PROM. Clinicians with sequence data can perform drug resistant profiling without getting into the hassles of installing and doing data pre-processing, which in most cases need expert/trained human resources.
Activities: PROM is expected to function as a centralised AMR models repository with existing machine learning models and input datasets (genome sequences, drug sensitivity and other annotation datasets)available as an open source platform for priority pathogens to begin with. Examples packageswill be provided for available models to test the system. New machine learning models will be built using the transfer learning methods. A network of research and clinician community will be engaged in testing the platform and contributing datasets/models to PROM.
(ii) The platform is expected to introduce good practices of data and model sharing as the models may be tested easily and the datasets are available in a ready to use format. This also provides an easy way for building reproducible workflows,for comparing different models on same input datasetsand for selecting best models in anygiven context. Given the global challenge of AMR, it will also provide insights into how strain variability can lead to prediction of novel drug resistant determinants in different geographical settings. PROM will also help in avoiding reinventing the wheel for several steps involved in process.

Details of proposal – evaluation plan

1. The first phase will deal with setting up platform for the critical pathogens (Acinetobacter baumannii; Pseudomonas aeruginosa; Enterobacteriaceae). This task will be divided into two steps:I) crowdsourcing dataof these pathogens with respect to their drug sensitivityprofile from literature as per the European Committee on Antimicrobial Susceptibility Testing -EUCAST standards. The PI has implemented several crowdsourcing projects and will assemble a team of 6-8 members for curating DST data on these pathogens.II) Simultaneously integrationof existing machine learning packages will be done alongwith the input datasets as shared libraries.2. Once the Galaxy platform is customised, benchmarking will be done with available models. This will ensure that the research community gets familiar with the platform and how it functions so that they can participate inusing /contributing to the same.3. Introduce predictive models in cases where the datasets are not sufficient using transfer learning approaches.4. Outreachactivities will be performed with clinicians and research communities to provide training and hands-on workshops(atconferenceslikeEpidemics/ICPIC). This will help in further customising the platform for the benefit of the research / clinical community and will also identify potential contributors. This step will also define the minimum metadata components required to reproduce the existing methods.5. APIs willbe used to track the number of users, datasets that are contributed and also workflows thatare generated on the platform to asses the usage/impactof PROM.6. After finishing onecycle, several crowdsourcing activities will be launched simultaneously for generating data on the other priority pathogens.7. The outcome will be anopen web-based platform with data on at least critical pathogens,their drug sensitivity profile, machine learning models and at least 3-5 workflows for predicting the drug resistance determinants with existing methods.

Decision

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

Title

Online Mentoring Consortium on writing for publication in Low &Middle Income Countries(LMIC)

Lead Applicant

Dr Bibha Simkhada

Details of proposal – team members and collaborators

Bournemouth University Team:

Dr Bibha Simkhada-Project lead has experience of running research and academic writing training to Nursing academics in Nepal. She will develop resourcesfor online consortium, workingwith the Learning Technologist to developVirtual Learning Environment (VLE). She will also help coordinatetheexperienced international researchers and editors to run online journal club.

Prof Edwin van Teijlingen-Sociologist has a long experience in writingand academicpublishing sessions internationally.He also has long track record aseditorofinternational journals. He will supervisetheoverallproject.

Mr John Moran-Learning Technologist to develop the online consortium.John willbe involved in day –to-day maintenance of the VLE.

LMICTeam:

DrAliya Naheed, icddr,b,Bangladesh-She is a recipientof D71 NIH training grant and leads a programmein Bangladesh for building research and publishing skills among junior researchers. Shewill be Local Publication Championto promote theonline platformin Bangladesh.

Prof Bhimsen Devkota, Tribhuvan University, Nepal-will be Local Publication Championto promotethe online platforminNepal.

Dr Quazi Syed Zahiruddin, Datta Meghe Institute of Medical Sciences, India-will be Local Publication Championto promotethe online platformin India.

Details of proposal – vision, aims and influence on open research

Vision: Promoting a positive research publication culture to disseminate research

Aim is to develop a resource to aidpublishingfor health professionals and academics working in health and related research in LMIC.

Background:Capacity development in health research and publication is high priority in many LMIC.There is a slowly growing research capacity in LMIC but still is a short fallin publishingby local researchers. Furthermore, there is highdemandfor skills trainingin writing for and publishinginpeer reviewed journals. Publishing health research is important to generate evidence at local, national and international levelsin LMIC. Therefore this project plansto develop an Online Mentoring Consortiumto offer health professionals and academicsopportunities to engage inresearch writing through an international support network. It is designed to engage and empower researchers producingresearch publicationsin health and related research and practice. Evidence shows that health workersoften receive little training and support to conduct research and produce publications(LeBaron et al.2015). This project will also benefit societyespecially in LMIC through spreading researchfindingsand good practice.

Deliverable:(1) an online platform, to interact with researchers in LMIC, based on free version ofVLEBlackboard(CourseSites);(2)online journal clubs every two monthsin each country. Generally the journal clubs are educational meetings where individuals meet regularly to critically evaluate recent scientific literature (Esisi 2007). In addition, it isa means of keeping up to date with the literature; promoting evidence-based healthcare; and teaching critical appraisal skills. This online mentoringapproachisunique and easily accessible to all researchers. It providesopportunity to meet experienced researchers to enhance academic writing. It will also offer novice editors toworkwithexperienced editorsto improve academicpublishing in LMIC. Both approaches will help todevelop an open research writing communitythrough this platform.

A good VLE allows users to co-construct their learning environments together over time. Itcan also act as aplace to facilitate a researchwriting training. This project is planned for open access purpose to use CourseSites which is operated by Blackboard. CourseSites is a free open source

VLE while additional products can be purchased, the basic VLE is free. Access to the "Online consortium" can be determined by the site administrator and permissions adjusted to allow external and internal participants varying degrees of access to the course. The structure of the course can be as simple or complicated as required.Access can be granted via email invites or providing courselinks forregistration. This allows for controlled access to the research project to be able to monitor who comes in and out. Creation of content within CourseSites is relatively straight forward and offers itself to open sourced tools which can be great for bringing in interactive elements that are not considered standard within a VLE. Blog posts, discussion forums, journal articles, downloadable documents and course customisation are also a basic part of CourseSites and fulfil the requirements of this project but alsowithenough sustainability to lend itto going forward.

References

LeBaronV, Iribarren S, Perri S and Beck SL (2015) A practical fieldguide to conducting nursing research in low and middle-income countries. Nurs Outlook; 63(4):462-473.

EsisiM (2007) Journal clubs. BMJ.335:s138. In https://www.bmj.com/content/335/7623/s138.

Details of proposal – evaluation plan

Target:We expect to engage at least 50 participants (health professionalsand academics)from each LMIC country. We will run four Journal Clubsin 12 months.We aim to haveat least 150 health professionals/academics benefitingfrom this pilot project.

Only registered participantswill getaccess to onlineconsortiumresources. The registration process will require participants'email and organisation detailfor monitoring purpose. We will set up agreement forms to not to breach personal data and to also coincide with the pre-existing policy on Coursites and also General Data Protection Regulation(GDPR) will be followed.

The effectiveness of the CourseSites(VLE) will be tested for usability and usefulness of the online resourcesand platformin Nepal, Bangladesh and India. The available data of the VLEwill bemonitored for the access andtime spentwithinthe platform. Engagement of the participants will also be recordedfor the purpose of evaluation.The success of the purposed activities such asuse of journal club will be measured throughthe feedback survey after the each journal club.The impact of the online mentoring consortium will be measured through auser-testing survey andtwo online focus group discussionswith online platform participants and 3-5 Skypeinterviewswith journal clubsparticipants. We will obtain written consentviaemail and ethical approval fromBournemouth UniversityEthics Committeeand appropriateEthics Committees in country as required

Decision

Not shortlisted

Comment on decision from Wellcome

This was a potentially impactful proposal aiming to build capacity in writing research publications in LMICs. However the extent to which openness would be supported or improved was unclear.

Title

Making the Addicts Index Database FAIR

Lead Applicant

Dr Caroline Copeland

Details of proposal – team members and collaborators

Christine Goodair, Substance Misuse Programmes Manager, SGUL

Christine is responsible for the Addicts Index database and accompanying microfilmed pdfs of all the records.

Michelle Harricharan, Research Data Support Manager, SGUL

Michelle will provide advice and expertise to ensure best practice is applied to make the Index data findable, accessible, interoperable and reusable (FAIR).

Kirsten Hylan, Records Manger, SGUL

Kirsten will ensure the Index meets SGUL institutional governance standards for records and preservation.

Carly Manson, Archivist, SGUL

Carly will develop a digital preservation/sustainability plan for the recovered database and ensure that the resource is preserved in accordance with archival standards.

John Corkery – School of Medical and Life Sciences, University of Hertfordshire

John will hold a consultancy role to bring expertise relating to the history of drug addiction and knowledge of the Index.

Details of proposal – vision, aims and influence on open research

The Addicts Index was created by the Home Office and comprises paper records regarding individuals seeking treatment for drug dependence, including their personal details (e.g. name, gender, occupation, drug problems, drugs used in treatment etc.), details of those providing their treatment (e.g. Drug Dependence Units, GPs, prison and police doctors, A&E units etc.) as well as information on prescribers and inspections of chemists and pharmaceutical companies. There are around 120,000 paper files ranging from the 1940s up until 1994 when the database was closed, which come under the provisions of the relevant Public Records Acts. Official custody of these paper files was transferred by the Head of the Home Office Drugs Inspectorate with the prior consent of the Home Office Records Branch and Public Record Office (now known as the National Archives) to the then Centre for Addiction Studies (now known as the International Centre for Drug Policy or ICDP) at SGUL in 1997/8. It was agreed to preserve these documents in accordance with archival standards.

The contents are 'medical in confidence' as they can contain very sensitive information regarding drug dependence, treatment regimen, criminal records (where applicable), etc. Therefore, applications for access from bona fide researchers are considered in accordance with an existing protocol agreed by the Home Office and the ICDP. An Excel index of these documents and an Access database was created, where the 120,000 records were converted into the more accessible format of pdfs so that they can be interrogated more easily, for example to search for specific notifications, addicts or notifiers, or to extract numerical or statistical information. An interface was developed to link the digitised records, via the Excel index. However, we are currently unable to access the database on its external hard drive, thus making the Index data inaccessible to researchers.

Proposal: To collaborate with research software engineers to produce a reliable code for software sustainability and research reproducibility to recover the Index and to develop and implement a sustainability plan to ensure continued access to the asset in the future.

Aim: To make the Addicts Index FAIR for researchers to interrogate.

Method: Modernisation and re-architecture of this legacy database will make the resource FAIR. Once complete, the Index will be thoroughly documented to meet the Open Archival Information System (OAIS) standard and then made available on a controlled access basis on the SGUL Research Data Repository. New tools will be trialled to facilitate access to the resource, including

using the SGUL Data Safe Haven to enable access to and analysis of the data. A digital preservation plan will be implemented so the resource does not become inaccessible in the future. A student project will evaluate the impact, risks and benefits of our work when the project is nearing completion.
Target audiences: For addiction specialists, medical historians and sociologists this collection provides access to a unique source of information about prescribing practices and policies related to outcomes in terms of treatment modalities and service provision. Aggregated data can be used to look at patterns of drug use of individuals or specific populations over time and to examine impacts on health and societal issues, for example, pathways to treatment. The way in which changes were made in the past and the lessons learned can help to inform implementation of changes in the future. These records are a unique resource of the formative years of the British system for monitoring drug addiction and differing approaches to prescribing.
Open Research Good Practice: This project is a cross-sector, interdisciplinary effort to recover an inaccessible historical database, and to bring that database to the highest standards in open data and digital preservation. It will trial, evaluate and share innovative approaches to making multi-format, sensitive historical information findable, accessible, interoperable and reusable for the long term.

Details of proposal – evaluation plan

Project Monitoring
Monthly team meetings with the specialist software engineers will be held to discuss project progression, issues arising and solution pathways. A project timeline (detailed below) will be followed.
Success Evaluation
Once the Index has been recovered and reformatted, a student project will be performed to test the usability of the database. Feedback from this project will enable us to assess database performance, and address any outstanding issues prior to project completion and promotion for use to collaborators, external researchers and social historians.
Conduction of research projects using the database will cement the Index as a valuable resource, and also act to promote its use further. For example, we ourselves are interested in the impact of the 1964 Drugs (Prevention of Misuse) Act in comparison to that of the 1971 Misuse of Drugs Act with regards to amphetamine use given the recent surge in novel psychoactive cathinone synthesis and availability, and the diversion of pharmaceutical stimulants for use as ‘study drugs’. Understanding how drug policies have influenced behaviours in the past will aid in defining new drug policies in the future.
Restoring the search function of the database will also enable addict careers to be tracked beyond the timeframe covered by the Index (1940s-1994) to other programmes such as the Dead Addicts Database, and the National Programme on Substance Abuse Deaths, which are also hosted at SGUL, enhancing the interoperable aspect of the Index.
Ultimate success will be achievement in making the Index data FAIR and formatted to ensure sustainable access in the future.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to increase the accessibility and useability of a historical database on drug addiction, with clear potential for impact in this field. However, the level of innovation proposed was considered limited.

Title

Linking neuronal function to cell identity through novel whole-brain neurochemical datasets and comparative analysis tools

Lead Applicant

Dr Chintan Trivedi

Details of proposal – team members and collaborators

(i) Professor Stephen Wilson (Wellcome Trust Investigator, UCL, London, UK): Professor Wilson's lab will provide the facilities for whole-brain fluorescence immunohistochemistry, in-situ hybridization and high-throughput microscopy to generate the core novel dataset for the project. His group will also provide expertise in anatomical annotation of the data.
(ii) Dr. Isaac Bianco (Wellcome Trust Henry Dale Fellow, UCL, London, UK): Dr. Bianco's lab will provide test datasets, imaging and brain-registration expertise, and assist in the development and testing of the open-source analysis tool.
(iii) Dr. Jason Rihel (UCL, London, UK): Dr. Rihel's lab will provide plasmids for several candidate markers for the novel dataset and generate test data for validation of our analysis tool.
(iv) Dr. Harold Burgess (National Institute of Child Health and Human Development, NIH, Bethesda, USA): Dr. Burgess' lab will host our novel dataset on the publicly accessible, web-based zebrafish brain browser. Dr. Burgess will also provide guidance on development of the open-source analysis tool.

Details of proposal – vision, aims and influence on open research

Vision and Aims: The project aims to characterise comprehensively neurochemical expression in the larval fish brain and develop methods to enable comparison between whole-brain neuronal activity, neuroanatomy, neurochemistry and gene expression. Among widely used model systems, larval zebrafish offer the unique opportunity to produce whole-brain datasets at high temporal and spatial (cellular) resolution with high throughput, and reproducibility across individuals. The registration of whole-brain datasets onto zebrafish brain atlases is possible due to advances in brain-registration techniques. These atlases (http://zbbrowser.com, https://engertlab.fas.harvard.edu/Z-Brain/home/ and https://fishatlas.neuro.mpg.de/) host repositories visualizing transgene expression and neuronal morphology. However, routinely generated whole-brain datasets often encompass neuronal activity, gene/protein expression and cell-type specificity. The tremendous potential of these diverse dataset types to accelerate neuroscience through holistic, comparative analyses is largely unexploited. Our vision is to leverage statistically validated, voxel-wise analyses between these datasets to empower researchers to link neuronal function to molecular/genetic identity, a central goal of neuroscience. We have identified potential challenges that impede the achievement of this goal: (i) Lack of availability of comprehensive neurochemistry specific whole-brain datasets limiting integration of neuronal activity with molecular identity (ii) Lack of open-access tools to perform statistically-validated, comparative analysis between diverse datasets (i.e. functional, anatomical and gene/protein expression) limiting interpretation of data
We aim to address these challenges by providing the community with the following solutions: (i) Generation of whole-brain neurochemical data: We will generate a novel data repository, consisting of whole-brain labels for 75 candidate markers encoding for neurotransmitters, neuromodulators and neuropeptides. These candidates have been chosen after an extensive literature survey and are conserved among vertebrates. We have already acquired pilot data by employing whole-brain histochemistry with our collaborators. This novel dataset will be made available on Zebrafish Brain Browser (http://zbbrowser.com), hosted by our collaborator. We will analyse co-expression patterns between all 75 markers which will serve to accelerate the field's goal of identifying cellular properties based on molecular identity. (ii) Analysis of diverse datasets: To lower the entry barrier for researchers with limited computational expertise, we will develop a desktop-based, graphical interface to perform statistically-validated, voxel-wise analyses of

whole-brain datasets. The tool will facilitate integration of processed whole-brain neuronal activity with datasets representing additional features, thereby allowing the user to tie function to underlying neurochemistry, gene expression and anatomy. For example, if a whole-brain dataset identifies all neurons with activity correlated to a specific behaviour or stimulus, the tool could readily answer the following questions: (a) What proportion of these neurons are in which specific brain region? (b) Which neurotransmitter, modulator and/or peptide likely represents the molecular identity of these neurons in each brain region? (c) Are there gene expression patterns that specifically overlap with these cells? By generating output spreadsheets for brain-region specific statistics for all repository labels, we will facilitate data-driven generation of novel insights and hypotheses. Target Audience and activities: The primary target audience for our tools and dataset is the zebrafish neuroscience community, particularly labs that generate microscopy datasets. Over 50 labs attended an international workshop on Zebrafish Neural Circuits held in November 2017, more than 60 at the Zebrafish Brain conference held in December 2018, and the field continues to grow rapidly. We will present tools and data created through this project at the Zebrafish Brain conference to encourage uptake. We will generate an online tutorial for the analysis tool with a linked forum to address user issues. We will organize a workshop at UCL and a linked webinar for optimizing acquisition, standardization and analysis of whole-brain datasets. Open Research Practices: We will host the new set of comprehensive whole-brain data (novel dataset of 75 markers and existing repository with >250 brain-wide expression patterns) accessible and downloadable through http://zebrafishucl.org/ and http://zbbrowser.com. We will encourage users to submit new functional/gene expression datasets to keep expanding this online data repository. This will strengthen the adoption of open-research practices within the community. The entire code will be hosted on Bitbucket (https://bitbucket.org/) to enable ease of modification by expert users and community-driven improvement of the tool. This will also ensure automated credit assignment to the community developers for their efforts. The tool will be developed using PyViz (a set of open-source visualization and analysis packages in Python https://pyviz.org/).

Details of proposal – evaluation plan

As we progress through the project, the alpha and the beta versions of our tool as well as the novel dataset will be released to all our collaborators and their lab members. This will ensure rapid bug-fixing and offer an opportunity to receive user feedback on features to be added to the graphical user interface to enhance experience. We aim to release the fully operational version of the analysis tool and the novel dataset at the Zebrafish Brain conference in November-December 2020. Currently, researchers generating whole-brain datasets rely on custom-written scripts for each specific dataset type, thereby elevating the entry barrier for other researchers and limiting reproducibility. The workshop and the online tutorial will allow us to inform the community about couching our tool within existing open-source workflows for whole-brain registration and data processing. This will ensure that the community has access to a completely open pipeline starting with generating data, processing and registering it to existing brain atlases and culminating with comparative whole-brain analyses. Therefore, we expect broad adoption of the tool as well as our novel neurochemistry specific whole-brain dataset by the community. The success of the tool and new data resource will depend on adoption by the community of zebrafish researchers. We will monitor and evaluate success through the following measures: (i) number of tool downloads (ii) number of data download requests on our server (iii) number of downloads of the source code (iv) number of contributors on Bitbucket (v) the number of participants in the discussion group and (vi) number of citations

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Development of standalone and web hosted software to identify probable synergistic drug indications and contraindications from co-expressed gene modules for drug repositioning

Lead Applicant

Dr Chittabrata Mal

Details of proposal – team members and collaborators

1. Dr. Chittabrata Mal, Assistant Professor, Amity Institute of Biotechnology, Amity University Kolkata. Dr. Mal will lead the bioinformatics part.
2. Dr. Mohit Mazumder, Director of Business Development, Pine Biotech, New Orleans, LA. Dr Mazumder will lead the machine learning implementation and analysis part.
3. Dr. Anirban Mitra, Associate Professor, Computer Science and Engineering Department, Amity University Kolkata and Senior Member of IEEE. Dr. Mitra will lead the software development part.

Details of proposal – vision, aims and influence on open research

Vision: This proposal aims to develop an application software to identify probable synergistic drug indications and contraindications from co-expressed gene modules. The applications of this software will range from identification of repositionable drugs using machine learning to evaluate and further develop drug-disease network. This will also help analyse the publicly available RNA-Seq expression data to identify probable drug indications and contraindications. One of the applications of a web-based tool is that it can be integrated within the RNA-seq expression database in future.

Background: Drug repurposing is not a new concept at all, a practice widely accepted especially in healthcare & medicine. Moving past the era of empirically measured science to today's multi-omics technologies has opened up a new dimension to the drug repurposing methods in a dramatic way. Disease-centric approach (i.e., health disorders linked to similar dysfunctional proteins may be treated with the same drugs) is not well accepted as the drug target might be involved in other diseases as well. To overcome the problem of specificity, network biology approaches could help in designing experiments with precision. For example, in 2016, Vitali et al. implemented a network-based study to identify drug repurposing opportunities against triple negative breast cancer, a subtype of breast cancer whose biology is still poorly understood. The gene network approach could significantly improve the prediction compared to the traditional single gene approaches. The genome-wide transcriptional profiling of disease samples, gene co-expression network analysis can identify modules (i.e., sets of co-expressed genes) as candidate regulators and drivers of disease states. The network-based drug discovery aims to harness this knowledge that could help in identifying drugs capable of helping many. In the proposed study, disease gene modules will be categorised from differentially co-expressed, positively co-expressed, negatively co-expressed, and contra co-expressed gene networks. The weighted gene co-expression and eigengene modules will be used to identify probable drug targets. Experimentally verified drug target interaction will be obtained from different databases. Each co-expressed module will be ranked statistically and analysed to find indications and contraindications.

Aims: To create a user-friendly software to identify probable synergistic drug indications and contraindications from co-expressed gene modules. In order to validate and apply the developed tool the benchmark RNA-seq datasets will be used to re-analysed the data and derive new meanings. Furthermore, a dataset with liver cancer data will be analysed to identify probable synergistic drug indications and contraindications. The tool will be useful in the analysis of drug target data which is available publicly at Connectivity Map, DrugBank and STITCH databases. Machine learning will be used to differentiate between probable synergistic indications and contraindications.

Target audience: Academics, Researchers, including post graduates in relevant fields, clinical pharmacologists and applied computer scientists, the pharmaceutical industry, and medicines regulators.

Activities: 1. Develop user friendly software to analyse RNA-seq data to identify probable synergistic drug indications or contraindications. The RNA-Seq pipelines available at t-bioinfo platform of Pine Biotech will be used to analyse and validate the results. The pipelines will be used to further develop a custom tool. 2. Addition of multiple algorithm to measure the gene co-expression. 3. Develop a ML model to differentiate between synergistic indications or contraindications in collaboration with the Pine Biotech Team. 4. Hosting of the tool freely available to user to download or analyse data using cloud infrastructure of Pine Biotech 5. Publication & awareness amongst the community by presenting in workshops & conferences.

Influence open research practices: Large amount of RNA-seq data and experimentally verified Drug-Target data are available publicly. But standalone software to identify probable synergistic drug indications or contraindications from gene coexpression modules is lacking. A standardised machine learning algorithm to identify drug indications will attract the drug repurposing community.

Details of proposal – evaluation plan

A: Data collection and framing of software
B:Data collection, acquisition, normalization, gene coexpression matrix development
C:Designing and planning for development of proposed Software, gene module identification, mapping drug target to the each of the modules
D:Incorporation of Statistical analysis techniques
E: Implementation of Machine learning techniques
F:Optimization of developed software and its validation
G:Hosting the optimised and validated software into an appropriate platform (preferably an open source platform) and report preparation

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to create a piece of software relating to drug repurposing. The potential impact of this proposal on advancing open practices in health research was unclear, and concerns were raised about the feasibility of the proposed activities.

Title

OpenStrat: The open population stratification engine

Lead Applicant

Dr Daniel Lawson

Details of proposal – team members and collaborators

Daniel John Lawson, University of Bristol

Principal Investigator, a Lecturer in Data Science in the School of Statistics with expertise in how to infer population structure. He has recently focussed on its consequences on genetic studies including genome-wide association, heritability, and prediction. He has experience with efficient computation and applied data science and is a Sir Henry Wellcome Fellow until Sept 2019.

Aliya Sarmanova, University of Bristol

Research Associate in the MRC Integrative Epidemiology Unit, with expertise in epidemiology and machine learning. She will undertake the majority of the coding work, and has first hand experience conducting GWAS and examining population structure

Details of proposal – vision, aims and influence on open research

(i) Our vision is to provide the genetics community with a freely-available, open-access software which promotes standardization and reproducibility of GWAS results by reducing bias due to population stratification. In the future the data content will grow from this initial work using worldwide diversity in UK Biobank to encompass extensive resources outside of UK.

Background: Recently it has become apparent that GWAS results based on large scale meta analysis (GIANT) and the UK Biobank have been at least partially misled due to inadequate correction for confounding by population stratification. This has led to ambiguous conclusions regarding selection on height (Yengo et al. 2018), the genetic factors influencing education (Haworth et al. 2019), and casts doubt on the use of genetic prediction of traits and disease (Reisberg et al. 2017).

Meta-analysis combines evidence from multiple small datasets (Evangelou and Ioannidis 2013), in which population structure is hard to detect (Lawson et al. 2019). That article calls for meta-analyses to be corrected in a standardized way – using a large external reference dataset to define worldwide variation, against which local variation within a single study can be compared. We have recently demonstrated that correcting for stratification in this way can massively decrease bias in GWAS, and we developed a tool to perform correction for other datasets. The technology to perform this correction is straightforward but the process of standardizing and imputing to the reference is fraught with complication.

Aims:

To create a pipeline which will make the process of correcting for population stratification in GWAS straightforward and mistake-free without requiring complex and costly imputation.

For GWAS researchers to engage with open research through a writable shared repository of population data, via the deposit of new population references.

To create a framework for future developments in accuracy. This will pave the way for new data, and methods such as Chromosome Painting (Lawson and Falush 2012; Lawson et al. 2019) which can correct for cryptic structure.

Sharing of information is limited to how to combine SNP level information into a predictor for population structure; analogously to GWAS summary statistics, no individual level data is required.

Activities/objectives:

1. Data creation. We will create/calculate principal components of worldwide genetic variation derived from the UK Biobank. We have already constructed these PCs but need to create a make a predictor for each common genotyping array, eliminating the requirement for imputation.

2. Develop software/pipeline for standardizing and genotype prediction. The key features will be ease of use, extendibility, easy inclusion into existing workflows and interoperability. It must be

efficient and robust to data coding problems such as genome build and strand. We will future-proof for anticipated rich descriptions of population structure.
3. Publicise:
- Collaboration with beta-testers, including a GoDMC proposal (methylation meta-analysis) and partners holding valuable data including CPDR (South Africa).
- Create a tutorial publication to allow other researchers to download the pipeline and incorporate it privately into their own workflow;
- Create a high-profile publication demonstrating the gains that can be made;
- Distribute the code through GitHub;
- Host a workshop at the Bristol MRC-IEU to bring together GWAS users and methods developers.
The target audience will include researchers working in genetics including genetic-driven drug discovery, genetic epidemiology and methods development. The project will also target epidemiologists and public health researchers using genetic data for causal inference.
(ii) This proposal will influence open research practices through:
Increasing research efficiency, reliability and reproducibility of GWAS results by using a fully tested standardized description of population with a simple pipeline separated from the algorithmic complexity.
Increasing academic advancement and engagement across countries, diseases and research groups by providing publicly available data analytics software and facilitating scientific collaboration.
Increasing productivity and innovation in scientific research by facilitating a community of applied users and methods developers engaged in dialog and building a world-wide community of data literate researchers.
NB This is a separate and complementary proposal to that of Dr Zheng. Dr Zheng's proposal focusses on combining data from leading groups. Dr Lawson's proposal focusses on making that data usable to more researchers. Both are independently valuable, but would amplify the others' impact.

Details of proposal – evaluation plan

Active monitoring will be performed at each stage of the development and implementation process: data creation, software/pipeline development and dissemination.
Milestones:
1. Distribution of prototype to project partners. Once the software/pipeline developed and tested we are planning to share a downloadable, documented prototype for feedback. The software/pipeline then will be refined and further developed if needed.
2. Workshop occurrence. We will seek feedback on overall usefulness and relevance of the pipeline methodology and training materials; the ease of set-up, installation, and on-going use; and whether individuals achieve independence in using the software for their own data analysis.
3. Community engagement. Github has facilities for easily disseminating, changing and monitoring code use.
4. Publications. Tutorial expected on release of the prototype, with the application expected at the end of the project.

The success of the project will be measured by adoption of the developed software/pipeline in research practice.
Additional success indicators:
- Long-term indicators (beyond the lifetime of the grant) will include number of citations to software and accompanying papers.
- Attracting further funding/collaboration to extend data content and algorithm development.
We believe that significant additional research activity can follow this application, but that this work will enable immediate value to the community.

Decision

Not shortlisted
Comment on decision from Wellcome This was an interesting and clear proposal, demonstrating a strong commitment to advancing openness. However, the evaluation plan would have benefited from more detail

Title

Participatory Research Ethics Analysis (PREA) Tool Development and Piloting (PREA-DAP)

Lead Applicant

Dr Donal O'Mathuna

Details of proposal – team members and collaborators

Dónal O'Mathúna will supervise a postdoctoral researcher who will be hired for this project. He is the PI for the PREA Project which conducted a systematic review of ethical issues in health research in humanitarian contexts, qualitative research interviews with stakeholders in humanitarian research, and in-country training in research ethics. The findings informed the development of the PREA Tool to stimulate reflection and analysis on research ethics. The knowledge and expertise gained from this research will be used to oversee the researcher as he/she develops and pilots the Tool.

Professor Tine Van Bortel, University of East London and University of Cambridge, led the qualitative research for PREA. Her experience in conducting humanitarian research, and collaborators in relevant settings, will inform and advise this project.

Sudarshan Pyakurel spent 18 years as a refugee from Bhutan. He has participated in research for many years and is a community advocate for refugee mental health. He will represent the interests and concerns of vulnerable and marginalised groups of research participants in this project.

Other members of the PREA consortium (http://www.preaportal.org/team/) are available to advise this project as needed since the team continues to work together and seek further funding for related research.

Details of proposal – vision, aims and influence on open research

(i) Ethical issues in humanitarian health research are important but understudied, with few resources openly available to guide and support researchers in this developing field. As disasters and conflict lead to humanitarian crises, responders need evidence to support the effectiveness of interventions. Research and other evidence-generating activities are proliferating, often by organisations with limited experience conducting research. These projects raise ethical issues for which practitioners often feel poorly prepared. The skeleton for the Participatory Research Ethics Analysis (PREA) Tool was developed from the R2HC-funded PREA project. It is based on evidence generated from the PREA systematic review, qualitative interviews with research stakeholders, and in-country training and feedback workshops in Nepal, Afghanistan, South Sudan and Ethiopia (publications in preparation). The PREA-DAP project aims to continue development of the PREA Tool, pilot it with key stakeholders, and evaluate its impact in addressing ethical issues in humanitarian health research. Additional funding has already been secured for all IT support costs for the PREA Tool until the end of 2020.

The PREA Tool is innovative in focusing on identification of and reflection on humanitarian research ethics. Approaches to research ethics often focus on higher-level principles, regulatory compliance, and utilitarian calculations of harms and benefits. Macfarlane, in Researching with Integrity, notes that contemporary research ethics misses "how to positively encourage ethical conduct. Developing an understanding of what to do is always a more challenging prospect than issuing edicts about what is not right" (2009, 3). To encourage reflection and ethical conduct, and develop ethics decision-making skills, innovative tools should support decision-making during ethical dilemmas, especially in humanitarian contexts, and openly promote lessons learned.

The PREA project developed the R2HC Ethics Framework to encourage humanitarian researchers to identify ethical issues (https://www.elrha.org/researchdatabase/r2hc-ethics-framework-2-0/). These questions address all research phases, and were combined with PREA research findings to design the PREA Tool framework (www.preatool.com). This open resource allows stakeholders in humanitarian research, e.g. researchers, participants or advocates, to identify potential ethical issues based on key variables in the project, and points to key resources and tools, mostly available open-access.

Within PREA-DAP, a curriculum, slide presentations, and guidance notes will be developed to allow independent usage of the PREA Tool. Also, the Tool will be piloted in workshops held for this purpose. In these, participants will be asked to prepare case studies focused on ethical issues they encounter in their research. These will be developed during the workshops, and innovative case studies added to the PREA collection so they can be used in further ethics training for humanitarian researchers. O'Mathúna is currently involved in another funded project using evidence-based educational principles to develop ethics case studies and training materials, and these approaches will be incorporated into PREA-DAP. O'Mathúna works closely with colleagues at OSU with expertise in implementation science who will advise on objective and qualitative methods of evaluating the usefulness and effectiveness for ethics reflection of the PREA Tool.
At the PREA in-country training, participants expressed interest in engaging in piloting and evaluating the PREA Tool. PREA-DAP will organise workshops in these countries (named above). Travel funds are sought for some, but not all, of these, as PREA consortium members work in some of these countries and have been trained to conduct research ethics workshops. The PREA Tool will also be piloted at major conferences (e.g. Global Forum on Bioethics in Research, World Congress of Bioethics 2020).
(ii) Our team is committed to open research practices with our PREA Tool available open access, as are our conference recordings https://kb.osu.edu/handle/1811/87554), and our publications will be. Within research ethics, concerns exist about how open researchers can be about the challenges and limitations they face with ethics. Our vision for the PREA Tool is that it will promote and support open discussions about ethical issues with research in humanitarian settings. At workshops or during team meetings, the Tool will prompt consideration of ethical issues identified by others in similar contexts. Through these, and aided by training and support materials that PREA-DAP will develop, openness will be encouraged to promote lesson learning from the experiences of others working in similar settings. These lessons can then be captured in case studies which will be openly accessible on the PREA website (http://www.preaportal.org/case-studies/).

Details of proposal – evaluation plan

Monitoring of the project will include weekly meetings between O'Mathúna and the postdoctoral researcher to set targets for adding content to the Tool, develop related materials and organise workshops. Bimonthly meetings will be held with the advisors, Van Bortel and Pyakurel, with email and phone contact available as needed.
Key success indicators will be the results of the evaluations and feedback from those involved in the piloting workshops, particularly around their willingness to use the PREA Tool further within their research, ethics training and support mechanisms. Interest within participant communities will also be a key indicator of success, especially if they see the PREA Tool helping their communities to understand research ethics and engage in discussions around ethical issues in relevant research.
The research leading to the PREA Tool was funded by R2HC, and there is interest in them using it with their grantees and applications. Adoption in this way by R2HC would be an important indicator of success. The PREA consortium has links with other international organisations such as WHO, International Office for Migration, the International Association of Bioethics, Global Forum on Bioethics in Research, etc. Adoption or recommendation of the PREA Tool by these, and the PREA consortium's own organisations, would indicate success. Our aim would be to have at least two of our own team's organisations adopt the PREA Tool by the end of PREA-DAP, with presentations arranged at other team organisations and at least one international organizations. In the year after PREA-DAP ends, we aim to have at least one in-country Ministry actively involved in testing the Tool. Other indicators would be presentations at international conferences and professional organizations.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aims to enhance the utility of an important and valuable tool. However, the level of innnovation was felt to be limited and the evaluation plan could have benefits from further development

Title

An open-access repository of bio-images to improve AI-based diagnosis of infectious diseases

Lead Applicant

Dr Elmer Llanos-Cuentas

Details of proposal – team members and collaborators

Development and Deployment: Pierre G. Padilla-Huamantinco, Jose A. Zapana-García, Anthony A. Campos-Zavala-Health Innovation Lab (Institute of Tropical Medicine "Alexander von Humboldt"). Study Design and Clinical validation: Fiorela Y. Alvarez-Romero -Malaria and Leishmaniasis Research Group, and Gabriel Carrasco-Escobar -Health Innovation Lab (Institute of Tropical Medicine "Alexander von Humboldt"). Sponsorship: Richard Bowmanand Julian Stirling-Bath Open Instrumentation Group (University of Bath), and Sharada Mohanty -Digital Epidemiology Lab (École Polytechnique Fédérale de Lausanne).

Details of proposal – vision, aims and influence on open research

(i)Vision: Our vision for this project is to build awebrepositorybased on images collected by low-costdigitalmicroscopy in order to provide clinical sampleson neglected tropical diseasesfor Artificial Intelligence(AI)-baseddiagnostics. We will leverage advances in computer visionand open science hardware to set up open-source, low-cost and portable stations in healthcare facilities (laboratory units)

where health personnel can easily read, label and upload samples as part of routine activities. At the same time, this providean open image repositorywhere the scientific community can use them through the platform for the development of disease diagnostics.(a)Aims: 1. We will constructand validate the OpenFlexure Microscope(OFM) in clinical settings. The OFM is an open-source and 3D printed microscope which includes a precise mechanical stage to move the samples and focus the optics. The software for controlling the microscope runs on an affordable Raspberry Pi computerandallows users to see a live preview of the microscope camera.We will use 3D printablefiles from OFM GitHub repository and build it according to the documentation. Then, OFM will be evaluated by the development team following the Mechanical and Optics performance tests published by Sharkey et. Al (2016). After functionaltests, we will proceed to evaluate the resolution of the OFM with a positive resolution target and get some images from different specimens to be checked by a laboratory technician.2. We will design and implement an open image repositorybased on Common Objects in Context (COCO)dataset format and DICOM Standard for Pathology. COCO is a large-scale object detection, segmentation, and captioning dataset which was designed as a new kind of dataset for computer vision research. DICOM Standard defines the protocols for exchanging information (interoperability) in medical imaging and makes it possible for the Pathology domain to be a part of the whole healthcare process.According to clinical guidelines and lab technician's expertise, we will define categories and dataset format to enable a practical collection of images and interaction with different visual interfaces.3. We will design and implement a web platform for data mining based on the principles of Bioimages Informaticsand Human-Centered Design(HCD).We will create a sitemap of the web platform and design the wireframes and mock-ups. They will be evaluated by a small group from our target audience and we will redesign the prototype based on the users' feedback. After some iterations, we will proceed to develop the front-end, then the back-end will be implemented by our developing team to interact with the image database. At this stage, we will add other specific features that were identifiedduring co-design process.(b)Target audiences: 1. We will tailor our platform to direct use specifically by (1) Scientists (Global Health, Computer Science, Biomedical Engineering, etc.), (2) Health personnel (clinicians, technicians, pathologists, etc.) in limited-resource settings. 2. We will also target two specific audiences by disease area: (1) Malaria and (2) Leishmaniasis.(ii)How your proposal will influence open research practices:Data size are growing from day to day in an increasingly connected world. Nowadays, the need to understand large, complex, information enriched data sets has increased in healthcare. The ability to extract useful knowledge hidden in these large

amounts of data and to act on the knowledge is becoming increasingly important especially in early detection, diagnosis, and medical decision making. However, the research and training are conditioned to a few open access datasets. This situation limits advances in healthcare for patients and support for medical practitioners. The proposal intent of making our platform (hardware and software) available to the global research community. Users will find the documentation in a GitHub repo to set up these low-cost stations in health facilities and to collaborate to an open image dataset (crowdsourcing) by following protocols of sample preparation, digitization, and labeling. Users will be able to modify hardware component and to add other open-source microscopes (e.g. FlyPi). Besides the technology, users can also suggest protocols for new neglected tropical diseases or other kinds wheremicroscopy is a gold standard fordiagnosis. This community may become self-sustaining, building knowledge and new staff to improve health, and providing useful information (images) to help the next research generation.

Details of proposal – evaluation plan

(iii) How we will monitor and evaluate our proposal, included success factors: (Aim 1)Comparison between OFM and a conventional microscope. To evaluate OFM performance against standard microscopy, a laboratory technician will be tasked to observe specimens (parasites samples) with OFM and with a traditional microscope and to make speciesidentification. The lab technicianself-described observations through each method have to bethe same. (Aim 2)Evaluation of image dataset by experts. Images will be taken at two experimental research stations (a research lab and a public health facility) associated with The Institute of Tropical Medicine "Alexander von Humboldt" in Lima, Peru. Experimental research stations offer the possibility of taking many images in a reduced amount of time. At this stage, specialists from the research team will check and clean the dataset removing poor quality images and fixing inconsistencies in data. (Aim 3)Evaluation of User Experience. The content will be written by the research team, reflecting information sourced from the scientific literature. However, as the site is targeted to a wide community, rather the professional pathologists, great care has to be taken to write the content in a way that is easy to understand. The most content will be written in Spanish (validation stage) and English (release stage). After HCD process and getting the final version of the platform, we will recruit two groups of participants: UX experts (3 participants) and the target audience (6 participants in total –3per category). To evaluate the web platform, the participants will fill a standardized Usability Questionnaire, then they will join to focus groups to share their opinions about the UX platform.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Benefits of gastric cancer prevention and screening: an open-source microsimulation model

Lead Applicant

Dr Filip Meheus

Details of proposal – team members and collaborators

International Agency for Research on Cancer (IARC), World Health Organization
Early Detection and Prevention Section, Lyon, France - Filip Meheus, Dr., Jin Young Park, Dr., Viktoria Knaze
Erasmus MC, Erasmus University Rotterdam
Department of Public Health - Iris Lansdorp-Vogelaar, Prof. dr., Andrea Gini
Korean National Cancer Centre (NCC)
National Cancer Control Institute , Goyang, South Korea - Il Ju Choi, Dr.

Role in the proposed research

The study is a joint collaboration between IARC, Erasmus University Rotterdam (EUR) and the Korean National Cancer Centre (NCC). The project will be coordinated by F. Meheus who will also be responsible for model development with A. Gini (currently completing a PhD at EUR and will join IARC as a post-doctoral fellow in the context of this project). Inputs for model development will be provided by JY Park and V. Knaze for the epidemiology of gastric cancer, I. Lansdorp-Vogelaar for technical expertise in the field of microsimulation modelling and MISCAN, and IJ Choi for the biology and management of gastric cancer, as well as for expert input on the epidemiology and management of gastric cancer in South Korea.

Details of proposal – vision, aims and influence on open research

Problem statement.

Gastric cancer (GC) is the third leading cause of cancer deaths worldwide (800,000 deaths, both sexes). In particular in East Asian countries but also in Europe, Central and South America, GC is an important public health problem. Current efforts have focused on screening tests to find GC at an early stage. For instance, in South Korea, endoscopy screening is part of a nationwide GC screening programme. However, screening for GC is expensive and there are uncertainties surrounding its efficacy. An alternative to screening would be to prevent GC from developing by providing antibiotic treatment to individuals testing positive for H. pylori infection, which is a Group 1 carcinogen for GC development. While randomized controlled trials (RCT) provide the strongest evidence, they cannot incorporate all factors related to e.g. the natural history, demography or epidemiology of GC necessary to assess and compare the effectiveness of different prevention and screening strategies. In these circumstances, mathematical simulation modelling of the natural history of disease can be a valuable tool since all strategies can be simulated using valid data, and long-term benefits and harms can be quantified.

The overall aim is to develop an open-source microsimulation model of the natural history of GC combined with a user-friendly interface to inform evidence-based decision making in GC prevention and control. Our vision is that access to a validated, well documented and open-source model will provide useful and timely information that will assist policy makers worldwide in elaborating efficient and equitable cancer control strategies to improve health outcomes and decrease the burden of cancer.

Our aim will be achieved through the following two activities:

- Develop an innovative microsimulation model for GC extending and revising the structure of the open source R-based EPIMETHEOS platform developed at IARC[1] with gastric specific modules (H. pylori infection/prevalence, natural history of GC, GC screening, and H. pylori eradication). The additional modules will be structured using the established MISCAN know-how in microsimulation models (Erasmus MC)[2] and the gastric carcinogenesis sequence proposed by Correa.[3] The model will be calibrated using an approximate bayesian computing algorithm informed by data from cancer registries and cohort studies.

- Model validation. Achieving a valid model is key for informing and optimizing GC screening policies. With our collaborators, we will externally validate the model replicating the epidemiology of GC in South Korea using data from a range of sources, including a large RCT on the impact of H. pylori eradication on GC incidence conducted by IARC and the Korean National Cancer Centre (the HELPER study[4]) and data from the Korean National Cancer Screening Programme.
We target the following user groups:
1. Health sector planners, policy makers and regional and country officers of the World Health Organization: The evidence provided by the project will improve strategic and evidence-based decision making. We target users globally, with a particular focus on users in low -and middle-income countries that are involved in decision on what cancer control interventions to include in the benefit package in the context of universal health coverage.
2. Research community, both academic and from public or autonomous bodies such as health technology assessment agencies. In addition to a user interface, we will also make the source code of the model available so that the research community can use the model to include additional screening alternatives and/or adapt the model to different country contexts.
Influence on open research practices: Many of the current models for GC are proprietary, limiting our capacity to independently verify the model assumptions and conclusions. Providing the open-source code of the model will maximize transparency and reproducibility but also improve acceptability of the model conclusions to policy makers. It will foster the exchange of knowledge and information through continuous engagement with the modelling community. In general, this proposal is part of an effort at IARC to conduct open science with increased availability of software, data and publications. Similar efforts at the Agency include the availability of global cancer statistics (http://gco.iarc.fr/) or repositories on the characteristics and performance of cancer screening programmes across the globe (https://canscreen5.iarc.fr/). Through IARC's leadership role in cancer research, we hope that the current project will synergize cancer modelling efforts worldwide towards the development and use of open-source approaches.

Details of proposal – evaluation plan

The main target of the project is the successful development of a natural history model that is validated against epidemiological data from South Korea. Success indicators will include (i) feedback on the usefulness of the model by researchers and policy makers in South Korea (e.g. the National Evidence-based healthcare Collaborating Agency), (ii) engagement with the gastric cancer community at the International Workshop on Helicobacter & Microbiota in Inflammation and Cancer and during either the Asian Pacific Digestive Week or the International Cancer Screening Network (ICSN) conference. These events will provide an opportunity to receive feedback on progress in model structure and development. And (iii) the ability to attract additional funding to add a module to estimate the cost and cost-effectiveness of the different GC prevention and screening strategies. Upon completion of the model, a dedicated website will be created using the shiny R-package (not budgeted).
References: [1] Baussano I, Lazzarato F, Ronco G, Franceschi S. Impacts of human papillomavirus vaccination for different populations: A modeling study. Int J Cancer. 2018;143: 1086-1092. [2] Vogelaar I, van Balegooijen M, Zauber AG, et al. Model profiler of the MISCAN-Colon miscosimulation model for colorectal cancer. Deparment of Public Health, Erasmus MC 2004; available from: http://cisnet.flexkb.net/mp/pub/cisnet_colorectal_sloankettering_profile.pdf [3] Correa P. Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res. 1992;52: 6735-6740. [4] Helicobacter Pylori Eradication for Gastric Cancer Prevention in the General Population (HELPER). Available at: https://clinicaltrials.gov/ct2/show/NCT02112214.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal from a strong team. However, it was felt the level of impact could be limited and it was unclear how this proposal related to other activities.

Title

Open StatisticsConsortium(OSC) for Young Researcher

Lead Applicant

Dr Gayatri Vishwakarma

Details of proposal – team members and collaborators

1. Dr. Gayatri Vishwakarma: Head Biostatistics, Indian Spinal Injuries Centre, New Delhi 110070 – Concept Designing, implementation, hostingand Coordinatingthe project
2. Dr. Abhaya Indrayan:Founder Professor of Biostatistics, and Head (Retd.)Department of Biostatistics and Medical Informatics, Delhi University College of Medical Sciences, Dilshad Garden, Delhi 110 095–Expert Member
3. Dr. R.M. Pandey: Head Biostatistics, All India of Institute of Medical Sciences, New Delhi – Expert Member
4. Dr.Karan P Singh:Professor and Chair, Department of Epidemiology and , The University of Texas, Health Science Center at Tyler, 11937 U.S Hwy 271, Tyler, Texas 75708
5. Dr. Shrikant I Bangdiwala:Statistics Director, Population Health Research Institute, McMaster University ON Canada L8N 3Z5 –Expert Member
6. Dr.Subhash Chandra: Chief Biometrician, Agriculture Research Division, Agriculture Victoria – DJPR, Tatura 3616, Victoria, Australia–Expert Member

Details of proposal – vision, aims and influence on open research

(i) VISION:The visionof this proposal is to create aweb-basedopenand distinguish platform of statistics to bring inspiration and innovation to every young researcherespecially for low and middle income countries.

AIM:The aim of the project is to provide an open forum of statistics (one-stop-solution) where every researcher gets answer to their query. We believe that this forumwill help researchers to deal with common statistical problems such as sample size calculation, identification of appropriate randomization technique, identification of appropriate statistical method for their study data to minimize statistical errors in biomedical research.

TARGET AUDIENCES:The target audience will be researchersin biomedical field or related areaseeking help in statistical conceptand applications.

ACTIVITIES:The proposed expert consortium will be available to researcher on a click. User has to login after creating an account on Open Statistics Consortium (OSC)website. User has to put query on any of the statistical concept by selecting themes and subthemes. Query will be sent to expert members and user will get answer to the query within 24 hrs. OSC will be having experienced expertsfrom various renowned institutes. Currently we have got six experts to start with. The expert member will get reasonable annual remuneration to sustain them in OSC. We will utilize the available networks in the statistics community to increase the expertsfrom various fields of statistics. OSCwill act as the connecting bridge for researchers working in biomedical research and looking forthestatistical expertiseespecially from low and middle income countries. We will have expert cluster in each theme and subtheme of statistics such as sample size estimation, statistical analysis(subtheme: inferential statistics, predictive modelling etc.)and statistical writing/interpretation so that query from a user can be answered quickly. Automated algorithms will be created through machine learning to generatenew theme by analysing query databases. The backend of thisplatform will be managed by an Open StatisticsConsortium (OSC) consisting of eminent personalities working in the field of Statistics.Access to section one will be free and access to section two will be chargeable. Every specific query resolution will have specific cost. This revenue generated will be used toremunerate the backend team.Other than the query, theOSC will be containing blogs, videos and online chat options to resolve quick query on statistical concept. Advertising will be done using available social media networks. Success of OSC will be measured through “Web Performance Monitoring”and “Real User Monitoring”. Both arepassive approach thatreports performance data as experienced by website's actual users. Based on the analysis of success and satisfaction by the users, further this site may be

recommended to all reputed journals to incorporate statistical review from OSC before publishing manuscripts in their journal.
ii) Quality of how statistical assumptionsand methods are applied in biomedical research is quite poor. Statistical errors are being published in literature and most common error quoted was “not choosing appropriate statistical method for study data” followed by “misinterpretation of P-values and main study results”. Another example is from JAMA Internal Medicine manuscripttitled “Effect of Statin Treatment vs Usual Care on Primary Cardiovascular Prevention Among Older Adults”makes the classic statistical error of attempting subgrouping rather than by correctly modeling interactions. Theerror was compounded instead of adjusting for covariates when comparing treatments in the subgroups.This may be because of lack of knowledge ornon-availability of expert statisticians. OSC is designed where user will get full support discussing their protocol/objectives to get answers of their queryquickly. This will help research community to minimize statistical errorsand enhance biomedical research by reducing deaths dueto medical errors.

Details of proposal – evaluation plan

Monitoring plan will begin with identification of key performance indicators such as deciding what to track; tracking&managing dataand turning information into actionable insights. The most important and basic open source monitoring tool to monitor OSC will be Google Analytics(Google's Webmaster Tools or HubSpot's Website Grader)to trackuserreach and impact. By measuring users and comparing numbersperiodically (monthly, quarterly, annually), wecan determine if OSCuseris growing, and if so, how quicklyor slowly. We will also analyse traffic sources and measure bounce rate and session time. Some of the performance indicator will be uptime, time-to-first-byte, Full-page-load-time, broken links, user journey, database performance and website quality.Content measurement technique and content performance will be the integral part of the monitoring the success of the OSC. One of the key indicators will be to determinethe cost ofconversions and overall return on investment.
Measuring the conversion rate will help tohave an idea of how many users are visiting OSCwebsite, how many times they are visiting, where they are coming from, and how long they are staying.How many are sending query and how many of them got satisfied with the services they got? Or user is not taking any action?

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to provide an open forum for statistics support. However the potential impact and level of innovation proposed were considered limited.

Title **The applicant opted not to share this information**
Lead Applicant **Dr George Karystianis**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

Metadata-enriched knowledge extraction from medical research publications (M-Extract)

Lead Applicant

Dr Gustaf Nelhans

Details of proposal – team members and collaborators

The SSLIS (https://www.hb.se/sslis) team focuses on mining scientometric data from clinical guidelines and building semantic links with scientific publications. SSLIS work will be led by Dr Gustaf Nelhans, whose research focuses on scientometrics and citation analysis, with a specific interest on 'professional impact', together with Dr Johan Eklund whose work involves semantic methods of text mining and information visualisation.

ARC (https://www.athenarc.gr) activities focus on mining scientific publications, enriching them with semantic annotations, interlinked with other metadata and scientific literature objects. ARC work will be led by Dr Haris Papageorgiou, Research Director at ARC, who has extensive experience in text mining and knowledge discovery for biomedical texts.

Dr Ioanna Grypari has a strong background in R&D impact evaluation; she will be working on the analysis and semantic enrichment of scientific literature, and,

Aris Fergadis, Research Associate at ARC, with background and programming experience in Deep Learning methods and NLP; he will be working on the data modelling and text mining tasks.

Both teams have participated in Data4Impact (Horizon 2020, No. 770531), using text mining methods offering innovative ways of knowledge discovery in the scientific literature. The two institutions possess the necessary infrastructure and complementary expertise for carrying out the proposed research.

Details of proposal – vision, aims and influence on open research

Finding relevant and up-to-date elements of information in large volumes of research is a demanding task for researchers and practitioners alike. A question answering system (QAS) is designed to retrieve specific responses, in contrast with a traditional information system that provides general query answers. For instance, in the context of medical information, a surgeon in need for critical information for a procedure, or a researcher who needs the latest corroborated knowledge regarding a specific aspect of a medical condition would need precise information that matches the professional's information needs at the moment.

To address these kinds of issues, we propose the development of a workflow for extracting and analysing findings and claims presented in medical research articles and storing this information in an open machine-readable format such as RDF triples.

The novel approach suggested here is to extract contextual information in the referring documents ("citances") and combine it with information extracted from the cited documents (knowledge extraction). An integral component of this approach is to enrich the information with topical keywords based on subject labels obtained through topic modelling of the content of the cited sources. Our approach will enhance and improve existing information services for medical professionals. Specifically, we will use clinical guideline documents as a source, which, in addition to cited references, provides qualitative information about excluded references and additional references, as well as a grading system for grading clinical evidence that will complement the contextual information extracted around cited references in the text.

Using the vast amount of medical research published in the MEDLINE database and made available through the PubMed search interface we have access to approximately 29.8 million records currently available in the PubMed index, with about 7.1 million records that are complemented with a link to the free full-text article.

Based on our experience in the Data4Impact project, we will be exploiting our collection of health publication data. This collection encompasses:

Free full-text in XML format with metadata from PubMed;

Additional Open access publications from OpenAIRE;

Clinical guidelines from national and international providers.

From the scientific publications, we extract citing sentences (citances) and group them by the paper they reference. Citances are considered a summarization of its key points and also its key contributions and importance within an academic community. They offer a scientific summarization of the cited paper which complement the reader's context, possibly as an author of a literature review.
Another important aspect of the knowledge extraction process is the identification of outcomes (such as methods, devices, and products) and relations (such as treatments and therapies). We base our methods on well-established techniques and tools for named entity recognition and relation extraction. Named entities and relations are the base components needed to create an ontology. Ontological Engineering is useful because it gives the systems the ability to recognize the context they are operating in and reason about those contexts. Ontology is becoming the pivotal methodology to represent domain-specific conceptual knowledge in order to promote the semantic capability of a QAS.
One critical aspect of the knowledge extraction process is the presence of hedged statements in the texts, i.e. statements that are not framed as scientific claims but express uncertainty or speculation. The challenge of identifying such statements will be addressed using established hedge cue detection mechanisms and will be incorporated in the proposed workflow.
The final outcome will be:
A model for generating workflows using non-proprietary software for extracting and using latent information in full-text collections.
A performant knowledge representation scheme based on the use of knowledge graphs stored in the open RDF standard as semantic triples (consisting of statements on the form subject, predicate and object). The RDF triples constitute a standardized language for question answering systems and decision support systems which could be used by practitioners and researchers in the medical field.
Ontologies generated from the full-text collections containing relations between concepts extracted from the texts.
A question answering system using logical inference would then be able to provide answers to specific information needs of the following kinds:
'What are the most probable conditions underlying this particular set of symptoms?'
'How efficient is treatment X for condition Y?'
'What are the known adverse effects of medicine X?'

Details of proposal – evaluation plan

The proposed work will enable and support researchers in knowledge discovery and generation, offering efficient ways to search, extract, interlink and summarize scientific literature, thus, going much beyond the capabilities of current keyword-based / bibliographic/metadata services.
To evaluate the performance of the proposed model for extracting structured information from research articles in full-text, we will develop a search service based on an RDF search engine. Using such a search service, we will evaluate the knowledge graphs produced by the proposed workflow by means of metrics focusing on the proportion of correct statements produced by the search service, such as precision, recall, and mean reciprocal rank. The evaluation will be performed in cooperation with a group of clinical practitioners working in prehospital care. This search engine will also facilitate the development of a dialog-based voice search service encompassing conversational language (cf. Google RankBrain)
Moreover, ARC, through direct connections to the European Open Science Cloud/EOSC (i.e., OpenAIRE, OpenMinted, e-InfraCentral, EOSC Pilot, NI4OS Europe) and participation in the EOSC Executive Board, will build and solidify bridges between the proposed Project and the open science content and infrastructures produced and built in the European Research Area.
Additionally, previous SSLIS research on clinical guideline citation data has been used as a component in the so-called "ALF funding", the national performance-based funding of clinical

research provided by the Swedish Research Council, that will be invited to monitor the results of the present work.
Decision Not shortlisted
Comment on decision from Wellcome This was felt to be an innovative and potentially impactful proposal. However, the level of openness of the outputs was unclear and the approach was lacking in detail.

Title

Opendataclinica as a platform for the exchange of open data from clinical studies and biomedical information.

Lead Applicant

Dr Heinz Nicolai

Details of proposal – team members and collaborators

Patricio Araneda, Dataclinica SpA, Software Engineer and Master in Medical Informatics.Project leader in IT area. He will coordinate the aspects of data design and requirements analysis. Programming and analytics.

Rodrigo Galvez, Dataclinica SpA, Software Engineer and Master in Medical Informatics. In charge of programming mobile applications related to patient access.

Marcela Jara, Nurse, specialist in Diabetes and cardiovascular disease.Medical advisor in the area of chronic diabetes and cardiovascular pathologies. In charge of collecting medical requirements for specific pathologies.

Arlette Cassot, Graphic designer and public relations specialist. In charge of interface design and user experience. Will manage the tasks of community manager in the relationship with patients.

Details of proposal – vision, aims and influence on open research

In clinical registries, it is essential to integrate, standardize and harmonize the information that each health institution collects, in order to achieve real data integration. The lack of specialized and integrated programs makes it difficult or almost impossible to exchange data to generate new knowledge in the patient's health.

Both in Chile and in Latin America, there are no policies or culture of data exchange in the area of clinical research, which delays scientific progress and leads to mismanagement of economic resources. This is due, in part, to the lack of initiatives and appropriate tools. Therefore, we intend to centralize and/or unify the information in order to carry out more efficient management of research efforts, both at a national and regional level, following the guidelines defined by the European Open Science Cloud (EOSC) and Fair data in health (FAIR4Health project) among others.

The main objective is to implement an integrated biomedical information tool for the management of clinical and biomedical studies and to generate a set of open, anonymous and standardized

research data that allows new comparative research. For this purpose, the initiatives of the European Commission related to the generation of FAIR data (Findable, Accessible, Interoperable and Reusable.) and indications from the Research Data Alliance to promoting the use of FAIR data in health will be used.

It also seeks to promote the commitment of patients diagnosed with specific pathologies to a process of active registration of their health information (PHR) and participation in clinical and/or pharmacological studies. To this end, a patient repository will be managed with truthful and efficient information and with approved informed consent for participation in clinical studies. This initiative is aimed at clinical researchers, patients and physicians:

Clinical researchers: They can collect and organize information from any area of clinical and/or biomedical research. Allowing the monitoring of their progress as well as the administration of informed consent and all necessary documentation to comply with legal and safety regulations. The information collected can be grouped into specific areas as needed by the researcher like clinical, laboratory, epidemiological surveys and others.

Patients: Allows associate patients to actively participate in health information records (PHRs) as well as own the information and be able to use that information in other health care benefits that can be useful or necessary. In this way, the patient can integrate patient groups with their health records and allow their participation in academic and/or pharmaceutical clinical studies.

Physicians: Active participation in the definition of the information that needs to be registered for the specific pathology, according to a standardized registration model. The physician will have at

his disposal a detailed history of the patients registered in this system and will be able to carry out detailed analyses of information in any area needed like image, laboratory, clinic and surveys.
Main activities
The acquisition of a software license of Opendataclinica is contemplated, to generate on it a series of additional applications to approach the different pathologies to study. In order to evaluate the flexibility of the system, three (3) pathologies will be implemented initially: diabetes, childhood epilepsy and prostate cancer. For which specific forms will be defined as well as the development of mobile applications oriented to the registration of information from the
patient.
Patient recruitment and informed consent will be carried out through patient associations.
In order to evaluate the flexibility of this system, three (3) mobile applications will be implemented that will store the information towards Opendataclinica. These mobile applications are oriented to
collect information from patients with specific pathologies, among them are initially: diabetes, childhood epilepsy and prostate cancer

Details of proposal – evaluation plan

The evaluation of the system shall consist of the following measurements:
Number of clinical studies
We consider having at least three (3) approved and/or ongoing clinical studies at the end of theperiod.
Registered Patients
It is considered to have registered at this stage, at least two hundred (150) patients with approvedinformed consent.
Mobile applications
Three mobile applications working in diabetes, childhood epilepsy and cancer.
Publications/congresses
It is considered to have at least one (1) approved publication in an international journal and at leastone (1) presentation in an international congress.

Decision

Not shortlisted

Comment on decision from Wellcome

This was felt to be an innovative and interesting proposal. However, the level of openness was unclear and there were concerns over the feasibility of the proposal

Title

Generic Ordinary Differential Equation based modelling tooL (GODEL)

Lead Applicant

Dr Ian Hall

Details of proposal – team members and collaborators

The project will be led by Ian Hall (Joint appointment University of Manchester and Public Health England PHE). He will manage day to day interactions with software developers and be responsible for delivery.

Dr Thomas Finnie (principal modeller PHE). Currently lead developer in PHE of the PyGOM tool and is researching a spatial modelling extension. He will be involved in steering group of project.

Dr Thomas House (University of Manchester). Provide advice around modelling of infection in subgroups and coding for multi-core architectures.

Dr Rob Haines (University of Manchester). Head of Research IT, will manage the research software engineer.

Professor Caroline Jay (university of Manchester). Will advise the research software engineer locally on human computer interactions and visualisations

Research software engineer (university of Manchester). Will conduct work and be named in later stages of application.

Details of proposal – vision, aims and influence on open research

i) This proposal is to develop existing computer based model development systems to account for user interface and connectivity with other systems to improve usage uptake. Transparent and reproducible disease modelling is essential for Government modelling with lapses in coding and model assumptions causing issues with model credibility (for example see the Aqua book, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/416478/aqua_book_final_web.pdf).

Specifically during major international disease outbreaks a wide community of analysts identify data and provide estimates of future case numbers and impact of interventions. Not all members of this community have an extensive track record in the domain and so estimates may be of variable quality. The main workhorse of epidemic modelling, especially under rapid timescales, is the compartmental model framework using ordinary differential equations. PHE modellers developed a rapid development tool named PyGOM (https://github.com/PublicHealthEngland/pygom and https://arxiv.org/abs/1803.06934) and a data interchange format EpiJSON (https://www.sciencedirect.com/science/article/pii/S1755436515000973) to provide rapid quality assurance of outputs.

This has been used on a number of projects to date with extension to spatial modelling and parameter estimation. However to reach out to a wider community the tool requires an enhanced user interface (either web based or using tools like Kivi (https://kivy.org/#home) to give desktop + mobile) which would greatly increase the utility and the ease with which newcomers could use PyGOM. Using software developer time on the interface would take the modeller focus from the coding and place it on the model saving valuable development time. Separate to the UI, but possibly related, the model description should be file based rather than code based with meta-data summarising key assumptions. This would allow the easy interchange of models between groups/people. Developer time can also support extension of the modelling framework to allow for additional subgroups of the population.

The audience is intended to be both the expert disease modelling community, building on initiatives like Epirecipes, the novice developers in early career stages and government analysts that need to review and validate model outputs for senior decision makers. We propose to use the funds available to fund a research software engineer over 12 months, as well as host 3 workshops. The first workshop would be with a select group of government and academic researchers to define the key gaps, the second workshop will be with the wider user community

to demonstrate concepts and development opportunities and a final workshop to demonstrate the working software. Travel and subsistence for these meeting will be costed in for attendees. As well as these wider workshops (forming milestones below) we would have monthly skype meetings with PHE to road test features and integrate the UI with PyGOM under open source licensing.
ii) By providing an open access software tool we intend to incentivise researchers to practise open research by improving how good practice is embedded in the research cycle, sharing of scripts can support transparency during peer review process as well as enhance uptake of findings after publication.
We intend also to provide a platform where research outputs in the form of models and data underpinning them are made accessible, re-usable and reproducible. This will be done by enabling a standardise input, building on tools such as EpiJSON and leveraging existing communities/platforms such as Epirecipes.
The development of open tools, will enable combination or repurpose of datasets, providing a citable link for disease parameterisations.

Details of proposal – evaluation plan

The principal investigator will have fortnightly meetings with the research software engineer with every other meeting involving a skype call with colleagues in PHE. On top of the physical meetings below.
We propose a 12 month project with the following milestones:
Milestone 1 – month 1. Steering group meeting to define scope of community needs (mix of academic and government based analysts)
Milestone 2 – month 6. Meeting of steering group and wider community to review interim progress and map out translation of tools to the user community.
Milestone 3 – month 12. Workshop with user community to evaluate impact, benefits and risks of toolbox.
Success will be measured by uptake of the tools between milestone 2 and 3 and the sustainability of the user community over longer term. We intend to have a web presence which can monitor visits and downloads of materials available. Each workshop will have attendee feedback forms to evaluate user perceptions and opinions.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a clear proposal to enhance an existing software tool. However, the level of utility and impact was considered limited, and the evaluation plan could have benefited from more detail

Title

Providing open access to data in Kenya to support Pathology and Lab Medicine ('Open PALM')

Lead Applicant

Dr Jacob McKnight

Details of proposal – team members and collaborators

Jacob McKnight, NDM Trop Med, Oxford – Overall PI; app design; database design; exploring opportunities for sharing data; team management; stakeholder engagement. Jake is an experienced health system researcher and entrepreneur.

Mike English, NDM Trop Med, Oxford and KEMRI-Wellcome – Guidance on overall project; monitoring and evaluation; negotiations with Kenyan partners; stakeholder links and research design. Mike is a Wellcome Trust Senior Research Fellow.

Felix Bahati, KEMRI-Wellcome – Database design, collection of data, and negotiations with Kenyan partners. Felix is a driven, talented young researcher and he will be involved in all aspects of research design and data collection.

Chris Paton, NDM, Oxford – app design process; IT integration and data sharing. Chris leads the Global Health Informatics research group which investigates open source and open data projects including DHIS2 and OpenMRS in Kenya.

Shahin Sayed, Pathologist, Aga Khan University – making links to lab registries and lab networks, senior pathologists. Shahin is a senior Kenyan pathologist involved in the many pathology and lab networks in Kenya and the region.

Ken Fleming, Pathologist, Oxford University – lab registries, lab networks, pathology. Ken is a well-connected pathologist who is involved in the Essential Diagnostics List (EDL) process and has campaigned for open data in pathology.

Details of proposal – vision, aims and influence on open research

This project was awarded a Gates Global Challenge Exploration grant (OPP1180942), but the work we outline here is additional, and speaks to making medical lab test information more accessible. We illustrate this by showing how the work will inform the development of the WHO Essential Diagnostics List (EDL) and, by extension, other lab research in general. Anonymised usage data will be made available to all through a highly novel platform and the project is designed so as to encourage broader data sharing in the important new digital health 'marketplace' industry.

Vision

Our Vision is to create marketplace app that provides improved data on lab test price, TAT and availability to patients and clinicians, but also provides much needed data on lab functionality to regulators, lab providers and researchers. This app, which is already designed and tested, naturally collects data on the demand for, and supply of, lab tests. By collecting rich, highly specific data on lab usage through the app, we can use this entirely new source of data to improve on the data available to the EDL process and Pathology and Lab Medicine (PALM) research both in Kenya and internationally.

Background

A host of new 'marketplace' apps have sought to take advantage of the same logic that underpins Uber, Ebay and Amazon. These companies all combine rich information with customer feedback and ratings to allow users to make a more informed choice about healthcare providers. We think we can use the system we have developed to generate data on lab test usage that will inform the development of the EDL. Additionally, we envision pioneering data sharing in this space and encouraging some of the largest apps to also share their data.

Aims

Establish a design for the data 'dashboard' with stakeholder input

Produce a working app/dashboard for viewing PALM data

Collect data through the app by initiating trials at two Nairobi hospitals

Share the data with the EDL team

Share the experience and approach with Practo, 1mg.com and other marketplace apps that have millions of monthly users on their systems.

Target Audiences

Our first audience is the EDL team. There are approximately 100 people working on developing both the international EDL (now in its second iteration) and many more working on national variations. They tend to rely on lab registries which show simply the tests done. Unfortunately, this is a very limited source of data because it does not show demand – if a test is not available, it will not appear at all in the registry. As such, the app could provide much deeper, more useful data: it could record lab requests, location of requests, and the availability of the tests.

Our second audience are the new marketplace apps working in this space. We want to appeal to this emerging sector to share their data in the same way. We are in discussions with 1mg.com who alone have 7million monthly users in India. We want to free their data using our platform to support broad PALM research.

Activities

Stakeholder (online) meeting with EDL members to co-design dashboard: we are connected to some of the key players in the EDL process.

Make the dashboard with the help of a professional developer

Conduct two trials at hospitals in Nairobi: 3-5 clinicians in each site, using the app for 1month.

Test dashboard and produce data and make data available to EDL planners

Share approach with 1mg.com and other large sectoral stakeholders

How will the proposed work influence the field?

The Essential Diagnostics List holds enormous potential for medical practice around the world, and particularly in LMICs. Our work will make more useful data about lab usage available to those working in this space, and will guide the next iteration of the EDL. Additionally, the data we produce will be of particular use to the development of a Kenyan EDL.

We also aim to influence the fast-developing field of online marketplace apps. These apps collect vast amounts of data on PALM usage, but as yet do not share their data. By establishing a template to do this, we will set a standard for sharing open data in this important area.

Details of proposal – evaluation plan

We will draw together a small stakeholder group from the network of people working on the EDL to provide overall governance and guidance to the project.

a) Conduct Stakeholder meeting to co-design PALM data dashboard – Oct, 2019.

The outputs for this objective will be a template wireframe of the EDL dashboard approved by the stakeholders we draw together.

b) Build app/dashboard – Dec 2019.

A Minimum Viable Product of the app is already complete. The developer is in place and well-tested. The output of this section will be the completed app and dashboard.

c) Collect data through the app by initiating trials at two Nairobi hospitals (currently have permission for two public, one private) – Jan, 2020

Ethical permission has been granted and we have already run one-day sessions as a proof of principle, but this stage will feature a longer term – one month in two locations – to test the app on clinicians' phones. The output will be the data they input into the app, and a successful trial (regular use by clinicians and no abandonment).

d) Share the data with the EDL team – Feb, 2020

We will use the dashboard to share data with the stakeholder team and the wider EDL community. The output will be their feedback.

e) Share the experience and approach with Practo, 1mg.com and other marketplace apps that have millions of monthly users on their systems – April, 2020

The output for this section will be feedback and engagement from each of the targeted marketplace app providers.

Mike English is a Wellcome Trust Senior Research Fellow and is very well aware of what is required in terms of reporting. He will guide the process above.
Decision Not shortlisted
Comment on decision from Wellcome This was felt to be an innovative proposal. However, there were concerns over the feasibility of the proposal approach, and the level of utility and take-up that would be achieved.

Title

Open, reproducible analysis and reporting of data provenance forhigh-security health and administrative data

Lead Applicant

Dr Jessica Butler

Details of proposal – team members and collaborators

Professor Corri Black: Clinical Lead –Grampian Data Safe Haven (DaSH)*, Director –Aberdeen Centre for Health Data Science, Associate Director – Health Data Research UK, expertise in data governance, secure data environments, health informatics methodology

Dr Jessica Butler: Research Fellow – Aberdeen Centre for Health Data Science, Open Research Champion –University of Aberdeen, local lead – UK Reproducibility Network, expertise in large-scale data linkage of health and administrative data and in open research practices

Ms Carole Morris: Director of Electronic Data Research Information Service (eDRIS)*, Acting Head of Service at NHS National Services Scotland, expertise in secure data environments and the extraction and preparation of health and administrative data for research

Dr Nir Oren: Head of Department of Computing Science, expertise in accountability and provenance in intelligent systems

Dr Magdalena Rzewuska: Research Fellow – Health Service Research Unit, expertise in multidisciplinary and participatory research, improvement and implementation evaluation methods

Ms Katie Wilde: Technical Lead – Grampian Data Safe Haven*, expertise in secure data environments and the extraction and preparation of health and administrative data for research

*DaSH and eDRIS are secure, accredited research facilities which provide access to health and administrative data for research. They support access to full electronic health records and other national datasets.

Details of proposal – vision, aims and influence on open research

Aim: To codesign, pilot, and evaluate FAIR (findable, accessible, interoperable, reproducible) methods for tracking and reporting data provenance in high-security data research settings

Background: A wide variety of national-level health and administrative data are available for research, including hospital admissions, outpatient visits, prescribing, education, work and pension, and census data. Access is strictly governed because the data are sensitive and participants have not explicitly consented to their use. Governance of this data focusses on preserving anonymity and confidentiality by strictly segregating the data and processing from their use for research. Raw data, cross-dataset linkage, data extraction and cleaning, and anonymisation are done separately from research by diverse data custodians (NHS information services, local authorities, government departments and national records agencies). Researchers themselves access a minimal, processed, anonymised dataset within secure environments.The opportunity cost of this strategy is loss of transparency of the data origins and processing history, and, more seriously, an increased risk of undetected error propagation. Current procedures for capturing and reporting data provenance are fragmented across data custodians, often labour intensive, and rarely shared with researchers. The team involved in this project includes both data guardians and researchers who have discovered and resolved errors in high-security data processing that were difficult to detect due to lack of reporting of provenance. We believe that the risks to research quality due to the opacity of data handling are as large as the risks of privacy breaches. If data provenance is not available, the resulting research can be impossible to assess and reproduce, wasting both government resources and public good-will for data sharing.Here we propose to improve data provenance tracking and reporting within the high-security data setting of the NHS Grampian Safe Haven.

Objectives

1) Develop and test automated methods of capturing data provenance within a high-security data setting

2) Co-develop tools to report provenance that are acceptable to data guardians and also meets FAIR guidelines for transparency and reproducibility

Activities

Objective 1:Capturing provenance involves tracking the entities and activities which generated a data set. We will develop and evaluate automated methods to: 1) identify source datasets and relate them to derived datasets; 2) record the logic used to link datasets and create extracts; 3) report quality measures for linkages; and 4) record the people/institutions responsible. This data provenance will be provided for all research projects within the Safe Haven. This record is absolutely necessary to evaluate the quality, reliability and trustworthiness of the data, but it may not be the ideal format for public dissemination, which we address below.

Objective 2: Provenance reporting that meets FAIR guidelines is necessary for the evaluation of health data research. We will run two workshops with data guardians and researchers working with data in high-security settings. Participants will be given examples of the data provenance captured in Objective 1 and asked how to balance privacy protection with the need to meet guidelines for transparency and reproducibility. The output of these workshops will be tools to concisely summarise data provenance for transparent and reproducible reporting.

Target audience: Both guardians of and researchers using unconsented, high-security data in settings like NHS Data Safe Haven, Health Data Research UK, and Administrative Data Research Partnership sites.

Influencing Open Research Practice: Tools for transparent capture and reporting of data provenance will be a step-change in open science practice in health data science. None currently exist.

We will make the methods, code, and stakeholder feedback publicly available immediately (MedRxiv/Github). We will implement the methodology on all future projects in NHS Grampian Safe Haven (lead by co-applicants Black and Wilde), and begin testing within the NHS National Safe Haven (lead by co-applicant Morris), with the goal of implementation across NHS Safe Havens and Health Data Research UK sites. We will also share the methods with data guardians of a health service secure setting in Brazil (project lead by Black and Rzewuska). Finally, this work will provide information needed for deciding how to govern sensitive data. Structured data provenance from these complex and high-security settings will allow researchers to study how storage and processing methods effect data quality. This type of assessment is necessary for deciding how to most responsibly govern use of high-security data.

Details of proposal – evaluation plan

The project team are experienced collaborators with a range of expertise managing and using high-security health and administrative data. Co-development with data guardians, researchers, and policy makers is at the foundation of our way of working. This participatory approach allows the key users of the main output to have a voice and ensuresmore effective results. Jessica Butler, Nir Oren, and Katie Wilde will work primarily on the capture of data provenance. Corri Black, Carole Morris, and Magdalena Rzewuska will work primarily on the stakeholder workshops and communicating provenance.

Activities across the 12 months will be as follows:

Identify exemplar high-security data use cases

Workshop 1 -risks & benefits of reporting provenance

Map data processing pathways

Develop provenance capture methods

Workshop 2 -communicating provenance FAIRly

Develop methods to query and communicate provenance

Outputs for the project will be the method (algorithm, metadata) for capturing provenance, the method for reporting a summary of the provenance that meets FAIR standards, as well as a report of the co-design process with stakeholders. All will be made publicly available immediately via

preprint server and will be submitted for journal publication. In addition, as a second phase following on from this work, they will be implemented in other NHS Safe Havens.
Decision Funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

PSI: a web portal for sharing population stratification information in large-scale biobanks

Lead Applicant

Dr Jie Zheng

Details of proposal – team members and collaborators

Dr Daniel Lawson, Sir Henry Dale Wellcome Trust Research Fellow at the University of Bristol until Sept 2019 and Lecturer in Data Science afterwards. His fellowship covers "Statistical methodology for population genetics inference from massive datasets with applications in epidemiology" and has led to the conclusion this tool is necessary. (Aim1,2,3 and 4).

Dr Bilal Ashraf, Research Associate in Population Genetics at University of Bristol. His current work focuses on GWAS and meta-analyses of different population groups world-wide using UK Biobank and simulated data from 1000 genomes project. He will play a key role in developing the GWAS results using available population resources (Aim1,2 and 3).

Professor Tom Gaunt is the leader of the translational informatics group in MRC IEU and has contributed comprehensively to open research during his research career. He will be the senior advisor of the proposed project. (Aim3 and 4).

Dr Ben Michael Brumpton, Research Fellow in K.G. Jebsen Center for Genetic Epidemiology, NTNU. Dr Brumpton is working with GWASs in the HUNT cohort. He has extensive expertise using HUNT data combined with other health registries from his research in Norway (Aim1).

Dr Benjamin Neale, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Broad Institute of MIT and Harvard. Dr Neale is the principal investigator of the UK Biobank GWAS release from the Broad Institute (Aim1).

Details of proposal – vision, aims and influence on open research

In human health research, large-scale biobanks, such as the UK Biobank and the HUNT study, are increasingly measuring a large number of phenotypes in the sample from different populations. It is therefore likely to become more common for large-scale GWAS studies of diverse phenotypes to be published from the same set of participants. Our collaborators Dr Benjamin Neale and Dr Ben Michael Brumpton have extensive experience leading large-scale GWAS in these biobanks. However, GWAS in such biobanks may bring issues such as subpopulation stratification. Our recent results (Lawson et al 2019; Haworth et al 2018) suggested that this can bias some genetic association results, yet this vital information remains unencoded in shared summary statistics. To allow correct comparison between studies there is an urgent need to develop software to enrich and standardise metadata. Average gender, average & covariance of geolocation, average and variance of age, and standardized representations of population structure are all necessary to quantify potential confounding.

Population structure can be standardized by constructing genetic predictions using standardized external datasets, for which predictors can be constructed using only SNP level information. Dr Lawson has developed such a measure using the international component of the UK Biobank as part of his Sir Henry Dale Wellcome Trust fellowship.

In this application, we propose to create an open source tool for constructing metadata describing population stratification and other confounding and develop a web portal to enrich and improve reusability of the metadata, which Dr Zheng have rich experiences in developing open source platforms such as LD Hub (Zheng et al 2017) and MR-Base (Hemani et al 2018).

We have four overarching aims:

Aim 1: a workshop to setup metadata standardization and sharing.

Aim 2: a case study using UK Biobank and HUNT to demonstrate the importance of obtaining and sharing standardised summary statistics using existing UK Biobank data.

Aim 3: develop software to gather standardized output from the formats of key software to be distributed via github.

Aim 4: develop a web portal to facilitate sharing and reporting metadata. Data users and a wider audience can interact with data generators.

Target audiences:
Funders - data sharing standards for future grant application.
Journal editors - data sharing policy for future GWAS publications
Principal investigators - of major cohort studies and GWAS consortia (in addition to UK Biobank and HUNT), such as Avon Longitudinal Study of Parents and Children (ALSPAC) cohort, Genetic Investigation of ANthropometric Traits (GIANT) Consortium and Psychiatric Genomics Consortium (PGC).
GWAS researchers - encouraged to share vital metadata.
Activities:
1. Establishing the core committee, comprising major stakeholders from MRC IEU, Broad Institute and HUNT. Organise a workshop to bring key stakeholders together to discuss and establish a standard for population structure of GWAS.
2. Conducting a case study using UK Biobank and HUNT to demonstrate the importance of obtaining and sharing standardised summary statistics using existing UK Biobank data.
3. Developing software to merge standardized output from the formats of key software (BOLT-LMM and PLINK) to be distributed via github.
4. Developing a portal to collect and harmonise genetic summaries of populations in a standardized way
5. Publishing findings in a leading journal and present them at a major conference.
Our project will influence open research in the following areas:
Change the way GWAS metadata is published (Aim 1,2). Essential quantitative population information is missing for most GWAS publications. Our project will further motivate researchers to report standardised population information in future GWASs.
Encourage researchers to practise open research (Aim 2). By starting this open practice for UK Biobank and HUNT, researcher will understand the importance of sharing population structure information.
Develop open source tool and platform (Aim 3 and 4). Open source software and a web portal will enable researchers to share their GWAS metadata.
This project aims to change researcher's behaviour towards open research and data sharing. Our tool and web portal will standardise reporting GWAS metadata and greatly improve the findability, accessibility, reusability and reproducibility of GWASs.
NB This is a separate and complementary proposal to that of Dr Lawson. Dr Zheng's proposal focusses on worldwide structure and simple tools. Dr Lawson's proposal focusses on subtle confounding. Both are important and would have independent impact, but would amplify the others' effect.

Details of proposal – evaluation plan

The proposed project will provide a reporting standard, software with portal, and encourage data sharing for subpopulation stratification. The impact plan for this project is:
1 Open source software to the health science research community
In the short term, this project will provide a tool to merge metadata to quantify potential confounding in UK Biobank and HUNT study. Measure: release of the software and portal.
In the longer term, the proposed project will provide an international portal for sharing and reporting metadata. Measure: monitor usage statistics via github and web analytics.
2 Data sharing to benefit the wider research community
The reporting standard of the population structure metadata will be expected by users of our software. Measure: share UK biobank and HUNT statistics with the wider research community via our web portal.
3 Academic impact
High profile journal publications and conference presentation that promote our software and web portal will establish our reputation globally. Measure: publications, citations, conference talks and posters.

Risks

Aim 1 will not bring enough stakeholders. Concern: low, since we have existing collaborations with major research groups, e.g. Broad Institute, HUNT. An appropriate UK Biobank project approval (Aim 2) is held in-house. Collaborator Ben Brumpton has access to, and will coordinate analysis in the HUNT study.

Aim 2 fails to demonstrate value. Concern: low, as Dr Lawson has (Wellcome supported) preliminary results demonstrating otherwise.

Slow development of the software and platform (Aim 3 and 4). Concern: medium that it is hard to totally automate data upload and validation from third parties. low risk that the platform will not be delivered, as the PI and collaborator Professor Gaunt has extensive experience having developed two platforms for open research.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal aiming to establish new standards for use in genome-wide association studies. However, the level of demand, and therefore the potential impact, was unclear.

Title

Development of a prototype of IARC's open-source research platform for worldwide H. pylori infection and gastric cancer epidemiology (Helicobacter In 5 Continents)

Lead Applicant

Dr Jin Young Park

Details of proposal – team members and collaborators

International Agency for Research on Cancer (IARC), World Health Organization:
Early Detection and Prevention Section, Lyon, France - Jin Young Park, Dr., Rolando Herrero, Dr., Viktoria Knaze
Genetic Cancer Susceptibility Group - Behnoush Abedi-Ardekani, Dr.
Information Technology Services - Philippe Boutarin (IT adviser), IT programmer (subcontractor)
This project will be coordinated by J.Y. Park who is also responsible for data collection together with R. Herrero and V. Knaze by conducting the ENIGMA studies of the International Agency for Research on Cancer of the World Health Organization (IARC). R. Herrero provides senior leadership to the project. V. Knaze is a project coordinator and data manager of the ENIGMA studies. B. Abedi-Ardekani will lead the whole slide imaging and web-based pathology review as the main ENIGMA pathologist. P. Boutarin will take a main advisory role with his extensive experiences in developing research platforms for large projects such as the Mutographs project and will supervise an external IT programmer who will be later recruited and develop an online platform dedicated to ENIGMA consistently to other open source web platforms such as GLOBOCAN or Cancer Incidence in 5 Continents, see http://gco.iarc.fr/.

Details of proposal – vision, aims and influence on open research

Background: Gastric cancer (GC) causes almost 800,000 yearly deaths worldwide, and despite declining trends, burden will not decline for decades because of population growth and aging. GC exhibits extreme between- and within-country variations. The main risk factor for GC is chronic Helicobacter pylori (H.pylori) infection usually acquired in childhood and generally persists without symptoms for life. H.pylori causes chronic gastritis and peptic ulcer disease, but also to premalignant lesions (atrophic gastritis, intestinal metaplasia, dysplasia) leading to GC in the series of events known as the Correa cascade.
Problems: The study of worldwide GC epidemiology has been hampered by lack of comparable data on H.pylori prevalence and premalignant lesions in each part of the world. The ENIGMA (Epidemiological iNvestigatIon of Gastric Malignancies) study led by IARC aims to investigate the epidemiology of H.pylori infection and GC in high and low-risk areas worldwide. ENIGMA has two components: prevalence surveys of H.pylori infection in population based samples (ENIGMA I); and prevalence studies of gastric mucosal changes (ENIGMA II). ENIGMA has been conducted in Chile and Iran, with ongoing plans to expand to Uganda, Colombia, Costa Rica, China, New Zealand with the ultimate goal of mapping H.pylori prevalence worldwide as an input for establishing public health interventions.
Vision: We expect to build a major resource for GC research by making standardised epidemiological data available (e.g. lifestyle, dietary, anthropometric) incorporating the ENIGMA Biobank, with blood, stool and urine samples together with histopathological images of gastric premalignant conditions, as a research and information tool for GC prevention which would target researchers and policy makers. Aim: We intend to develop a prototype for an open research platform (Helicobacter in 5 Continents) to share worldwide epidemiological, biological, and histopathological images from the ENIGMA studies centralized at IARC.
Brief design of the ENIGMA I & II: We recruit 700 participants (1-69 years old) from each site with contrasting GC risk using a population-based sampling method (ENIGMA I). The participants provide extensive interview data and blood, urine and stool specimens. The interview and H.pylori data will be stored using REDCap, which is a secure web application for building and managing online surveys and databases. ENIGMA I participants 40 years old and older are invited to endoscopic examinations and provide mapped gastric biopsies for detailed histopathological

examination of gastric mucosal changes (ENIGMA II). Half of the collected stool, urine and blood samples from the ENIGMA I and formalin-fixed paraffin-embedded (FFPE) gastric biopsies from ENIGMA II are shipped to IARC as per protocol. The transferred biospecimens are stored in IARC's biobank and tracked using an in-house sample management system (SAMI). At IARC, we will apply whole slide imaging and web-based pathology review for detailed standardized microscopic evaluation of gastric biopsies and scoring of gastritis. Digital pathology allows for sharing of the images with expert pathologists thus consolidating our pathology expertise at IARC. The processing of FFPE blocks, scanning of the glass slides to create high resolution images, data inspection/cleaning, task assignment and call for consensus review in difficult cases will also be conducted centrally by IARC through established SOPs and based on our previous experiences (e.g. Mutographs).

Proposed activities: We aim to develop a prototype for an interactive research platform which will link selected epidemiological and biological data from REDCap and digital pathology images (Helicobacter In 5 Continents). The prototype will be designed to enable global H.pylori prevalence estimates and stratified analyses of the prevalence and its risk factors by country, site, and the background GC risk to investigate both local and global epidemiology of GC. As the study evolves, selected results from relevant biomarker analyses, e.g. H.pylori genetics, metabolomics, and microbiome will be integrated to the platform expanding the scope of analyses.

Impact: Similarly to IARC's GLOBOCAN, Helicobacter In 5 Continents will be hosted via IARC's research website where anyone can freely access the ENIGMA study sites profiles. This open research platform will serve as an extensive tool for health research, especially concerning GC epidemiology for which exist limited understanding and scarce epidemiological and biological resources from different parts of the world. This online platform would help equity of access to the data thus facilitating research collaborations with the ultimate aims of eliminating GC worldwide.

Details of proposal – evaluation plan

The prototype of the Helicobacter In 5 Continents platform will be developed and evaluated within the following framework of deliverables:

1. Development of detailed functional specifications (4 months)

- Website specifications - expected pages, menus, links & interactions
- Databases specifications - to identify all the relevant databases, e.g. ENIGMA REDCap database, histopathological images in SlidPath
- Interface specifications- to define how to link the different databases, selection of data that the website will pick up, and whether the interface will be asynchronous/synchronous
- Definition of resources: technical (CPU, disks, memory and network capacity), persons (website developer, website editor, database administrator, etc.)
- Expected deliverable: detailed functional specifications, validated by the scientific and IT teams and ready for technical implementation and dissemination for request of quotations

2. Creation of mock-ups (2 months)

- Examples of expected layout for the menu and the pages for the website based on static images and data
- Detailed description of web page and interactivity zones
- Expected deliverable: visual representation of the detailed specification

3. Technical and functional implementation of the prototype (6 months)

- After writing of specifications and creation of the mock-ups the development phase of the specified modules and features will begin. As part of the functional implementation, throughout the project period we will collect and gather the scientific data for the platform.
- Populating the modules using a subset of the existing data kept at IARC from ENIGMA Chile and Iran or incoming data from ongoing/future ENIGMA countries, such as Uganda and Costa Rica, depending on their study progress, will be made.

- Expected deliverable: working website on a limited subset of data as a proof of concept before potential industrialisation. This website will be tested by IARC staff and some identified ENIGMA partners.
4. Feedback collection (2 months)
- Through specific surveys collect partners and website users' feedback on features and requirements for future development

Decision

Not shortlisted

Comment on decision from Wellcome

This was felt to be a good proposal with potential for significant impact. However, there were concerns over the level of openness of the work and it was felt the evaluation plan could have benefited from further development

Title

Implementation of a new model to associate mtDNA variation with complex traits

Lead Applicant

Dr Joanna Elson

Details of proposal – team members and collaborators

Dr Joanna Elson (project lead)

Research associate who will conduct the work detailed below to allow implantation of the variant load model without having to collaborate directly with the Elson group. Requests for such collaboration have been come too large to support and represent an analysis bottleneck.

Details of proposal – vision, aims and influence on open research

Mitochondria are cellular organelles that generate energy. They contain a small chromosome mitochondrial DNA (mtDNA); this codes for 13 proteins essential for energy generation. Mutations of mtDNA cause rare disease in about 1 in 4000. Additionally mitochondrial DNA (mtDNA) variation has been linked to susceptibility or course of disease for complex traits including diabetes, neurodegenerative disorders and multiple sclerosis. The vast majority of these studies have been conducted using the "haplogroup association" model. MtDNA undergoes strict maternal inheritance resulting in the evolution of mtDNA being defined by the emergence of distinct lineages, called haplogroups. The results of haplogroup association studies have been controversial with a myriad of conflicting results [1] emerging in the literature. One of the major reasons is population stratification of the common variants upon which the haplogroup association model is based.

I have developed a new model to link mtDNA variation to complex traits that produces fewer false positive associations and has greater statistical power. The new model considers complete mtDNA sequence data, analysing variation using up-to-date computational methods producing a single numerical metric to summarize the predicted level of 'mildly deleterious' variation on a cohort member's mtDNA. Allowing the application of parametric rather than non- parametric analysis, this permits better correlation of mtDNA variation to phenotypes measured in the cohort. Additionally the method focuses on variants predicted to be 'mildly deleterious', most of which are rare. These rare variants are not subject to problematic levels of population stratification thus fewer false positive associations are produced by this model. This model is termed the 'variant load hypothesis'. My group has applied this model to Alzheimer's [2], Myalgic encephalomyelitis [3] and Parkinson's disease (unpublished data). This method can be considered a rare-variant common disease model. Currently the implementation of this model is not automated restricting its application.

Aim 1: The principal aim of the project will be to produce a fully automated version of the model that can be used by the community as a tool via a simple web based interface. Investigators in the mitochondrial field and those with mtDNA data without experience in the field would both be able to use the tool. This tool will also allow groups with historical data to quickly re-examine their data. It would also support meta-analysis of datasets centred on a single phenotype. In this way the method would re-cycle prior data and allow for efficient analysis of newly generated data and thus confirm or refute prior results produced using the haplogroup association hypothesis. There has already been considerable interest in my model, with other groups having initiated collaboration resulting in published studies in Atherosclerosis [4], Oxidative stress and inflammation [5] and hypertension and diabetes [6]. Groups that I have worked with have recently started to apply this method independently [7]. The monies requested here would be used to pay for a staff member to produce a tool to allow this model to become available to all.

Aim 2: We will conduct a set of simulation studies to fully define the power of the new model, in a similar fashion to the work of Samuels et al 2008, who defined the power of the haplogroup association model. Specifically the power of the model will be tested for the following conditions: Hypothesis 1: An individual is more likely to be in the control group if they do not possess any mildly deleterious variants (variant load score = 0). Hypothesis 2: An individual is more likely to be

in the case group if they possess any mildly deleterious variants (variant load score >0.49). Hypothesis 3: An individual is more likely to be in the control group if they do not possess any mildly deleterious variants (variant load score = 0), and more likely to be in the case group if they possess any mildly deleterious variants (variant load score >0.49). Hypothesis 4: An individual is more likely to be in the case group if they possess two or more mildly deleterious variants (variant load score ≥1).

Thus this project would make a new model growing in popularity easily available to the community, allowing the re-analysis of existing data as well as the efficient use of new data.

1. Salas, A. and J.L. Elson, Raising Doubts about the Pathogenicity of Mitochondrial DNA Mutation m.3308T>C in Left Ventricular Hypertraveculation/Noncompaction. Cardiology, 2012. 122(2): p. 113-115.
2. Pienaar, I.S., N. Howell, and J.L. Elson, MutPred mutational load analysis shows mildly deleterious mitochondrial DNA variants are not more prevalent in Alzheimer's patients, but may be under-represented in healthy older individuals. Mitochondrion, 2017. 34: p. 141-146.
3. Venter, M., et al., MtDNA population variation in Myalgic encephalomyelitis/Chronic fatigue syndrome in two populations: a study of mildly deleterious variants. Sci Rep, 2019. 9(1): p. 2914.
4. Piotrowska-Nowak, A., et al., New mtDNA Association Model, MutPred Variant Load, Suggests Individuals With Multiple Mildly Deleterious mtDNA Variants Are More Likely to Suffer From Atherosclerosis. Front Genet, 2018. 9: p. 702.
5. Venter, M., et al., Implementing a new variant load model to investigate the role of mtDNA in oxidative stress and inflammation in a bi-ethnic cohort: the SABPA study. Mitochondrial DNA A DNA Mapp Seq Anal, 2019. 30(3): p. 440-447.
6. Venter, M., et al., Using MutPred derived mtDNA load scores to evaluate mtDNA variation in hypertension and diabetes in a two-population cohort: The SABPA study. J Genet Genomics, 2017. 44(3): p. 139-149.
7. Piotrowska-Nowak, A., et al., Investigation of whole mitochondrial genome variation in normal tension glaucoma. Exp Eye Res, 2019. 178: p. 186-197.

Details of proposal – evaluation plan

This project is important due to the large volume of conflicting studies in the literature concerning mtDNA variation and its relationship to complex traits, with this problem having been recognized for some time [3, 4, 12]. The success of the activity will be measured firstly by the number of groups that apply the model; thus far it has been used by myself [6, 7], by others in collaboration with myself [8-10], and papers are now emerging having applied the methods independently from myself [11]. There are a number of other groups wanting to work with me on similar projects. Another important measure of success will be if the tool is able resolve some of the conflicts in the literature by re-analysis of existing data using the new model. Users will need to sign up to use the tool. This will allow them to opt into receiving updates. They will also be asked the phenotypes that they are interested in. If the user then opts into giving this information they will be able to see a list of other users and the phenotypes upon which they work. This will be done with a view to strengthening the re-analysis of data. Specifically we will

*Require people to register to use the site. This will also allow us to keep them informed of updates should they opt in, as well as offering aid with interpretation and analysis.

*Ask them to cite the paper that will be written to report and publicise the tool. I anticipate reporting the re-evaluation of a number of datasets with the tool. This will act as an example of how the tool should be applied and boost its impact.

*After a three and five year period conduct a meta-analysis of the literature to determine the take-up and successes of the tool in resolving the conflict in the literature.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a proposal to create a resource with potentially high value to the mitochondrial DNA research community, with a good evaluation plan. However the level of innovation proposed was limited.

Title

Developing an open access journal authenticator tool

Lead Applicant

Dr Kelly Cobey

Details of proposal – team members and collaborators

Primary Investigators

These individuals will co-lead the day-to-day activities of the project ensuring its progress and success.

Dr. Kelly Cobey (Investigator, Centre for Journalology, Ottawa Hospital Research Institute)

Dr. David Moher (Senior Scientist, Director, Centre for Journalology, Ottawa Hospital Research Institute)

Collaborators

These individuals will support the theoretical development and testing of the tool, and with the knowledge translation activities.

Dr. Matthias Egger (President, National Research Council, SNSF); content expertise, knowledge translation, and policy expertise

Dr. Deborah Poff (Chair, Committee on Publication Ethics (COPE)); content expertise and knowledge translation

Dr. Tom Olyhoek (Director, Directory of Open Access Journals (DOAJ)); content expertise and knowledge translation

Dr. Jelte Wicherts (Professor, School of Social and Behavioural Sciences, Tilburg University, the Netherlands); content expertise, measurement expertise

Details of proposal – vision, aims and influence on open research

The open access (OA) publication model has enabled increased equity in access to information, and accelerated discovery by enabling work to be freely built upon. These and other benefits of OA publishing have led numerous funders globally to commit to ensuring all work they support is made available publicly. Plan S, an open science publishing initiative launched and supported by a consortium of funders, including the Wellcome Trust, is an example of a progressive policy change in this area. Public sharing can be accomplished via online platforms, but in many disciplines, sharing is most commonly achieved through publication of findings in OA journals.

Despite these positive policy changes, the integrity of OA publishing is being threatened. Predatory journals/publishers have entered this space. Many OA journals charge an article processing charge (APC) to publish accepted articles. Predatory journals/publishers have their own self-interest in mind, and they typically look to make profit from APCs, with little regard for what they publish. Several studies have shown that predatory journals/publishers are increasing, with their impact being felt globally. These journals do not aspire to best publication practices.

A likely reason for the increased penetration of predatory journals is that researchers, particularly early career researchers (ECRs), find it difficult to distinguish between legitimate journals and fake ones. Providing prospective ECR authors with information about journals, such as whether they are a member of the Committee on Publication Ethics, is likely to help them avoid predatory journals. Publication of work in predatory journals represents a waste of resources; the work is unlikely to be found or used.

vision, aims, target audience, activities

The vision for the proposed work is to develop an online journal authenticator tool. The proposed tool could be downloaded as a computer browser plug-in. Then, when viewing a journal homepage, a user could click the tool button in their browser tab and would obtain best publication practices information about journals. This would provide a signal to the user about which best practice standards the journal does, or does not, uphold. The tool would provide information such as: whether the journal is listed in the Directory of Open Access Journals (DOAJ), whether the journal is a member of the Committee on Publication Ethics (COPE), where the

journal is indexed, whether the journal has evidence of conducting peer review, and whether the journal clearly states its operation policies.
The work will address three aims:
To develop a prototype journal authenticator tool
To evaluate the journal authenticator tool
To disseminate the tool to stakeholders
The research team recently convened a 2-day summit on predatory journals attended by diverse stakeholders from around the globe. As part of this summit a Delphi survey was conducted, in which attendees and experts participated in three rounds of voting on issues related to predatory journals. Our international experts reached consensus (>80% agreement) that a journal authenticator tool, as described in this application, would be a positive resource needed to address the threat of predatory journals.
The target audience for the proposed tool is researchers, particularly ECRs. However, end users may also include funders and research institutions, who may be looking to ensure work they support is published in credible outlets. These stakeholders will also be approached to contribute to knowledge translation activities to help inform the research community of the tool. Such activities will include: a published commentary describing the authenticator tool, press releases, a social media campaign, and outreach via listed collaborators and related listservs.
Influencing open research practices
The proposed journal authenticator tool would impact how researchers publish by supporting them in making responsible OA journal submission decisions. Predatory journals have been known to dupe unsuspecting researchers into submission. One recent study indicated that 5% of scholars seeking academic promotion in Italy had publications on their CV that were from predatory journals. Support to select a OA journal will be of increasing value as the research environment moves towards mandating OA publishing.
The tool would be openly available, and free to implement. This tool would provide researchers with an easy resolution to determine the authenticity of a journal, without creating barriers for use or education needs.

Details of proposal – evaluation plan

Aim 1: develop the journal authenticator
Months 1-8
The research team will first establish the key criteria the authenticator tool should report for each journal.
Subsequently, they will work with internal computer science co-op students, computer and engineering faculty, and external consultants, to facilitate the creation of the tool.
The team will schedule regular conference calls to discuss progress on the development of the authenticator tool during this time, and to address any issues, should they arise.
Aim 2: evaluate the journal authenticator
Months 9-11
Once the tool is developed the core team will do a range of internal testing, and the tool will be modified accordingly. Subsequently, a pilot test group of researchers and related stakeholders will be identified, a protocol written, and a study to obtain feedback and validation of the tool will be conducted. Changes, if needed, will then be integrated into the tool.
Aim 3: disseminate the journal authenticator
Month 12
A knowledge translation plan will be developed by the core team in collaboration with the communication offices of our collaborator groups.
A commentary manuscript will be prepared for publication in an OA journal.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a proposal to create a browser plugin to identify predatory open access journals. The level of innovation, as well as the potential impact of this proposal to transform health research through openness was limited.

Title

Facilitating exposure to quantitative research findings through app-based interaction

Lead Applicant

Dr Laura Skrip

Details of proposal – team members and collaborators

Team members:

Benjamin Althouse (Co-chair of Epidemiology, Institute for Disease Modeling [IDM], Affiliate Assistant Professor, iSchool, University of Washington, Affiliate Faculty, Biology, New Mexico State University) – Contribute expertise in modelling of infectious disease with a focus on human behaviour and facilitate interactions with quantitative science students at the University of Washington

Edward Wenger (Director of Global Health Research, IDM) – Advocate for use of the application in the modelling community and among IDM collaborating field groups in Burkina Faso; contribute modelling expertise and experiences with best practices for open research

Benoît Raybaud (Engineering Manager, Software Team, IDM) – Guide user design process and development of free and open software (app and online interface)

Laura Skrip (Postdoctoral Research Scientist, IDM) – Coordinate across target audiences, leveraging past app development and capacity-building experience in West Africa

Mosoka Fallah (Deputy Director General for Technical Services, National Public Health Institute of Liberia and Director, Public Health Programs, University of Liberia) – Facilitate buy-in and participation from public health students, FETP leadership, and policy makers across West Africa

Olayinka Stephen Ilesanmi (Lecturer of Community Medicine, University of Ibadan) – Facilitate buy-in and participation from public health and clinical medicine students and researchers in Nigeria

Details of proposal – vision, aims and influence on open research

Mathematical modelling is a quantitative research tool that is frequently used to assess infectious disease dynamics and intervention effectiveness in resource-constrained settings. Despite the potential of modelling to inform more impactful public health policy and practice, understanding and adoption of findings are often challenged by how, where, and thus to whom they are communicated, as individuals in resource-constrained settings may have limited access to and/or require guidance in how to interpret scientific outcomes.

The vision for the proposal is to develop new mobile technology to engage public health students, frontline health workers, and other health sector stakeholders in settings with historically limited opportunities for quantitative research capacity building in an iterative process that exposes them to analytical findings and encourages them to offer critical feedback from their own experiences.

Specifically, the project aims to

1 Improve the foundational understanding of how and why mathematical modelling is an important tool for studying infectious diseases among populations;

2 Motivate individual-level use of quantitative information in decision-making about health behaviours; and

3 Create a mechanism for active engagement of local expertise in identifying and filling information gaps in modelling analyses, to improve validity and relevance of such analyses as presented on an open platform

Target audiences will include (1) university students majoring in public health, surveillance officers, and other health sector stakeholders in resource-constrained settings, as well as (2) graduate students and research scientists with a focus in data science or mathematical modelling. By working within university systems (e.g., University of Liberia, University of Ibadan, and University of Washington) and other research- or practice-focused institutions (e.g., IDM and the Field Epidemiology Training Program, respectively), there will be quick access to cohorts of individuals who can shift current practices in the modelling field to not only encourage more

interaction with current and future stakeholders but also build capacity and interest in the methodological process.
Activities:
(Months 1-3) IDM (http://idmod.org/software) will draw on its extensive experience of developing open software (e.g., EMOD), web-based interfaces (e.g., COMPS), and visualization toolkits (e.g., Vis-Tools) to create user personas and storyboards that will be shared in focus groups with individuals in both Target Audience Groups.
(Months 4-5) IDM will develop functional prototypes of the mobile application and a visual interface. Data summarizing user responses and interactions, without any personal identification, will be presented through an interface hosted on the IDM website. Collaborators will work with the application lead to finalize a deployment strategy across target audiences.
(Month 6) A pilot phase will be used to gather information on baseline level of quantitative understanding across different settings (Target Audience 1) and initial pool of modelling assumptions/results (Target Audience 2). The application algorithm will be adaptive in complexity to build interest and capacity at a pace that matches the level of engagement among users. Research scientists will be encouraged to submit modelling assumptions and findings with relevance to the settings involved in the project.
(Months 7-11) Large-scale deployment of the application across collaborating sites in Liberia, Nigeria, and Burkina Faso, with user-driven opportunities for wider spread.
(Months 11-12) Dissemination of information from the open interface to a larger network of public health policy makers and modelling programs.
Applications exist for soliciting ground truth data for analytical or research purposes (e.g., PREMISE). However, simultaneously using app-based interaction to provide feedback and learning opportunities creates mutual benefits to students and frontline workers with contextual expertise and modellers with technical expertise to initiate a shift in how quantitative research findings are generated, communicated, and applied. This vision depends upon and will encourage use of open practices. Via the mobile application, scientific findings will be presented in a way that is easily accessed for not only quicker, but also more effective consumption to facilitate its translation into action. Likewise, the web-based interface will display new information that addresses data gaps and how the incorporation of such information may have changed findings. It is expected that open display of these process results will encourage wider spread buy-in of decision-makers and researchers to motivate systemic change.

Details of proposal – evaluation plan

Focus groups will be conducted with approximately 100 and 25 individuals in the two target audiences, respectively. Feedback on feasibility, usability, and utility of the application (e.g., internet access, language requirements, content/navigation difficulty, data use policies) will be solicited.
During the pilot phase, success will be measured in terms time spent using the app (target: >30 minutes per month), percentage completeness of responses (>50%), and percentage of positive responses to prompts about advancing to new features (>50%). Short quizzes will be used to assess baseline quantitative skills and knowledge of mathematical modelling and its potential.
During scaled deployment, success will be evaluated independently for the two target audiences. For Target Audience 1: The effectiveness of the app in promoting evidence-driven behaviour will be measured as the relative odds of changes in reported health behaviour upon exposure to modelling findings versus exposure to other app features, such as quizzes on quantitative reasoning (target effect: 50% higher odds). The effectiveness of the app in exposing users to the results of quantitative studies and soliciting their critical feedback will be measured in terms of the percentage of all users (number of downloads) who submit at least one comment in reaction to short results from a quantitative research study with relevance to their context. For Target Audience 2: The utility of interactions via the app for enhancing the realism and translation to practice of results will be assessed using short embedded “user satisfaction” surveys on the how

relevant feedback from Target Audience 1 is perceived to be and the likelihood that such feedback will lead to a change in methodology on current and/or future work (>50% responses indicating highly relevant and highly likely). Additional metrics include rate of referrals to the app (>50 monthly) and number of model updates presented on the interface (>25).
Decision Not shortlisted
Comment on decision from Wellcome This was an innovative proposal with potential to impact infectious disease modelling through engaging local expertise. However, the commitment to openness was unclear.

Title

Healthcare for older adults: education, research and practice with safety and sustainability

Lead Applicant

Dr Luciana de Gouvêa Viana

Details of proposal – team members and collaborators

Edgar Nunes de Moraes: Geriatric Physician, PhD, Professor of Medicine, Universidade Federal de Minas Gerais – UFMG. Proposed research role: technical reference for clinical aspects in Geriatrics and Gerontology, including the development of clinical protocols and guidelines applied in the elderly population.

Carla Jorge Machado: Economist and Demographer, PhD, Professor of Medicine, Universidade Federal de Minas Gerais – UFMG. Proposed research role: technical reference for epidemiology and statistics, technical reference for epidemiology and statistics, including the validation of the frailty assessment instruments applied in the elderly population.

Letícia Maria de Henriques Resende: Clinical Pathologist, Master, Professor of Medicine, Universidade Federal de Minas Gerais – UFMG. Proposed research role: technical reference for laboratory tests in Geriatrics and Gerontology, including the development of frailty assessment instruments using laboratory tests results and guidelines applied in the elderly population.

André Aguiar Souza Furtado de Toledo: Geriatric Physician, Hospital das Clínicas da UFMG. Proposed research role: technical reference for clinical aspects in Geriatrics and Gerontology, including the development of clinical protocols and guidelines applied in the elderly population.

Empreendimentos Digitais/MedLogic: Software company specialized on elderly healthcare and member of a Consortium supported by a Newton Fund Grant to promote elderly's wellbeing.Proposed research role: App Development/Digital Marketing.

Details of proposal – vision, aims and influence on open research

Our vision: Brazil has been experiencing one of the fastest demographic aging worldwide. Many other Latin America nations will experiment the same transition level in the next three decades. This demographic transition is occurring in a context of few resources and great social inequalities. In most countries, the elderly healthcare is already a challenge, implicating the need for investments in research and education to yield efficient and sustainable practices focusing on this population. Scientific and empirical evidence suggest that integrated health and social care for older people contributes to better health outcomes at a cost equivalent to usual care, with higher levels of return on public and private investments. Additionally, the proportion of physicians in Brazil is composed by one geriatrician to each 22,000 elderlies, when the WHO recommendation is a 1/1000 ratio. And this relation is even worse in most of the 5,570 Brazilian municipalities that have not a specialist access, at all. However, most of these cities, even the small ones, have Internet access. The main goal of this project is to consolidate high quality information, to balance the education deficit, empower the professionals in direct contact with elderlies, especially those that are frail and from low income families and inspiring researchers to generate qualified results related to the healthcare of the elderly people for application in public health. On the other hand, the practice of Evidence-based Medicine can help in building the best possible strategies for approaching the older adults, not only at the level of geriatrics and gerontology, but especially at the level of primary care physicians. The access to qualified scientific information, clinical protocols based on the best evidence and validated instruments for clinical and laboratory evaluation are key elements for healthcare safety and efficiency.

Proposal Aims: The aim of our proposal is to develop a mobile software solution (App) with free download, integrating platforms for general information for society, research, education and healthcare for elderly people. This idea was born from the public health experience of proponents, particularly in aging, and from the perception of the difficulty of health professionals to qualify and manage simple and practical assistance instruments to the elderly patient in their daily practice.

Target Audiences: The application will target health professionals, scientists, students and patients/society. It will be developed in Portuguese, Spanish and English. Thus, their penetration into various nations will be facilitated. Each platform will have tabs corresponding to the contents and functionalities. Regarding the research, it will be included links to databases of scientific publications and papers selected by the technical references, especially those with high level of evidence. There will also be access to databases of groups and research projects registered in the funding agencies. In the education tab, there will be a virtual library with books, classes, courses, case studies and events calendar. On the practice tab, there will be guidelines, protocols and evaluation nstruments. Health promotion education for older people will be the focus of general information tab. In addition, there will be an open channel for communication with App managers. The App and all content will be in compliance with the GDPR (General Data Protection Regulation) and the Brazilian LGPD (General Data Protection Law). After the final version is finished, it will be promoted in events and by using digital market strategy.
Activities: Compile the international recognized studies already available at our study group to the App format, research all other relevant and specific to our target worldwide studies, books and open courses, identify partnerships that can update related events on regular basis, develop an App compatible with iOS and Android in compliance with GDPR and LGPD, test and refine the App with pilot groups, create a digital marketing and dissemination strategy to promote the App. How our proposal will influence open research practices in your field or more broadly: This project will contribute to scientific knowledge dissemination and innovative practices in the healthcare for older people. Access to qualified information will made the engagement of health professionals in best practices easier, contributing to the health systems efficiency and sustainability. On the other hand, students will find ways to speed and increase their knowledge acquisition and training.

Details of proposal – evaluation plan

The number of App users will be continuously monitored as well as accesses in each features and functionality. Comments on the communication channel with administrators will also be monitored and analysed. Satisfaction surveys will be applied to App users including questions about performance and response time; features and functionality; reliability; appearance, usability and navigation; safety and quality of information. Focus groups will be applied in a virtual environment, called online focus groups, in order to highlight factors for updating the application.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was to create a database of published research. The level of innovation was considered limited

Title

Standarised Detailed Hierarchical XML data-file and WebGEOMap JavaScript Leaflet and OpenLayers plugins.

Lead Applicant

Dr Maksym Bondarenko

Details of proposal – team members and collaborators

Dr. Bondarenko Maksym - University of Southampton
Kerr David - University of Southampton
Ves Nikolaos - University of Southampton

The prime investigator of this project will be Dr. Maksym Bondarenko. The PIs experience in multi-disciplinary projects, including, engineering, physics, numerical simulations, computer science and web-based systems makes him very suitable to run and develop this project. This is an interdisciplinary project: merging of high level skills from the computing side (PI) and the Geography and Environmental Science expertise of his collaborators from the WorldPop group. David. K and Ves Nikolaos have a significant amount of experience in web-based collaboration systems and computational numerical analysis of complex systems. This experience in creating web-based centralised processing and database systems put the David K. and Nikolaos V. in a very good position to develop these WebGEOMap JavaScript Leaflet and OpenLayers plugins. Bondarenko. M will be responsible to lead the project and develop a new Standarised Detailed Hierarchical XML data-file (SDHXML).

Details of proposal – vision, aims and influence on open research

The use of high-resolution geospatial datasets has increased markedly over recent years, with the detail/resolution of datasets rising alongside the variety of datasets accessible. The availability and use of these datasets in demographic and population health field is no exception, as governments, funders, academics and other stakeholders strive to help low and middle-income countries achieve the United Nations' Sustainable Development Goals from different perspectives. These geospatial data help decision-makers and researchers focus their analyses on different populations within countries to ensure the most vulnerable and isolated are highlighted when planning interventions. Whilst researchers are invariably adept at understanding and making use of geospatial data, there is still a need to disseminate their discoveries in a manner in which stakeholders not accustomed to geospatial data can easily understand the findings and be able to make effective decisions quickly regarding these outputs. One contemporary method in which information extracted from analysis can be disseminated effectively to large audiences in a cost-effective manner is with web tools and portals. Web-GIS (Web-based Geospatial Information Systems) is a subcategory of this method, whereby geospatial data can be visualised on a web map, providing interactivity to allow users to carry out simple analyses, download data or generate reports. Although a range of JavaScript libraries provides web-developers with the tools to carry out these functions, there is still a need for them to undergo some training in front-end development and design in order to produce professional web-maps. In addition to the visualisation of the maps, the developers are also required to develop data structures to ensure that the data and functionality always performs as intended. The development of a web-map can typically take a considerable amount of time, and whilst most of these libraries are free and open source, considerable cost can be introduced when considering these complications. This time and cost can dissuade some bodies from providing these effective visualisations, thus limiting the scope and efficacy of their findings.

Leaflet and OpenLayers are two popular open-source JavaScript web-mapping libraries that facilitate a range of geospatial functionality and visualisations. Being open-source, the frameworks allow the development of customised plugins to be developed and hosted on their respective repositories. The WebGEOMap Leaflet and OpenLayers plugins propose to alleviate some of the time and financial burden on data providers/analysers by enabling those with limited web-development skills to access, visualise and disseminate a variety of geospatial data held on data

providers' repositories. By simply downloading a choice of plugin and including it in their web pages, users are able to link their applications to a detailed hierarchical file held on the data providers' servers. The hierarchical file will allow data providers to specify paths and options that can be linked to developers' plugins, allowing them to customise their applications with minimal configuration. The link between the user and the configuration file will be in the form of a URL that the user includes in their JavaScript file, specifying the options required for their particular application. To the best of our knowledge, this is a novel approach in linking data and configurations between data providers, developers and front-end users. The advantage of such a framework will allow a uniform method for easily implementing an application or dashboard linked to different data providers. An example output of the customisation could be a choropleth map that shows a metric specific to a country, summarised at the subnational level. Upon clicking on the subnational polygons, various additional information can be displayed in a pop-up window, along with the option to download the data from the data-provider repository in table or geospatial grid format. WorldPop SDI team hope to consult potential users/stakeholders in low/middle-income countries to provide further relevant functionality for the tool. It is hoped that the tool will help to increase the use of geospatial data in different settings, whilst promoting the benefit and reach of valuable datasets, in not only health/demographic settings, but all environments in which spatial elements play an important role, helping to improve the quality of life of citizens, and increase transparency and accountability. The plugin will be hosted on the repositories with full documentation of how to use it and a newly developed Standarised Detailed Hierarchical XML data-file.

Details of proposal – evaluation plan

Three main outcomes are expected from the development of such a tool/standard. The development of a new standard presented in the Standarised Detailed Hierarchical XML data-file (SDHXML) to store the information will potentially contribute towards global academic advancement, building a worldwide community and could lead into other developments such as R/Python packages and different plugins GIS software. In addition, JavaScript plugins for the Leaflet and Openlayers mapping libraries, which can be re-used in multiple web applications. To compliment these plugins, an API will be developed to aid developers to query the SDHXML and its associated data. Finally, upon the completion to the project, the framework will be detailed in a computer science journal publication.

The development of standards, hierarchical framework and software to be held on a GitHub repository, a platform that allows version control of software in addition to the option to submit requests to make amendments to the software by the software's users. We aim for collaborative development from the start, with regular interactions among collaborators to coordinate and prioritise activities. It is hoped to debut the final Beta version of the software to participants of the WorldPop winter school workshop, to be held in 2020, giving us the opportunity to gain valuable feedback from potential users of the software, and make the necessary changes to ensure the framework's success over time. 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

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title **From academic publications to medical practice: translating data driven risk calculations into actionable medical information**
Lead Applicant **Dr Mercedes Bunz**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

SPIN: a pervasive problem in medical science

Lead Applicant

Dr Naichan Su

Details of proposal – team members and collaborators

Prof. dr. Geert J.M.G van der Heijden, Chair and professor in Social Dentistry, Department of Social Dentistry, AcademicCentre for Dentistry Amsterdam (ACTA),University of Amsterdam, The Netherlands.

Prof. dr. Geert van der Heijden will oversee and guide the project team during all phases of thisproject.Prof.dr. Lex M. Bouter, Chair and professor in Methodology and Integrity, Department of Epidemiology and Biostatistics, VU Universiteit Medical Centre (VUmc), Amsterdam, The Netherlands.Prof. dr. Lex Bouter will co-guide and oversee thisproject in all phases of this project and in particular contribute to the development of the spin checklist, and make contributions to the development of the concept of spin, and the development and refinement of the checklist.

Prof. dr. John P.A. Ioannidis, the C.F. Rehnborg Chair and professor in Disease Preventionand professor in Health Research and Policy (Epidemiology) and in Statistics, Stanford Prevention Research Center, Department of Medicine, Stanford University, Stanford, California, USA.Prof. dr. John Ioannidis will contributewith his vast expertise on the utility of methodologies in medical research and their importance in terms of the science and practiceof medicine and healthcare.

Details of proposal – vision, aims and influence on open research

Background-Spin is pervasive in society, from intentionally providing distorted information (e.g. fake news) to unintentional misinterpreting data. Spin can be defined as a propagandizing or deceiving practice that distorts the meaning of information, and results in misleading interpretation and conclusions. In science, sharing and exchanging information from research is essential in the transfer of knowledge. In the dynamics of science, spin can be a key factor in scientific communication that may introduce bias in all phases of the scientific enterprise, from framing study questions to concluding on study findings. Nowadays, references and citations are used to convey the meaning of information, and spin, for example, may be related to selective citation resulting in biased research. As a result of spin, distorted information is conveyed, and this results in bias.

There is evidence that spin is manifest in the science and practiceof medicine and healthcare. It may have biased scientific communication and the transfer of knowledge to a larger extent than currently is acknowledged. It has been reported that spin is manifest in 57% of the published clinical trials.

Spin leads to a cascade of inflated and questionable evidence in the literature and then leads to skewed systematic reviews and misinformed clinical practice guidelines or health policies. The mechanism behind spin reinforces by publication bias, selective outcome reporting bias, and citation bias. Such biases may go unnoticed for policymakers and clinicians during their decision-making. Thereby spin has the potential to negatively impact population health and reduce the efficiency of medicine and healthcare. Spin also contributes to the reduced reproducibility of research, andmay thereby slow down the progress of scienceand reduce the return-of-investment of research funding. As such, spin is accountable for a huge waste of societal resources. Therefore, spin is hazardous to almost all the stakeholders in medicine and healthcare, notably, patients, professionals, policy-makers, researchers, funding organizations and medical companies.

Aims and activities-Our ambition is to ensure accurate interpretationand disseminationof medical and healthcare researchin order increase the return-of-investment(ROI) and to improve the value and outcomes of medicine and healthcare. To increase the reliability, transparency, and accuracy of translation, dissemination, and implementation of knowledge and evidence from medical and health research to practice, this project addresses the challenge of spin in the research and practiceof medicine and healthcare. In this we aim at the identification of spin as a conceptby,first,increase the awareness among practitioners, researchers, policymakers, editors

and peer reviewersof the importanceandwidespread prevalence of spin;second, develop methodologies in order to avoid and reduceits occurrence; third, establish approaches accordingly,to supportchanges in scientific communication, and finally, implement these in educational curricula.
To achieve this we seek to:develop an innovative checklist for identification of spin; evaluate and test the performance of the checklist; describe the system architecture for aweb-based service for identificationof spin.
Target audience-By addressing the above challenges, and fulfilling the objectives, this project generates potential breakthroughs for science and society, and for medical and healthcare practice. All stakeholders in medicine and healthcare, notably, patients, professionals, policy-makers, researchers, funding organizations and medical companies, will be able to benefit.
Impact on medical science-Thisprojectcan improve the stakeholders` awareness of spin in medical research.Thisprojecthelps the stakeholderseasily identify spinwhen reading literature and writing papers so thatthe translation, dissemination, and implementation of knowledge and evidence from medical and health research to practicecan be more reliable, transparent and accurate. The data and results will be shared at the platform of the Center for Open Science (https://cos.io/).
Broader impact-Thisproject may improvethe reliability and reproducibility of research and acceleratethe progress of scienceThereby, this researchmay ultimatelyimprove the value and outcomes of medicine and healthcareand potentially boost its ROI. Therefore, this research may increasethe health gains of the population and the Gross Domestic Product gains of a nationincrease.

Details of proposal – evaluation plan

We seek to deliver on the following project objectives:
1.Development of a checklist for Identification of spinBuilding on prior work by Chiu, et al. and further evidence scoping, a longlist of potential checklist itemswill be compiled. For selecting pertinent items researchersin the field will be invited for a Delphi study. Shortlisted items will be discussed in aconsensus meeting with key experts. All the selected items will be elaborated and refined. This results in a checklist manual, including explanatory notes and examples for each item. To establishtheproof of concept,the checklist will be piloted for its validity and feasibility. To elaborate the rationaleand background, an explanatory commentary will be drafted. The checklist, the manual and accompanying explanatory commentarywill be based on consensus among the key experts andwill be published in an open access scientific journal.
2.Proof of principle evaluation of the checklist To explore the performance of the checklist, it will be evaluated in randomized controlled trials (RCTs) in three societal important fields of medicine and healthcare.Based upon this evaluation, recommendations on spin preventionwill originate to aid editors, reviewers, researchers, healthcare professionals, educators and healthcare policy makers.
3.Describing the system architecture for aweb-based service for spin identificationThe checklist will provide the framework for identification of spin in medical publications. Based on this framework, we will describe a system architecture for automaticprocessing of spin identification, includinga working methodology for developing a prototype of a graphical interface and web-based service. Based on this system architecture, a source code for a graphical interface for automated and technology assisted identificationof spin in medical scientific publications can then be prototyped for aweb-based service.(Please note: prototyping is beyond this proposal)

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting and potentially impactful proposal aiming to identify spin in medical research publications. However, the level of innovation proposed was considered limited.

Title

Meta-data inventory

Lead Applicant

Dr Natalia Dutra

Details of proposal – team members and collaborators

Gonzalez-Marquez, Monica. Uniklinik RWTH - Aachen University, Germany, Department of Neurology, Editorial Assistant at the Journal of Neurochemisty. Project manager and idea originator. Will coordinate integration of all aspects of the project as well as communication between the various stakeholders.

Dutra, Natalia B. Psychology Department, Durham University, United Kingdom. Training of Brazilian editors in open science practices and support with the implementation of the meta-data inventory in the journal Estudos em Psicologia. Advertising the project in Brazil, and training Brazilian scientists to provide input and use the inventory when submitting their papers.

Tsuji, Sho. Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure. Project consultant. Creation of open access analysis and visualisation platform, expertise in meta-analysis, acquiring data for meta-analysis, meta-analysis trainings.

Christina Bergmann. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Supervision of the team and progress monitoring. Creation of open access analysis and visualisation platform.

Philipp Brauner, RWTH - Aachen University, Germany, Human-Computer Interaction Center. Will oversee and advise on the human-computer interaction aspects of the project.

Details of proposal – vision, aims and influence on open research

Imagine being a medical researcher needing information about leukemia studies run on 6-9 year-old girls from 1980-2000 anywhere in the world and potentially published in multiple languages. Or a researcher doing a meta-analysis on short-term memory effects after marijuana use in computer programmers. At present, studies like these require a painstaking search of the literature, followed by the careful examination of each possible paper in search of the relevant data. This process is known to be error-prone, whether done manually or using machine-learning, often taking months, only to yield substandard results largely because documentation of studies is inconsistent and incomplete. Now imagine access to a database where all this information was accurate, complete and available in a matter of minutes. What we propose is the Meta-Data Inventory (MDI), a bibtex-like inventory form intended to data-ify the entire contents of a science article, as well as contain information necessary to evaluate scientific progress that is not typically included in the published literature. This tool, though simple in principle, would facilitate the eventual creation of a world-wide, searchable, semi-language independent database containing extensive internal details about research.

What sets the MDI apart from other efforts to gather meta-data is that it will:

first and foremost focus on building the human infrastructure. We've long had the technology to gather meta-data, and yet its quality has not improved (see Page & Moher, 2017). To succeed, this project will focus on involving a variety of stakeholders (authors, journal editors, etc.) from diverse scientific and demographic backgrounds (international and non-native English speakers). They will be part of the decision process for content and context, as well as the development of tailored training.

gather meta-data directly from authors as a condition of publication, as a pre-emptive measure. This exploits the expertise and unique knowledge of those who conducted the research, and imposes minimal costs on the research community.

gather information to provide a more complete picture about the production and content of research. Namely:

a. Labor and attribution - currently largely opaque, with even authorship somewhat arbitrarily determined and roles such as data collection, statistical consulting and manuscript editing rarely included in the scientific record.

b. Classification of research - currently dependent on an inconsistent keyword system.
c. Evaluation of science practice - including types of analyses, participant characteristics, research technology, code, open access status, etc.
begin as an international, multilingual effort (4 countries in 3 continents).
Implementation will involve partnering with three journals (two confirmed: Cognitive Linguistics (UK) and Estudos em Psicologia (Brazil), two pending outlets: a neuroscience journal (Germany)) and a pre-print archive (USA) to develop the MDI according to their specifications and language (overlap is expected). This will be done by developing the lists of items to be indexed taking into account the CRediT Taxonomy (https://casrai.org/credit/) and the APA JARS guidelines (http://www.apastyle.org/manual/related/JARS-MARS.pdf) in collaboration with guidance from meta-science experts and the journals. Then developing the computational form where the information will be input by the authors using guidance from our human-computer interaction expert. All partner outlets will agree to the MDI as a publication prerequisite.
Following these guidelines will create a product on two scales. At the micro scale each paper will be accompanied by its own machine-readable inventory.
On the macro level, the MDI will be directly fed into an open-access analysis and visualization platform, building on MetaLab (www.metalab.stanford.edu), which currently provides infrastructure for meta-analytic data. The data gathered will be accessible on an online platform that allows easy data access as well as basic analysis and visualization functions. Compared to extant meta-scientific studies and meta-analyses, this tool makes it possible to collect meta-data at an unprecedented low cost and high accuracy.
This project will create awareness of and implement science as a massively collective enterprise that must be internally transparent. It will provide the necessary data for continuous evaluation of scientific progress, in faster, more effective, and more transparent ways. The MDI will also affect how scientists think about essential aspects of their research and how they must report them, by incentivizing more accurate, careful description. Finally, MDI will greatly impact diversity through development and implementation across different countries and languages.

Details of proposal – evaluation plan

The success of the project will be determined by ease of use of the MDI in the context of the publication process, and the quality of the data produced. Because the human component is the focal point of the project, monitoring and evaluation of human interaction with the MDI will be continuous throughout the project. The MDI must be as effortless as possible. Our evaluation cycle will consist of:
Meetings to determine content or format;
Implementation;
mock experiments to evaluate effectiveness and ease, including surveying participants for level of 'annoyance' and time consumption;
Modification incorporating findings from mock experiments;
Repeat, adding components until the final product is complete.
This process will be supported by a human/computer interaction consultant. The foreign language/cultural components will involve back-translation and research into the most apt terminology, as well implementing the cycle to ensure efficacy. As to the quality of the meta-data, our data will be compared to typical data scrapped for meta-analyses and systematic reviews, and assessed for accuracy and completeness. If these requirements are satisfied, we will consider the project a success and will seek additional funding to expand the project to other journals and archives.
There is a risk that the MDI will fail to reduce author labor substantially, that data quality will remain low or that attempting to implement the MDI at four venues over one year will prove too much work. The fact remains that the project should be attempted as a first step toward standardization of empirical data across disciplines and languages. This type of standardization is

long overdue in science, and knowledge development is suffering for its lack. Even if the project were to fail, we will have learned a great deal about the implemention of such a system.
Decision Not shortlisted
Comment on decision from Wellcome This was an interesting proposal with clear potential impact. However, its success was dependent on securing high-levels of community support, and it was not clear how this would be achieved.

Title **Networked, smart micropublications to rapidly publish clinical case studies, incremental progress, and other research outputs on rare cancers and other clinical domains**
Lead Applicant **Dr Nate Jacobs**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

Open Synthesis: ensuring that systematic reviews are verifiable, repeatable and reusable

Lead Applicant

Dr Neal Haddaway

Details of proposal – team members and collaborators

Neal Haddaway, Stockholm Environment Institute – Neal will lead the project, coordinating the group's activities, coordinating the hackathon, and leading the drafting of working papers. Neal will also represent the Collaboration for Environmental Evidence, an organisation that publishes Open Access environmental systematic reviews and is interested in making these reviews more Open.

Tamara Lotfi, Global Evidence Synthesis Initiative, American University of Beirut – Tamara will co-convene the group with Neal, making use of her connections across the GESI Network and its partners.

Vivian Welch, Campbell Collaboration – Vivian will represent the Campbell Collaboration, an organisation dedicated to publishing Open Access systematic reviews in the social sciences, and interested in implementing Open Synthesis strategies developed by this group.

Jordi Pardo Pardo, Cochrane - Jordi will represent Cochrane, which publishes systematic reviews in health and is interested in making its reviews more Open by implementing strategies developed by this group.

Adam Dunn, Macquarie University – Adam will advise on informatics in relation to evidence synthesis.

James Thomas, University College London – will advise on systematic review management software.

Martin Westgate, Australian National University – Martin will support the coordination of the hackathon.

Elie Akl, American University of Beirut – Eli will represent the Living Evidence Network.

Details of proposal – vision, aims and influence on open research

SRs are vital for rigorous evidence-based policy and provide essential feedback to improve underlying primary research. All SRs generate huge quantities of data in addition to the published review – including lists of relevant articles, information on how articles are relevant, and their key findings – yet these data are almost never released. In rare cases where data are available, they are not standardised. These practices stifle replication and updating of SRs that is necessary to reduce research waste and ensure SRs are updated.

Aims: We will establish a community of practice (CoP) to define pathways to achieve Open Synthesis (i.e. Open Science principles applied to systematic reviews (SRs)), with the overall aim of facilitating the verification, reuse, efficiency optimisation and automation of SRs through the application of FAIR (Findable, Accessible, Interoperable, and Reusable) principles to SR data. Open Synthesis aims to maximise openness and reusability of SRs, reducing waste from repeating tasks already conducted by other researchers. Open Science is not new, but its application to SRs, to date, been limited.

Target audience: The project will build a CoP of methodologists and software developers interested in exploring the mechanisms of Open Synthesis. It will produce recommendations and guidance for the broader community who produce SRs (commissioners, funders, authors, editors, and publishers).

The project will rely on stakeholder engagement: the CoP will be involved throughout to co-design the project and its outputs. In particular, the hackathon (see point 3, below) will be a highly interactive, co-production event involving diverse stakeholders. Furthermore, although the primary aim of this project is to benefit healthcare SRs, we will engage with stakeholder across disciplines that use SRs (e.g. software engineering) to benefit from their knowledge and tools.

Activities: The CoP (that will include review management software developers, e.g. EPPI-Reviewer) will aim to facilitate Open Data/Methods adoption by defining data structure, types,

storage, and minimum requirements. Since most reviewers use these tools, substantial gains in Openness can be made by agreeing on standard interoperable file formats (e.g. for citation screening decisions and flow diagram information). This interoperability will allow the content of reviews to be reused in part or in full rapidly, without the need for human data reformatting or repetition of work.

1. We will assemble an Open Synthesis Working Group of leading experts in SRs across disciplines (but heavily focused on healthcare) to produce a definition of Open Synthesis that is widely accepted by different stakeholders. This process will involve a discussion of 'how Open is enough?', 'which processes can be Open?', and 'how can we move towards Openness, and 'how can we develop incentives and tools to support this transition and overcome any barriers' (e.g. giving credit for Openness, citation credit for data publishing, technical tools to facilitate Openness)?
2. We will develop suggestions of actions needed to attain Open Synthesis, including the production of standard data structures, data types and minimum required for the outputs of systematic reviews (e.g. lists of included studies with extracted data).
3. Finally, we will convene a highly interactive workshop in the style of the successful Evidence Synthesis Hackathon event series (www.eshackathon.org) to facilitate these discussions and develop tools to support Openness. We propose to host this workshop alongside the Cochrane Colloquium in Toronto in 2020. Such tools may include technology to transparently record web-based grey literature searches, for example.

Impact on Open practices: c. 3,000 SRs are published yearly, each requiring screening of c. 3,000 records, equating to 75m records examined annually, yet this screening information is not shared. This project will make better use of these efforts by supporting SR authors and publishers (e.g. Cochrane, Campbell, CEE) with guidance, tools and incentives to be more Open, facilitating the process of updating and reuse, and increasing efficiency. Furthermore, verification and replication will strengthen the accountability and robustness of SRs to external criticism. The project has a high likelihood of impact, since we have support from 4 global organisations (GESI, Cochrane, Campbell, CEE) with interest in and commitment to the project's aims. Their involvement will allow the outputs to be efficiently and effectively integrated into their workflows, rapidly changing practices of these CoPs (Cochrane has >11,000 members and >67,000 supporters, for example).

Details of proposal – evaluation plan

This project will require a range of metrics to monitor and evaluate success.

-Network metrics: membership, satisfaction with process, progress towards agreed upon activities

-Tools for Open Synthesis metrics: number of tools developed at hackathon, number of accesses/downloads of those tools in the year following the event, number people engaged in hackathon and their disciplinary background, methods for recognising and rewarding Open Synthesis practices

-Communication, outreach and engagement: metrics for impact of blogs, commentaries, people engaged

We will assess the general acceptance of the suggested practices produced within this project across the broader stakeholder group, and we will also appraise the reception in the broader community: we will do this using a combination of online surveys and key informant interviews. Feedback from this process will allow improvement of outputs and mechanisms for communication of the project's outputs.

We will also trial the recommendations and tools produced within the hackathon on a small number of cases study systematic reviews to understand what factors facilitate or inhibit successful implementation in real world examples. Again, feedback will allow improvement of the project outputs. In addition, we will use focus group discussions as part of ongoing training workshops provided by Cochrane, Campbell and CEE to better understand users' perceptions of the requirements and recommendations related to Open Synthesis. Since the framing of reporting

standards is instrumental in affecting uptake and success, we will particularly use these modes of feedback to tailor terminology and framing of communication media that aim to raise awareness of the project's outputs. Finally, the outputs of the project will be hosted on a dedicated cross-disciplinary website that provide explanations and tools to support Openness in evidence synthesis. We will monitor access to this website to improve the platform and increase impact.
Decision Shortlisted, not funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Online meta-analysis engine for electric brain stimulation

Lead Applicant

Dr Nick Davis

Details of proposal – team members and collaborators

Dr Nicholas Holmes, University of Nottingham. Dr Holmes' role will be to provide assistance and oversight in the aggregation and analysis of brain imaging and stimulation data, and in the creation and maintenance of the online resources.

Details of proposal – vision, aims and influence on open research

Electric brain stimulation (transcranial electric stimulation, tES) is a tool for non-invasively modulating the activity of the brain. tES works by passing a weak electric current between two electrodes on the head, which alters brain function. It is commonly used in research environments, and has been suggested as a treatment for a wide range of brain disorders, including depression, stroke and pain, and may even enhance cognitive performance in healthy younger and older people.

Some scientists have raised concerns about the safety of tES, with possible adverse effects including skin burns, changes in mood and cognition, and in extreme cases seizure. At present there is no central repository for collecting information about adverse effects of tDCS, except through "Letters to the Editor" in journals such as Brain Stimulation. Lab-based studies of the effect of tES are typically limited by their small sample size and by the lack of follow-up. For example, the NICE overview on the use of tES in depression (https://www.nice.org.uk/guidance/ipg530) notes that currently the largest sample size in any study of tES and depression is 60. As well as the risk of adverse effects, there is also a risk of missing beneficial effects of stimulation, either through under-powered studies or through incomplete understanding of the mechanisms of tES in the brain. This proposal is part of a wider programme to understand the sources of variance in brain stimulation studies, in order to improve safety and efficacy.

We will address this challenge by creating an online 'electric modelling' service for tES studies. The modulation of brain activity by tES relates directly to the electric field induced on the brain surface by the current passing between the electrodes. Currently the best way to target stimulation at a specific location on the brain is to simulate this electric field with a computational model. Since each person's brain differs in size, shape, folding and depth from the scalp, it is necessary to produce a model for each person, for each electrode set-up. Our online resource will create person-specific electric field models, as well as 'average' models for a sample of study participants.

The project will appeal to two main groups of users: scientists and home users. Scientists are people actively engaged in running brain stimulation studies with tES, who have collected data from participants. Scientists want to know how their stimulation protocol affects their participants' brains. With the proposed tool, scientists will upload 3D MRI brain images, plus information about the tES protocol used (e.g. electrode size and location, stimulation intensity), plus any observations such as adverse effects of stimulation. The tool will create an image, for each participant, of the electric field induced by the stimulation, plus an 'average' image for the study cohort as a whole. This average image will be added to an open repository, and assigned a stable identifier (DOI) to encourage sharing of images. The second group of users will be those who use (or want to use) a direct-to-consumer tES product a home. These users will be able to understand the effect of stimulating a particular target brain area, they can see any adverse effects reported by other users, and can contribute their own reports.

The major step towards openness in the field of brain stimulation will be the opportunity to accumulate and share images of the electric field induced by stimulation. Similar movements have occurred in the behavioural sciences (e.g. https://osf.io/ezcuj/wiki/home/) and in the brain imaging field (e.g. https://www.openfmri.org/, https://neurovault.org/), and have led to the

integration of user-contributed data with online meta-analysis engines (e.g. http://neurosynth.org/). This latter function means that a user can see which brain areas are active in association with a particular mental state; for example, Neurosynth can build a brain image corresponding to the aggregated findings of 449 studies of 'anxiety': http://neurosynth.org/analyses/terms/anxiety/. The goal of the project will be to provide a central repository for analyses of the effect of tES on the brain, and an ability to relate these effects to changes in behaviour induced by stimulation.
This proposed project will encourage scientists and home users to think about the effect of tES on an individual person's brain, and will encourage sharing of this information publicly and in academic contexts.

Details of proposal – evaluation plan

Popularity of the service: We will measure the number of page impressions received by the server as an index of the interest in the project, aiming for 1000 page impressions in the first month after release. We will also track social media mentions to understand how awareness of the project is disseminating in the neuroscience and brain stimulation communities.
Usage of the service: We will measure the number of electric field images created by the service as an index of usage. We will aim for 60 images in the first three months after release.
Academic impact: We will release two manuscripts that relate to the project. The first (MS1) will describe the website and its benefits, and will be released in preprint form on bioarxiv.org simultaneously with submission to an open-access journal. We will track the metrics provided by the preprint server, aiming for 100 reads (HTML or PDF) in the first three months. A second manuscript (MS2) will summarise the findings of the project after six months, and will describe the meta-analyses of electric field in different common tES protocols. We will submit MS2 to a journal that offers an open-access option, such as Brain Stimulation, and again will upload MS2 to a preprint server.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to create an online tool for transcranial electrial stimulation research. The methodology was not clearly described and the potential impact of this proposal to transform health research through openness was limited.

Title

SSVEPLAB: an open source toolbox for frequency-tagging electroencephalography

Lead Applicant

Dr Nika Adamian

Details of proposal – team members and collaborators

Dr. Nika Adamian, Postdoctoral Researcher, University of Aberdeen
Dr. Adamian is a postdoctoral researcher with experience in EEG data analysis. She will be responsible for package development, documentation and general code maintenance. She has been using SSVEP to study attention since 2017. She is currently leading a meta-analysis of multiple SSVEP datasets and develops novel visualisation techniques.

Dr. Søren Andersen, Senior Lecturer, University of Aberdeen
Dr. Andersen has extensive experience in EEG and SSVEP research. He will provide the core functions of the package which have been developed and tested within his lab. He will also host a workshop to present the tool to the community and receive feedback from the users.
Dr. Andersen has worked in this field for over 15 years and published over 25 papers using SSVEPs in leading journals. He is also responsible for teaching programming and MATLAB at undergraduate and postgraduate level.

Details of proposal – vision, aims and influence on open research

We propose to develop on open source toolbox for analysing electroencephalographic steady-state visual evoked potential (SSVEP) data whose design features inherently facilitate open science practices. SSVEPs are oscillatory brain responses driven by a flickering stimulus. They have the same frequency as the driving stimulus and can thus be used to separately measure the cortical processing of multiple stimuli flickering at different frequencies ('frequency-tagging'). In recent years, SSVEPs have proven a powerful tool in cognitive neuroscience (e.g. attention), visual perception (e.g. colour vision, face perception), and the development of brain computer interfaces (BCIs), with SSVEP-based BCIs being the dominant type. Despite this, the use of SSVEPs in basic and applied research is still restricted because there is no commonly available software solution for the analysis of SSVEP data, which requires specialised time-frequency analyses. Thus researchers in the field rely on lab-specific custom software which severely limits transparency and comparability between results in different labs and the ability to share datasets.
Our toolbox – SSVEPLAB – would provide a standard in the analysis and reporting of SSVEP data and thus go a long way in opening the field to more researchers and enhance comparability between data collected in different laboratories. It would go beyond that by fundamentally including functionality that facilitates data sharing and good reporting practices. Our approach builds on the core principle of object-oriented programming – a strong integration of the data and the operations applied to it. While analysing the data, SSVEPLAB would automatically store the steps and parameters of the analysis from data pre-processing to computation of figures and potentially statistical analysis. These analysis steps would not be stored simply as a text log, but as executable commands. Thus a researcher could share experimental data along with the autogenerated analysis log-file, allowing any other researcher with SSVEPLAB installed to rerun the entire analysis up to and including the figures of the published work.
SSVEPLAB will be based on the EEGLAB toolbox for MATLAB, which is also open source and de facto the standard for analysis of EEG data. This will allow us to build on a wide range of general purpose routines (such as importing data from different EEG systems) and focus our efforts on developing the specific SSVEPLAB functionality. SSVEPLAB will also provide a user-friendly interface for converting data into BIDS (Brain Imaging Data Structure) format, ready for sharing in open databases with comprehensive SSVEP-specific metadata.
The aims of the project are the following: 1) provide the SSVEP community with analysis software implementing most popular approaches to data treatment and novel visualization techniques; 2)

improve the quality of data sharing and methodology reporting in SSVEP studies; 3) make SSVEP technique more accessible to researchers with limited access to training.
The first two aims will be achieved by the core functionality of SSVEPLAB. We propose to use autogenerated log files linked to the analysis software instead of relying on the researcher to provide readable code, which may not always be possible due to complexity of procedures, lack of expertise or incentives. This approach will not only facilitate sharing of data and materials but will also improve reusability of published datasets by providing high-quality automatically generated metadata.
The third aim will be achieved through openly available code and teaching materials. In the final stage of the project we will organize a workshop to introduce the fully functional toolbox to the community, discuss the methodological approaches to SSVEP research and plan further development of the toolbox. We will also provide online tutorials on SSVEPLAB. The existence of the analysis toolbox which is open and specific to SSVEP is especially important to researchers who are new to the technique and are looking for good practices, for example to those in clinical settings.
Overall, SSVEPLAB is not only a much-needed set of tools, but also an opportunity to construct analysis software with data sharing and good reporting standards in mind. If this approach proves to be effective at the level of SSVEP studies, it will be feasible to extend it to broader EEG and other neuroscience data in the future.

Details of proposal – evaluation plan

We will be able to assess the interest in and later the uptake of the toolbox throughout the project. The ultimate goal is to ensure the project is adopted by the SSVEP community.

The toolbox itself will be openly released on GitHub after 6 months of initial development. The release will be announced on EEG-related mailing lists and social media. We will track downloads of the toolbox and views of online tutorials throughout the remainder of the project.

Beta-testers (10-15 labs) will be recruited among past and current collaborators of the lab. Beta-testers will provide more in-depth feedback and assess usability of the toolbox using a System Usability Scale (SUS) questionnaire. We expect to reach intermediate to high usability scores (above 75 out of 100) at this stage.

During the final month of the project we will organise a workshop to promote the toolbox, discuss SSVEP methodology and provide a starter course for those who want to use SSVEPs in their research. We expect 30-40 in person attendees and 200-300 views of recorded sessions in weeks following the workshop.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was from a strong team and aimed to produce an open source toolbox for use in the field of cognitive neuroscience. However, the proposal would have stronger if an entirely open source solution had been used.

Title **RAMSES: Development & Testing of an R Package Enabling Hospital Antimicrobial Stewardship Analytics**
Lead Applicant **Dr Peter Dutey-Magni**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

ContentMOOC

Lead Applicant

Dr Peter Murray-Rust

Details of proposal – team members and collaborators

The ContentMine Team (now 5 years old) has experience in managing virtual projects. They provide and support the infrastructure (software, integration, documents, dictionaries, tutorials). Through a variety of projects and small awards over 10 years PMR and ContentMine have generated a dynamic international virtual community of over 20 Early Career scientists committed to making science Open and available. (presented at Eastbio DTP last week: https://www.slideshare.net/petermurrayrust/early-career-reseachers-in-science-start-early-be-open-be-brave).

Jon Tennant (one of these) has founded OpenScience MOOC (Massively Open Online Courses), https://opensciencemooc.eu/ a multinational virtual community for scientists that is developing online community-driven courses and other resources. The MOOC provides a welcoming environment for newcomers and a place where concerted action happens.

We have been in regular contact with Crossref, Unpaywall and CORE about using their services and will benefit from their mentoring.

We've been funded by Wikimedia to develop annotation and Wikidata-based searching and will use this as a core technology; we are enthusiastically part of the Wikimedia community.

Details of proposal – vision, aims and influence on open research

(i) To develop an Open framework for automatically mining the biomedical literature.

To nurture a community of Early Career Researchers (Fellows and volunteers).

To develop and deploy technology and to evangelise the benefits.

Example Scenario:

ECR1: "Help! My PhD is on Zika and I've got to do a first-year literature review on mosquito vectors. I've got over 24000 hits in Google Scholar but many aren't relevant".

ECR2: "Have a look at ContentMOOC! They've got Open software that carries out searches on EuropePMC. You can download thousands of articles, search them for diseases, insects, drugs, insecticides, countries, funders … and analyze the results with RStudio."

ECR1: "I'll need help. No one else in my lab is working in this area."

ECR2: "You'll find help in the OpenScienceMOOC – it's a virtual community created for collaboration. They even run courses. And ContentMOOC offers online mentoring through a Fellowship programme. You could apply!."

This problem is repeated in healthcare areas every day. Over 5000 scientific articles, preprints, theses are published daily and researchers just can't keep up. Research establishments require ECRs to do formal literature searches, often on thousands of references taking months by hand. So the ContentMine community have developed a framework for making searching automatic - results within a morning. We started by successfully creating a resource for a 1-day workshop on Crop science [1] in India using EuropePMC content. Now we are extending this to biomedical and healthcare.

Now we are joining with OpenScienceMOOC to create ContentMOOC, where courses and technology are developed bottom-up by ECRs. We want to explore other Open literature sources and have been in contact with bioarxiv (to include preprints) and Crossref and Unpaywall to include as many "free to read" documents as possible. Part of the current proposal is to develop APIs that extend the (Open) ContentMine code to download and normalize these automatically. Alongside the technology, we've already developed and deployed a Fellowship program, tutorials and mentoring. We advertised globally and appointed Fellows from 6 countries including a 15-year old ECR and one from Brazil. OpenScienceMOOC (which supports multiple languages) gives us the opportunity to explore this on a wider basis and nurture a dynamic volunteer resource (seeded with Fellowships). We'll advertise for Fellows to (a) develop technology (APIs and

content converters) (b) create biomedical Wikidata dictionaries (c) define and execute literature projects.
Influence on Open Research practices.
Open Science is Better Science.
It's as simple as that. It's quicker; almost lossless; better defined; transparent; reproducible; with communal wisdom and expertise.
ContentMOOC will explore how virtual communities can create lasting Open value. We've worked with (and been funded by) Wikimedia[2] to develop Open literature-based science. Our latest Wikimedia project, ScienceSource [2], downloads and annotates the key papers in a discipline such as tropical diseases. We've already shown that we can create high-quality semantic Wikidata-based dictionaries within an hour or two – this means that anyone can collect a set of relevant dictionaries to annotate, interpret and search the key literature.
We expect ContentMOOC not only to explore the technology for literature searching but be an example of how ECRs can create a new vision of literature-based science. It will act as a high-profile centre in Open research .
[1] https://github.com/petermr/tigr2ess/blob/master/OVERVIEW.md
[2] https://meta.wikimedia.org/wiki/Grants:Project/ContentMine/ScienceSource

Details of proposal – evaluation plan

We'll call on our panel of ECR scientists, especially past Fellows, to evaluate the success of the program (some have done this evaluation before). It is always difficult to predict the outcome of volunteer-based work – reasonable targets are:
(1) 6 Fellows completing a six month mentored program
(2) Two Fellow-created software demonstrators of extraction of documents from new sources (Crossref, Unpaywall, CORE, *-arxivs).
(3) Four projects in biomedical literature research.
All projects will be managed through open repositories (e.g. Github) with tools such as RStudio and Jupyter. These will be Open for inspection by the funders and the wider world. The MOOC will be a good dissemination platform and we've also had a long tradition of contributing to Wikimedia news.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal from a strong team. However it was unclear which aspects of the existing resource would be developed through this new proposal.

Title

Open and reliable evidence from systematic reviews on post-traumatic stress disorder (PTSD); development of a free online PTSD library.

Lead Applicant

Dr Sandra Matheson

Details of proposal – team members and collaborators

Dr Sandra Matheson (NeuRA) will be responsible for the development of the library.
Prof. Peter Schofield (NeuRA) will oversee the project.

Details of proposal – vision, aims and influence on open research

PTSD involves a cluster of symptoms that develop in some people who have been directly or indirectly exposed to a traumatic event. Whether PTSD develops depends on the severity and frequency of exposure, and on various regional and personal characteristics. The primary aim of this project is to provide a free online library of evidence on PTSD that is reliable and comprehensive. The secondary aim is to identify gaps in the PTSD evidence by highlighting areas that need more research, in order to answer clinically relevant questions.

The PTSD Library will build on an established platform that currently contains the Schizophrenia and Bipolar Disorders Libraries (http://library.neura.edu.au/). These libraries follow a stringent methodology that is similar to that used to develop treatment guidelines, with information updated every 6 to 12 months. The libraries are unique because the information provided encompasses all topics that have been reviewed on the disorders, and the information is quality assessed using objective methods.

The PTSD Library will include information gained from well-conducted systematic reviews and meta-analyses covering topics of risk factors, treatments, diagnosis, symptoms, physical features of the disorder (such as changes in brain structure and functioning), prevalence and incidence rates, common co-occurring conditions, illness course and outcomes, and issues that affect families. Reviews are rated according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting checklist, and only those reporting more than 50% of the recommended items are included. The evidence contained in each review is quality assessed using the Grading of Recommendations Assessment,
Development and Evaluation (GRADE) Working Group approach. This ensures the quality assessments follow objective guidelines in order to identify low, moderate and high quality evidence.

The Schizophrenia Library currently contains around 450 topics, and the Bipolar Disorders Library contains around 350 topics. Preliminary database searches suggest the PTSD Library should initially contain around 250 topics. The PTSD Library will consist of two tiers of information for each topic. One-page factsheets in html and PDF formats aimed at patients and families contain an introduction and conclusion, including evidence quality. More detailed technical tables in PDF format aimed at clinicians, researchers and policy makers contain an introduction, detailed method, all review findings with relevant statistics, quality assessments, and conclusions. Full review citations are provided with links to the free abstracts online or the free full-text if available. The technical tables are the backbone of the library and ensure transparency in our reporting and quality assessments.

(ii) how your proposal will influence open research practices in your field or more broadly

The PTSD library will provide a 'one-stop shop' of all the information gained from systematic reviews. Whilst some reviews' full texts are available online via open access, most require journal subscription, and their abstracts do not contain all the relevant evidence, particularly null findings. The Library technical tables provide this detailed information, which contributes to open research practices by ensuring this evidence is freely available. Further, by identifying the gaps in the evidence, research is encouraged in areas needing investigation via systematic review (pending enough primary studies), which in turn is included in the library.

We also encourage collaboration via feedback from library users and by developing partnerships with carers and consumer groups to gain input into the design and content of the library. We have developed partnerships with policy makers and researchers to encourage collaborative reviews and evidence briefs based on library methodology.

Details of proposal – evaluation plan

Our target for the development of the PTSD Library is that all or most of the topics will be completed and available online by the end of the 12 month period. We will be monitoring the number of visits to the library, where the visits are coming from, and which topics are most viewed. As the Schizophrenia Library traffic has increased from around 200 visits per month when it was completed in 2013 to nearly 13,000 visits globally in the month of April 2019 alone, we envisage similar traffic increases to the PTSD Library over time. The Bipolar Disorders Library was launched mid May 2019, therefore statistics for that library are not yet available.

As the entire output of the PTSD Library will be rigorous reviews that result in high quality guidelines or recommendations, we envisage the potential to contribute input to public policy bodies such as Australia's Sax Institute and the Australian Healthcare and Hospitals Association's Deeble Institute, who have previously commissioned us to undertaken evidence checks using library methodology (for examples see: http://ahha.asn.au/publication/evidence-briefs/are-our-policies-and-laws-leading-treatmentdelays-people-schizophrenia, http://ahha.asn.au/publication/evidence-briefs/does-casemanagement-improve-outcomes-people-schizophrenia, and http://www.saxinstitute.org.au/publications/management-of-suicidal-behaviour/).

Other external validation will come from The Health On the Net Foundation, which has endorsed the existing libraries for providing reliable web-based health information (https://www.healthonnet.org/HONcode/Conduct.html).

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal addressing an important topic. However the level of innovation was felt to be limited and the potential impact of this proposal to transform health research through openness was considered limited.

Title

Distributed whole genome sequencing analyses for personal genomics

Lead Applicant

Dr Shaman Narayanasamy

Details of proposal – team members and collaborators

Megeno S.A. was formed in June 2018 by a group of accomplished Luxembourg-based scientists and experienced entrepreneurs from the Munich biotech/health sector. Our interdisciplinary team of 11 people encompasses experts from systems biomedicine, bioinformatics, data science, business development, information technology and corporate development. Five personnel from Megeno will be allocated to the project:

Dr. Robert Koesters

PhD in Molecular Biology and former experience, including an associate professorship with focus on various cancers. Developed a mouse model for renal diseases.

External communication.

Dr. Dheeraj Bobbili

PhD in Biology specializing in bioinformatics for human genomics. Involved in several epilepsy and neurodegenerative disease consortia.

Human -genomics and -genetic analyses. Software development.

Sorin Ivan

Experienced full stack developer with more than ten years of working with IBM.

Infrastructure design and web- and software- development.

Dr. Shaman Narayanasamy

PhD in Biology specializing in bioinformatics for microbiome studies. Experience with reproducibility in bioinformatic workflows.

Project design, coordination, communication, bioinformatic workflows and software development.

At present, Megeno is informally collaborating with various public institutions. We will formalize those existing collaborations and continue pursuing collaborative relationships within the scope of the

proposed project.

Details of proposal – vision, aims and influence on open research

Whole genome sequencing (WGS) is widely used in research and healthcare. Over the next few years, millions of Europeans will have their genomes sequenced thanks to increasing capacities and

diminishing costs. WGS data is special, as it will become relevant across a multitude of human diseases, beyond the relatively narrow focus of its primary research topic and/or clinical utility. The European life-sciences Infrastructure for biological Information (ELIXIR) and the Million European Genomes Alliance (MEGA) have been proponents of federated genomics research within Europe. Consequently, ELIXIR is actively developing a platform for distributed WGS data analyses to further promote European-wide federated genomics research. This platform enables researchers to securely access and analyze WGS data without transferring large volumes of WGS data on site. Furthermore, researchers can leverage the computing capacity (storage and processing) of highly capable institutions using containerized software (i.e. docker and singularity), ensuring high reproducibility. Last, but not least the sensitive WGS data remains secure within the confines of the responsible party.

Unfortunately, federated research does not bring direct benefit to research participants and patients (hereafter referred to as sequenced individuals). Our European-wide survey indicated that 15 out of 31 institutions (in 20 countries) that generate large amounts of sequencing data experienced cases of sequenced individuals requesting for their own raw WGS data (e.g. FASTQ, BAM, VCF formats). Interestingly, all surveyed institutions recognise the rights of sequenced individuals to access and control the usage of their own WGS data (manuscript in preparation).

Accordingly, the main advantage of individuals having access and control of their raw WGS data is future re-use for:

• Genome-informed clinical diagnosis and treatment (e.g. pharmacogenomics)

• Genome-informed prevention (e.g. actionable secondary findings such as hereditary breast and ovarian cancer or iron overload)

• Inheritance of WGS data by next of kin (e.g. through a digital will)

Institutions that generate large amounts of sequencing data are highly cautious on transferring WGS data sets outside of their environment, mainly due to security, internal policies, regulatory requirements and downstream implications on the sequenced individual. Cutting-edge distributed genomics infrastructure (i.e. distributed WGS data analyses) is an ideal solution to bypass this issue as data never leaves the original institution, but still offers the opportunity for secondary analyses. However, present institutions are unable to manage genomes on a personal level, for the benefit of individuals, due to various complexities associated with personal genome management, including i) identity management of individuals, ii) reversing pseudonymisation, iii) interoperability and iv) managing ethical, legal & social implications (ELSI).

We intend to develop an open framework that leverages distributed WGS data analyses for the personal benefit of sequenced individuals. In this project, we aim to systematically evaluate the potential impact, benefits and/or risks of the platform through a small scale pilot study by engaging all relevant stakeholders and beneficiaries. Accordingly, our target audience(s) include i) genome

sequenced individuals, ii) European institutions with large volumes of WGS data and genomic analyses capability and iii) experts in the field of human genetics/genomics. It is paramount to involve all stakeholders within the European genomics ecosystem to develop a compliant, trustworthy and highquality system for personal genomics. Briefly, the activities in this project are:

i) Collecting input from experts across Europe on potential implementation strategies of distributed genome data analyses

ii) Develop and test the tools relevant for distributed genomics analyses (e.g. APIs, bioinformatic pipelines, containerized software, etc)

iii) Onboard several of the aforementioned institutions as collaborators to test and evaluate the approach

iv) Performing a “real world” pilot with collaborating institutions, involving sequenced individuals (i.e. patients and/or research participants)

In summary, we describe a platform that will promote the reuse of existing WGS data sets by leveraging cutting-edge distributed genomic analyses. In the long term, this approach will enable research outputs to directly benefit sequenced individuals and promote a new paradigm of open practices in personalised genomic research and healthcare. Finally, this approach must be setup now before millions of individuals are sequenced in the upcoming “genomic wave”.

Details of proposal – evaluation plan

Successful outcome of the project depends on effective engagement with the relevant stakeholders.

Therefore, we will measure the success of this proposed project by:

i) The amount of institutional feedback collected - we will attempt to collect opinions from at least 10 representatives of distinguished across Europe

ii) Number of institutional collaborators - From the institutions in i), we will aim to onboard least three as pilot partners for the project to test and evaluate our approach

iii) Recruitment of sequenced individuals - We intend to build pilot cohort of approximately 100 people from the onboarded institutions to pilot the distributed genomic infrastructure

iv) Provide direct benefit to the cohort - Including a subset of the pilot cohort in an iron overload (e.g. hemochromatosis) study pioneered by Megeno, in collaboration with experts of the topic. Involvement of any sequenced individuals will be based on consent

Megeno has a track record of effective communication with distinguished institutions across Europe and therefore will be able to achieve the goals described in i) and ii). Additionally, Megeno has developed a web portal that is accessible by institutional and private (sequenced) individuals to bridge the gap of personal engagement, and hence will enable the execution of iii) and iv). Overall, the existing platform will serve as a strong foundation for the development and execution of the proposed projects. Finally, the results from successful outcomes will be reported within peer-review publications.
Decision Not shortlisted
Comment on decision from Wellcome This was an innovative proposal. However, it was felt there were important ethical questions which were not addressed and the relationship to other activities was not clear.

Title

A Practical Guide to Open Science

Lead Applicant

Dr Siouxsie Wiles

Details of proposal – team members and collaborators

Associate Professor Siouxsie Wiles (University of Auckland) – Project lead, 0.2 FTE

A/P Wiles will be responsible for the direction and evaluation of the project and will supervise the research assistant and organise the workshops. A/P Wiles heads the Bioluminescent Superbugs Lab which currently comprises a mix of postgraduate students, research technicians, and postdocs, and recently completed the Mozilla Open Leaders programme.

Research Assistant (to be appointed) – 1 FTE

The research assistant will be responsible for carrying out the project, including doing the literature reviews and developing the documentation.

Details of proposal – vision, aims and influence on open research

In this project, we aim to produce a step-by-step guide for designing a project so that each part of the research process can be made open for scrutiny and replication, and any data generated will be FAIR - Findable, Accessible, Interoperable, and Reusable. As the template, we will use a project in my lab focused on discovering new antimicrobial agents. While the document we produce will be tailored to the types of experiments used in antimicrobial drug discovery, the processes and the wider lessons we detail will be easily adaptable to other fields of medical and biological research.

Open but not FAIR…

I am a health researcher who would like to make my data and research processes more open and transparent. This is not something that was part of my scientific training, and I have discovered that opportunities to receive training are limited. As a result, I have sought informal advice from researchers currently practicing 'open science' and begun making the raw data from our recently published studies available on Figshare.[1] In May 2019 I completed the Mozilla Open Leaders programme[2] and now understand that while I have made my data open, it is not fully accessible, interoperable, or reusable because it unintentionally lacks much of its associated metadata.

As well as the clear gains achievable if we are able to reuse the vast amounts of health data produced worldwide, another driver for making health research more open is the 'reproducibility crisis', the finding that many published scientific studies are difficult or impossible to replicate/reproduce. This phenomenon has been well documented in psychology and cancer research.[3,4] In response, numerous labs have developed lab manuals that document their workflows and the code they use to analyse their data.[5,6] Similarly, the Open Science MOOC community is developing online modules to equip students and researchers with the skills they need to practise open research.[7] The community has already released the 'Open Research Software and Open Source' module[8] and plans to develop modules on 'Reproducible Research and Data Analysis' and 'Open Research Data'.

A limitation of these lab manuals and modules is their focus on the processes that convert raw data into analysed data; in general, the parts of the research cycle that lead to generation of the data itself are missing. How data is generated is as important as how it is analysed, especially if the intention is for the data to be reusable. Even the award-winning Open Source Malaria project,[9] an exemplar for open drug discovery, lacks crucial experimental details for some of the data reported. For example, while the results of toxicity testing are provided, details of how the experiments were carried out are missing.[10]

This is the gap that our project aims to fill. Using workshops, web searches, and literature reviews, we will develop detailed guidance on what metadata should be provided to fully capture how experimental data is generated. We will also provide details on how to fully document the research process, from literature searches, to the experimental equipment and reagents used. Our focus will be less on what open tools to use and more on what information should be

supplied. The documentation we develop will be suitable for all researchers in the biological sciences, especially those that are not familiar or comfortable with coding. All documentation will be made available online under a Creative Commons licence and will be suitable to form the basis of a new module for the Open Science MOOC.
Our vision is that the step-by-step guidance we produce will contribute to improving the quality of health research data made openly available by researchers worldwide, beginning with the data produced by my lab. However, through my role as a section editor for PeerJ, and by working with microbiology and other societies and journals worldwide, and the Open Science MOOC community, we aim to influence open research practices within the wider microbiology and biological research communities.
References:
1. https://auckland.figshare.com/articles/Effect_of_common_and_experimental_anti-tuberculosis_treatments_on_Mycobacterium_tuberculosis_growing_as_biofilms_/4097772
2. https://foundation.mozilla.org/en/opportunity/mozilla-open-leaders/round-7/
3. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.
4. https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology
5. https://github.com/alylab/labmanual
6. https://github.com/cbahlai/OSRR_course
7. Tennant J, et al (2018, September 23). OpenScienceMOOC/Module-5-Open-Research-Software-and-Open-Source (Version 2.0). Zenodo. http://doi.org/10.5281/zenodo.1434288.
8. https://opensciencemooc.eu/modules/
9. http://opensourcemalaria.org/
10. http://malaria.ourexperiment.org/biological_data/11081

Details of proposal – evaluation plan

We will monitor and evaluate the success of our project on four levels:
1. Uptake and use by members of the Bioluminescent Superbugs Lab. We will measure which parts of the guide are taken up by new and existing lab members, and for existing members how the guide changes their current practices. This will be done using a mixture of questionnaires, workshops, and interviews.
2. Uptake and use by the wider community. We will measure the number of page views, downloads, citations, and adaptations. Our target is for the guide to have been viewed over 500 times, downloaded 50 times, and to gain 5 citations/adaptations within 12 months of it being made publicly available.
3. Endorsement by relevant journals and societies. Our target is for 5-10 microbiology societies/ journals to endorse the guide within 12 months of it being made publicly available.
4. Suitability to be converted into an Open Science MOOC module. Our target is for the guide to be used as the source of a module within 6 months of it being made publicly available.

Decision

Not shortlisted

Comment on decision from Wellcome

This was a potentially impactful proposal from a strong applicant, aiming to develop guidance on good metadata collection. However concerns were raised about uptake, as well as the static nature of the guidance being produced.

Title

Developing and Operationalize Digital Access and Usage of Open Health Research Data in Tanzania

Lead Applicant

Dr Sydney Msonde

Details of proposal – team members and collaborators

1. Dr. Sydney Enock Msonde, Head of Technical Services, Research, and Training, at MUHAS. He will oversee the implementation of the entire project and ensure health data are collected, stored and made available for FAIR use to a wider community.
2. Dr. Raphael Z. Sangeda is a data Manager of Sickle Cell Disease Programme at MUHAS. He is involved in various data management projects such as FAIR project of the Pan African Bioinformatics Networks (H3ABIOnet). He will co-design data management plan for data storage in the repository.
3. Dr. Felix Sukums, a Director, Directorate of ICT at MUHAS. He is involved in a number of multicounty and multi-disciplinary research projects in information systems and eHealth. He will be responsible for platform requirements elicitation, analysis, design, development and integration with the existing systems.
4. Dr. Rehema Chande-Mallya is a Head of Reader Services: Directorate of Library Services at MUHAS, and a retired AHILA president. She is an expert in evidence-based training in Africa.
5. Dr. Paul Erasto Kazyoba is a Chief Research scientist and Director of Research Coordination and Promotion at the National Institute for Medical Research (NIMR), Tanzania. He will be involved in coordination, monitoring and evaluation of project activities.

Details of proposal – vision, aims and influence on open research

Every year researchers and students from Muhimbili University of Health and Allied Sciences (MUHAS) and other health institutions in Tanzania produce a big volume of research data that are owned by either individual researcher or a research project. In most cases the data collected are less used due to limited data sharing mechanisms among researchers, lack of trust and belief to share locally generated data online, and skills on how to use open health datasets. As such, there is a need for a locally managed health data repository which dissipates the fear of local researchers to openly share data. It is expected that after building confidence in local repository, researchers will get abreast of data sharing and extend the habit of using open datasets in their research. Therefore, this project will develop a scientific data sharing platform which will create a big data pool of health datasets that will allow other researchers to carry out comprehensive followup research in order to meet the increasing demand for health research. The platform is expected to handle data heterogeneity across disciplines within and outside the university. Similarly, the platform will promote multi-disciplinary collaborative research among MUHAS faculty and students as well as with other researchers in Tanzania. The shared data will be made available to the whole scientific community as per FAIR principles. The project will utilize the Tanzania Education and Research Institution Network (HERIN) to transmit the developed solution to other institutions. Such undertaking will allow continuity of research work to inform policy formulation and improvements of healthcare services. The overall aim of the project is to develop an open health data research platform in Tanzania, which will allow quick discovery of scientific research to improve people's health. Specifically the project intends:

1. To determine researchers' awareness and attitude towards Open research Data use in Tanzania
2. To develop researchers capacity on the FAIR use of open health research data in Tanzania
3. To develop standard operating procedures (SOPs) on using open health research data based on the Tanzania Open data policies.
4. To design and implement an open research platform for collecting and storing the available health data in Tanzania

5. To package research data with codes for easy accessibility and reusability using research electronic data capture (REDCap) for data from postgraduate students,' researchers and two previously
Wellcome trust funded projects.
6. To evaluate the extent of adoption of open research data platform in Tanzania
The target audience is the faculty members, researchers and postgraduate students at MUHAS and from two selected health institutions. We will also work with IRB bodies of public Universities and research institutions in Tanzania so that implementation of Open Data can be mandated during the authorization of ethical clearance. We hope that the outcomes of this project can be scaled up across the country and other countries that are yet to implement Open Data policies.
This project will have the following activities:
1. Conduct feasibility study to faculty, researchers and postgraduates on awareness and attitude towards Open research Data use.
2. Create standard operating procedures (SOPs) on using open health research data by building on the Tanzania open data policies and standards for data sharing plan.
3. Capacity building to create awareness and skills on the use of open health research data platform to MUHAS community and two selected health institutions in Tanzania
4. Designing, installation, configuration, testing and fine tuning of the platform for packaging data with specific codes for easy accessibility and reusability.
5. Develop and conduct a post-survey for adoption of Open health data platforms.
6. Conduct monitoring and evaluation of the efficiency of open Health Research platform in Tanzania.
The proposal will influence open research practice by first, increasing awareness, skills and change of beliefs among researchers on sharing and use of Open Research Data platforms in Tanzania. Second, researchers will get expertise to use open research data and publish the generated knowledge in open access outlets. Third, researchers will get opportunity to harness valuable health datasets and make quick scientific discovery in unexplored area at very minimal costs. Fourth, researchers will increase their productivity in research by reusing the available datasets and address the increasing demand for health research discoveries.

Details of proposal – evaluation plan

The monitoring and evaluation activities of the project will help to assess the impact of the implemented Open Research Data platform on the academia and its utilization within the country and abroad. Throughout the project monitoring will review the progress with the intention to improve the project design and functioning. This is by receiving feedback from the participating academicians, the IRB bodies of health institutions who authorize ethical clearance in Tanzania. Furthermore, we will evaluate the outcome of a project by assessing number of research datasets deposited, number of datasets published in data journals, and number of citations received from the published datasets among others. The end line evaluation of the outcomes will inform decision making for policy and practice on the publishing of health research datasets in the platform. We will also assess the impact of the published research data by looking at the usage statistics
of the health research data repository developed over time and changes in belief and practices among faculty and researchers' on sharing and use of open health research data in their academic and research works.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal was felt to have good potential to positively impact health research. However, it was lacking in some key details, the extent of openness was unclear and there were concerns over the sustainability of the proposed platform

Title

Improving the reuse of research results by text-mining open access research articles

Lead Applicant

Dr Sylvain Massip

Details of proposal – team members and collaborators

The project will be led Sylvain Massip and Charles Letaillieur, the two co-founders of Opscidia (see below). Sylvain will bring his more mathematical approach whereas Charles will bring the computer science point of view. A suitable post-doc level employee will complete the tam.

Dr Sylvain Massip holds a master's in applied mathematics and a PhD in Physics from the University of Cambridge. He has worked for more than 10 years at the interface between research and industry, as an academic researcher, and as a R&D manager in a start-up, with multiple academic collaboration. His experience made him see the potential of text-mining full-text research articles. In 2018, he completed a certified professional diploma in Data Science and Machine Learning at Ecole Polytechnique.

Mr Charles Letaillieur is an engineer from Telecom ParisTech (2007), majoring in artificial intelligence. He has 10 years' experience as a web professional in several professional contexts, such as Cap Gemini consulting as well as start'ups. Recently, he has specialized in e-democracy and digital transition of the public sector. In this context, he has developed an expertise in open data and big data.

Details of proposal – vision, aims and influence on open research

(i) Presentation of Opscidia and general vision

Opscidia is an Open Science company created in 2019. Our general vision is that the knowledge included in academic articles, if correctly reused, has enough value to fund the open access publication of research. Hence, we develop a free and open-access academic publication platform that will be funded by the development of technological intelligence tools based on text and data mining corpuses of open access publications and research data. For that purpose, Opscidia establishes partnerships with research communities in order to solve practical problems with high added value. For example, Opscidia has recently started projects in geotechnics and in material sciences.

(ii) Concept of the project and motivation

We propose to develop text-mining applications on the full text of open access articles in the domain of infectious diseases in order to favor the reuse of research results by the researchers themselves, and by healthcare professionals such as health agencies in developing countries. Following the 2014 Ebola outbreak in Liberia, Sierra Leone and Guinea, three Liberian healthcare professionals wrote an editorial in the New-York Times [Dahn et al., 2015] stating that "yes, we were warned about Ebola". They showed that an article from 1982 published in the annals of virology stated that Liberia had to be included in the Ebola virus endemic zone. Because of paywalled access to research papers, and because of the difficulty to retrieve information from massive amount of literature, Liberian healthcare professionals such as them did not have access to this information. An access to this information before the outbreak would have saved many lives.

The aim of this project is to try to prevent that such a case will not occur in the future.

(iii) Activities

A small proof of concept, focused on the Ebola case, was developed using simple machine learning tools as well as indexing tools. The main results were presented as a poster at ElPub and COAR 2019 conferences [Massip and Letaillieur, 2019]. In particular, co-occurrence of Ebola and Liberia, Sierra Leone or Guinea, was indeed found in the full-texts of several papers published before 2014. Hence, this proves that the risk of the Ebola outbreak could have been predicted, provided that weak signals had been correctly interpreted.

The project that we propose for the Open Research Fund is to extend this proof of concept to build complete tools, that provide more indicators, and deals with other major infectious diseases of Africa.
- We will team up with infectious diseases specialists, for example laboratories at Institut Pasteur or the French National Research Institute for Development (IRD) in order to select the most promising weak signal indicators that can be text-mined from the full text of research articles.
- We will retrieve a corpus of open access articles from open repositories such as EuroPubMed Central.
- We will apply to this corpus a mix of biology-specific text-mining techniques such as PathText [Kemper et al., 2010] and more general Natural Language Processing techniques such as BERT [Devlin et al. 2018], in order to extract the most relevant information from the corpus.
- This information will be made into quantitative indicators, that can be made dynamic to follow the evolution of the literature.
- Depending on the type of reuse and the community reusing them, these indicators can either be used directly to formulate new hypotheses, or built into dashboards to be able to follow the evolution of the literature.
(iv) Open research practices
First, this project will favour the reuse of research results by researchers across disciplines and by healthcare professionals.
Furthermore, it will demonstrate the value of open access academic publishing for society as a whole, and not just for researchers. On a more technical side, it will also demonstrate the value of using CCBY licences and interoperable formats for the full texts, such as JATS.
The code developed in this project will be open source, and our approach will be duplicable in other fields.
Finally, we believe that finding practical value to the research that is published open access will help fund open access, and is the best way to develop open access across all fields of research.

Details of proposal – evaluation plan

The aim of the project is to build information retrieval solutions that are of use for:
- Healthcare professionals: our system will help them search the scientific literature for information about the infectious diseases they might face.
- Professional Researchers: to help them formulate new hypothesis and discover implicit links.
Hence, the first measure of success will be the number of relevant indicators that will be built. We aim at building 10-15 text-mining modules that brings relevant insights retrieved from full text articles.
The second success indicator will be the reuse of the developed tools by the two target communities.
We will measure the number of downloads of the code that we will develop and that will be made open source, the number of citations that it will generate, and the number of collaborations that will
be started on that basis.
Our aim is to generate around 5 to 10 collaborations in total with at least 2-5 collaborations with healthcare professionals and 2-5 collaborations with infectious diseases researchers.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal demonstrated a strong commitment to openness. However, it was felt the objectives of the proposal and its potential impact were felt to be unclear

Title

Translate medical research publication metadata for easy access through Wikidata and Wikipedia

Lead Applicant

Lane Rasberry

Details of proposal – team members and collaborators

Data Science Institute, University of Virginia:
Daniel Mietchen, data scientist
Lane Rasberry, Wikimedian in Residence
The Data Science Institute at the University of Virginia is importing structured data into Wikidata, evaluating the quality of information in Wikipedia and Wikidata, and documenting best practices for university partnerships with the Wikimedia platform. This team identifies the research content for which there is a need in research discovery to be FAIR, available for query, and translated to promote global collaboration.
UVA Global, University of Virginia:
This is a language department at the university where faculty and classes will translate and publish structured data into Wikidata, where it will be FAIR and open in the semantic web.

Wikimedia community organizations:
These organizations provide community feedback on publication in Wikipedia and Wikidata and also on the translation process. These partnerships ensure participation among stakeholders and regional communities of users.

Details of proposal – vision, aims and influence on open research

This project seeks to conduct language translation on metadata labels for research publications, attribution data, and clinical trials information to make data about medical research queriable in underserved languages through Wikidata and the Linked Open Web. This project has the benefit of distributing content through Wikipedia and Wikidata, which already have an annual userbase of a billion users and which already have established actionable standards to practice diversity, inclusion, openness, FAIRness, and transparency about program development. The impact will be localized access to basic research information in various Global South languages to integrate with existing community efforts for establishing the same. Although Wikidata development in this direction seems inevitable, the cultural and social exchange required to establish global multilingual research partnerships could begin now with support rather than later as a second phase effort for including the developing world. Wikipedia and Wikidata are established forums with an existing active userbase for multilingual research collaboration, but the research practices there still are immature. By applying metadata expertise through this project, we will elevate the current amateur development with more stable Linked Open Data compatibility to English language databases. Using the wiki distribution and discussion platform to develop the global conversation about data sharing will set good precedents for the trend of global research collaboration.
Methodology
1. By default, adopt the established Wikipedia and Wikidata publishing and engagement practices for open, FAIR, documentation, receiving feedback in permanent public forums, and collaboration
2. Contribute to the documentation about the position of Wikipedia and Wikidata in the Linked Open Data ecosystem, particularly emphasizing university participation in import and export of research metadata in the Wikimedia platform and collecting impact metrics for doing so.
3. Within Wikidata, contribute to the WikiCite project which seeks to enrich data around citations and metadata, including PubMed research papers, and subsets of CrossRef, ORCID, and ClinicalTrials.gov. Explore and document possibilities to ingest non-United States clinical trial databases.

4. Identify the set of terms and concepts which are necessary to perform and visualize queries of medical research data, for example, "clinical research sites in a given country with the highest trial completion rates in infectious disease research"
5. Translate those terms to languages including Hindi, Bengali, and Swahili to the level of quality which is established as a norm by existing local community participants in Wikipedia and Wikidata
6. Use Wikipedia and Wikidata's native metrics reporting processes to measure the impact to users and the engagement of peer reviewers

Details of proposal – evaluation plan

This project will use Wikipedia's own established processes for monitoring and evaluation of university projects to develop and publish general reference information in Wikipedia and Wikidata. We will evaluate this project in these established ways:
1. Content metrics - Report the standard publishing metrics as measured by Wikimedia's own native metrics suite for publishers
2. Diversity and inclusion - Partner with established Wikimedia community organizations; confirm their oversight and approval
3. Impact metrics – Report audience readership as measured by Wikipedia's own native metrics suite for users
4. Quality review - university student researchers will evaluate and publish an evaluation of the source research metadata and the translation process
5. Bias evaluation – we will subjectively publish our opinions on bias we identify and its cause. One obvious source of bias will be availability of open data, as much research indexed in PubMed and ClinicalTrials.gov is not compliant with recommended metadata standards. This project favors institutions which apply FAIR principles, and we will identify these practices.
Wikipedia as a publishing, technology, and community platform continually introduces processes for content development and evaluation. In 2012, with the establishment of the Wiki Education Foundation, there was a major cultural shift to make Wikipedia compatible with university education and research. Today, that precedent has developed into a suite of open evaluation tools for measuring audience size, levels of engagement, use of fact-checking processes, and a culture applying metrics to perform critical review of Wikipedia's quality. These measurements and processes establish a precedent for this project to follow in doing publishing and content development, operationalizing ethics in digital governance, and publicly demonstrating community conversation in seeking feedback on this project's activities in the context of global Wikimedia content development.

Decision

Funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title **Harmonization of Patient Registries to Support Cross-Disease Analysis of The Barriers and Facilitators of Access to Rare Disease Care**
Lead Applicant **Marissa Schlemmer**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

BNA-UKRN: Improving the credibility of neuroscience research through open science

Lead Applicant

Miss Sophie Sykes-Jerrold

Details of proposal – team members and collaborators

The proposal is intended to support the work of the British Neuroscience Association (BNA), the leading UK neuroscience organisation, to increase the credibility of neuroscience research in the UK. It has been developed in partnership with the UK Reproducibility Network (UKRN). The project will be led by Georgina Hazell, Head of Policy and Campaigns at the BNA (and Enterprise Fellow at the University of Bristol) and guided by an Advisory Board of experts and representatives in credibility issues. Georgina's credibility work will be supported by Anne Cooke, BNA Chief Executive, and Sophie Sykes-Jerrold, BNA Development Director. The project will delivered in partnership with the UKRN, with Marcus Munafò, Chair of the UKRN Steering Group and Professor of Biological Psychology at the University of Bristol, acting as the primary liaison between the BNA and UKRN. Through this partnership we will be able to work with a range of other relevant UKRN stakeholders (e.g., other learned societies) and local networks from key UK institutions, along with the wider BNA membership and the organisations with which we already work closely on research credibility and reproducibility issues. A list of UKRN stakeholders and local networks is available at ukrn.org.

Details of proposal – vision, aims and influence on open research

In April, the BNA launched a consultation to help understand neuroscientists' perceptions of credible, open and reproducible research practices*. To date, we have had over 600 responses from researchers from a variety of neuroscience disciplines. Initial analysis reveals that most participants are in full support of open science credibility initiatives and believe they will improve the quality of neuroscience. However, very few participants have engaged directly with open and reproducible research activities and describe a lack of dedicated funding, positive incentives, training and support, and fear of being scooped as the main obstacles to credible research.

There is therefore a clear need for initiatives that will promote and support the uptake of open science practices within the neuroscience community across the UK. As the leading professional body for neuroscientists in the UK, we feel it is our duty to create an environment in which neuroscientists feel comfortable to conduct open and reproducible research, and address the concerns highlighted in our consultation. Working with the UKRN and informed by our credibility advisory board, we aim to remove the current barriers to credible neuroscience research. We will achieve this through a BNA-UKRN "roadshow" that we will deliver at major neuroscience institutions across the UK.

The roadshow will be focused on awareness raising and training delivery, and is intended for researchers at all career stages. These events will be linked to discussions with the senior management teams of institutions where the roadshows are held around how open science practices can be embedded and supported within those institutions. We will work closely with UKRN, who is currently developing guidelines for universities that want to create academic roles focused on research improvement. The content of the roadshow will be informed by discussions with UKRN stakeholders and researchers (through UKRN local networks). The roadshow content will include sessions on:

Selfish reasons to work reproducibly. This session will provide an overview of current debates around research credibility and open science, and highlight the various reasons for working practices that serve to improve reproducibility. This will be based on Five Selfish Reasons to Work Reproducibility by Florian Markowetz (published in Genome Biology).

Pre-registration and Registered Reports. This session will provide an overview of various forms of study pre-registration, including for different study designs (e.g., observational vs experimental) and ranging from self-registration of study protocols (e.g., on the Open Science Framework) through to the submission of a Registered Report to a journal that offers this format.

Data carpentry, data sharing and FAIR principles. This session will describe basic data carpentry skills that allow researchers to reproduce their own analyses, and curate their data sets to allow others to re-use their data. It will also provide an overview of FAIR principles of data sharing, and the repositories that exist that support data archiving.
Reproducible workflows and sharing materials. This session will extend on the principles of the data sharing session, and focus on tools for producing reproducible workflows (e.g., electronic lab notebooks, the NC3Rs experimental design assistant), and issues around sharing materials, code and other elements of the research workflow.
We will also use the roadshow as an opportunity to discuss these issues with the senior management team at each institution that hosts a roadshow. Specifically, we will discuss the creation of a senior role for research improvement (if one does not already exist), infrastructure support (e.g., a data repository), incentivising open science practices (e.g., through including these in promotion and hiring practices), and incorporating this training into staff development for researchers at all career stages.
Through the UKRN local networks and our membership we will open up the roadshow to researchers from other institutions in the region, if there is sufficient capacity. We will also develop online materials (e.g., recordings of lectures) to increase the reach of the roadshows beyond the host institutions and funded period. We will also host a parallel webinar series in partnership with ReproducibiliTea – an early career researcher led journal club initiative supported by UKRN.

Details of proposal – evaluation plan

Removing the barriers to credible research will undoubtably take more than the outlined 12 months. However, there are many ways to show that we are moving in the correct direction. Specifically, we will survey our membership before and after the delivery of the roadshows to assess:
Pre-registration / Registered Reports. The number of members reporting pre-registering a study protocol / submitting a Registered Report. We will also evaluate this at our journal.
Data sharing. The number of members reporting having shared data. We will also evaluate this at our journal.
Materials sharing. The number of members reporting having shared materials. We will also evaluate this at our journal.
Use of reproducibility tools. The number of members reporting having used a reproducibility tool (e.g., notebook or design assistant). We will also evaluate this at our journal.
Institutional change. We will monitor the number of institutions creating academic roles focused on research improvement, modifying promotion and hiring criteria, etc.
We will also leave our consultation open throughout the duration of the project and assess this for evidence of changes in understanding/perception of credible, open and reproducible research practices as we deliver the roadshow events.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to promote open research practices among the UK neuroscience community. However the level of innovation proposed was considered limited.

Title

Data Visualisation and Interactive Storytelling

Lead Applicant

Mr Alan Hyndman

Details of proposal – team members and collaborators

Alan Hyndman (Figshare) - Conceived of and shaped the project with key team members. He leads on delivery to ensure all objectives are met and stakeholders are listened to and satisfied.

Dr Erinma Ochu (University of Salford) - Academic lead, specialising in science communication including co-design to bridge the gap between policy and practice (Gold and Ochu, 2018). Acting as Figshare ambassador, Erinma will introduce relevant stakeholders, lead workshops and provide guidance, drawing on academic literature and advising on outcomes.

Megan Hardeman (Figshare) - the community lead at Figshare will liaise with researchers, devise and conduct training, solicit feedback, assist in writing up and sharing the research stories via social media.

Florin Apetrei (Figshare) - Head of design and UX at Figshare will lead the creative translation of data into compellingly designed, interactive stories.

Natasha Trotman - Inclusive design researcher and graduate in Information Experience Design (Royal College of Art), currently researching and designing inclusive spaces toolkit. She will contribute to co-design and user testing to ensure neurodiverse user perspectives are considered.

Reference

Gold, M. and Ochu, E.E. (2018). Creative collaboration in citizen science and the evolution of ThinkCamps. In: Citizen Science: Innovation in Open Science, Society and Policy , UCL Press, pp. 146-167.

Details of proposal – vision, aims and influence on open research

The Need

The recent rise in fake news challenges scientists and academic publishers to provide accessible, reliable, trusted sources of scientific knowledge to inform policy. Further, a Parliamentary Science and Technology committee exploring the communication of science to the public illustrates a clear desire for the public to know how science affects their daily lives, yet the public hold a strong belief that the media sensationalises science (The Science Communication and Engagement Report, 2017). This includes in relation to communicating insights from health research and climate science.

And yet, the results of scientific research have long been locked away behind paywalls, out of reach from the media, average citizens, scientists, policymakers and businesses who could potentially benefit from key findings. In spite of key developments in citizen science, where the public contribute to scientific research and learn about the process, the ability to analyse data is still out of reach of most citizens and new methods to communicate complex scientific data and ideas in a simple way without dumbing down the facts could enable people to access scientific knowledge.

This presents an opportunity for the research community, science communicators, science publishers and key users of science knowledge to share responsibility by offering trusted knowledge sources by co-designing innovative digital solutions. We believe that data visualisation and interactive storytelling is the perfect way to bridge this gap.

Data visualisation is a mode of 'brokered research communication' which serves the purpose of making research clear and accessible (Allen, 2018). At the same time, data visualisations must satisfy the informational requirements of scientist, academic and non-academic publishers, including the media, to provide a trusted source of information.

By involving and empowering researchers to develop interactive storytelling through data visualisation, as a trusted knowledge broker, informed by science communication research and practice,we can help bridge the gap between complex scientific ideas and key identified publics.

Vision

Our vision is to use data visualisation to reimagine how a scientific article can be communicated. We aim to innovate beyond the limited traditional PDF layout, with static figures. Why are papers so rudimentary when The Guardian can present their data like this?

Aims

There are pockets of good data visualisation happening in academia, we aim to make data visualisation a core competency of all researchers, so they know how to structure their data at the start of the project in order to visualise the data and share insights continuously. Researchers will grow confidence in visual communication to identified users, e.g. media and policymakers.

Good examples of academic data visualisation:

The Interactive Data Network at Oxford University

State of Open Data

NOAA

Going Critical

This project will provide data visualisation and storytelling services to researchers solving the big global health problems of our times.

Our project has 5 phases:

Phase 1 - Design

Working with Wellcome Open Research we will identify suitable data for visualisation. Then partnering the researchers to understand their data, the research story they are trying to tell and identify key audiences. We will review existing services, including for accessibility.

Phase 2 - Prototype and test

We use our internal expertise and collaborate with external visualisers to prototype different concepts, liaising with researchers and stakeholders through the process at key events.

Phase 3 - Develop

We will commission data visualisers and developers to execute the concept in the form of an interactive website and toolkit.

Phase 4 - Publish

All data will be made openly available, all code we generate will be made available open source and any associated story will be made available as a preprint on Figshare.

Phase 5 - Feedback, iterative, continuous improvements

Retrospective meetings with all stakeholders to identify how we can improve our process.

Publication of academic article documenting the process.

We will run this process iteratively, 2-3 times over 9 months. In the final quarter we will run a series of workshops, webinars and training sessions for researchers to gain skills and knowledge and test end-products. Figshare attends over 20 conferences every year and will be submitting to run this workshop at ~5-10suitable events.

The project will help all researchers and key public stakeholders grow closer together to understand and apply insights to tackle the big health challenges facing our society, from antimicrobial resistance, to mental health to the link between climate change and health.

Details of proposal – evaluation plan

We will commission evaluation to evaluate our project across a number of criteria:

1 - Participant feedback on the usefulness of the process, products, skills and knowledge gained through co-design activities, webinars, events and toolkits.

2 - Website visits, embeds and shares - the number of people who come to view, interact with and share the data visualisations.

3 - Altmetric attention - All data, code and preprints will be made available on Figshare so we can measure the views, downloads, citations and social media attention of all the outputs.

4 - Training Workshop and webinar attendance

Figshare will use our extensive network of researchers, institutions, publishers, funders and governments across the world to promote the importance of data visualisation. We will be looking

to run workshops at upto 10 institutions and online with our Figshare ambassador network to test the concept globally. 5 - Developing a sustainable service - We are looking to establish a sustainable data visualisation service at Figshare for the research community. A key metric of success of this project would be the sustained interest and user demand in continuing the project beyond 12 months.
Decision Not shortlisted
Comment on decision from Wellcome This was an interesting and innovative proposal. However, the potential impact to transform health research more broadly (beyond the projects directly involved) was felt to be limited.

Title

Back Your Scientific Stack

Lead Applicant

Mr Alex Morley

Details of proposal – team members and collaborators

Pia Mancini (CEO) & Alanna Irving (Executive Director) at Open Collective.

Open Collective is itself an open source tool designed for transparent funding, which has powerful functionality allowing the creation and maintenance of numerous project budgets, and the financial and legal infrastructure to accept and hold funds, handle taxes, provide real time reporting, and accountability for all stakeholders, including grantmakers, researchers, code contributors, users, and supporters—all with low overhead, so projects can stay focused on their mission.

As collaborators they will be providing both their expertise in transparent and accountable funding in the open source space as well as the infrastructure that Open Collective provides.

Details of proposal – vision, aims and influence on open research

Modern science is dependent on software, from data acquisition to data analysis and beyond. Often these software are are proprietary, and their cost eats away from already limited funding. Open scientific software (OSS) provides much better value for money for those funding research - being free to use and modify - and is often the only choice for researchers working in under-resourced institutions and countries. Importantly, investment in OSS has impact on a global scale as it democratizes access to software and research creating a global research community. Conversely, lack of investment in OSS results in higher direct costs for researchers paying software licensing fees, and higher indirect costs of lower research quality due to poor software.

We propose that in order to get the maximum value and sustainability from OSS some of the funds that would have been spent on software licenses instead be directed towards development and maintenance of existing open solutions.

In order to accomplish this we will take three complementary approaches: The first is to develop a tool to estimate the value of a given piece of open software by associating it with the most related proprietary solutions and ascertaining their current license cost. We have a functioning proof-of-concept tool already developed1. This tool was originally inspired by, and would build on the success of, a tool developed at Open Collective to make it easier for companies to donate to the open source projects they rely on2.

The second is to develop a reporting framework that can be used in grant applications to report the software that will be used in projects even when no license fees are being costed in. This could eventually result in aggregated “donations” by funding bodies to projects on which their grantees rely. This framework will be developed in partnership with key stakeholders including funding agencies, scientists and research software engineers as well as related non-profit organisations in the UK and the US with whom we have existing connections3.

Finally we will develop guidelines for costing development work for open projects in grant applications. Currently it is much easier to cost in payment for proprietary software than it is to pay for the work that would enable open software to reach the same standards. The advantage of the latter is that all researchers/grantees would benefit from investment in work on open projects, and the investment only needs to be made once. However this investment can take many forms: instead of no-strings-attached donations it could take the form of support contracts, or contracts for the development of specific features. These approaches lead to researchers being confident they will get the help they need to make the software meet the needs of their project with the side-effect of improving the software for everyone else. The guidelines would lay out these when these different approaches are most appropriate and extra considerations that might be involved.

While the reporting framework will be a resource for funding bodies and institutions, and the guidelines for costing development work will be aimed at tool developers and grant applicants, we

will use the same three-stage process to engage stakeholders from start to finish. Initially we will reach out for semi-structured one-to-one conversations about where these resources would fit in their processes and what solutions they already have in place. Next, we will organise a design sprint, a participatory process where we bring together 10 - 20 people in the same room to uncover in detail their current journeys and areas of uncertainty before iterating on some high-level ideas for solutions. Finally, after building on the information uncovered thus far to develop a prototypical solution, we will request individual feedback on where this solution could be improved.

Together we believe these approaches can have a positive impact on the research quality not simply by injecting funding in the short-term, but by building up the framework that is needed to sustainably support the plethora of open scientific software on which almost all research relies.

1. https://github.com/alexmorley/Back-Your-Scientific-Stack
2. https://backyourstack.com/
3. e.g. Software Sustainability Institute (UK), Code for Science and Society (US), NumFocus (US)

Details of proposal – evaluation plan

1. Tool to estimate the value of a given piece of open software

Key Deliverable: Tool should be built and released within 8 months of the start of the project.
Key Metric: Use - 5000 uses of the page to get an estimation of software value.
Goal: Raise awareness about the tangible value that open scientific software has.

2. Reporting framework for OSS use in research

Key Deliverable: A customisable framework that can be used in both grant applications and grant reporting. There should be broad consensus about the key aspects of the framework between stakeholders. The framework will be accompanied by a written report that is shared publicly.
Key Metric: Adoption - the framework to be piloted by 3 organisations for the project duration.
Goal: That the framework is adopted, or informs, future grant application/reporting forms with increased emphasis on evaluating which and where open software had an impact.

3. Guidelines for costing development work

Key Deliverable: Written guidelines, long-form, short-form and infographics that will be shared publicly.
Key Metric: Understanding - while these guidelines are being developed we will be doing user testing to ensure that readers take away the key messages.
Goal: Ensure that both developers and grant applicants have an understanding about what options are available and appropriate for funding open development in a research setting.
Stretch Metric / Goal: Approval - to get some form of the guidelines approved by a national funding agency - that they support the measures and permit the costs in some grant applications.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal with the potential to create a useful tool. However, it was felt that the proposed approach might not be successful in achieving buy-in and the impact would be limited.

Title

Accelerating the adoption of open research practices with a collective action platform for academics

Lead Applicant

Mr Cooper Smout

Details of proposal – team members and collaborators

Cooper Smout, B.Sc., is a Cognitive Neuroscience Ph.D. candidate at the University of Queensland (Australia), an Instructor at the Open Science MOOC, and a Researcher in Training at IGDORE. Cooper founded this project in September 2018 and has since grown a large support base through conference presentations, personal contact, the project website, and Twitter. He personally funded the minimal viable product (MVP) and plans to work on the project full time after completing his Ph.D. (October 2019). Cooper will direct the project development and promote it throughout the community.

Alfred Garcia, M.Sc.IT, is a software engineer with 16 years of experience and co-founder of Codi Cooperatiu SCCL (Spain). Alfred developed the MVP for this project. He will lead the platform development and deployment, as well as coordinate hosting, data security and data management.

Jonathan Tennant, Ph.D., is a research fellow at the Center for Research and Interdisciplinarity (France) and founder of the Open Science MOOC. He will coordinate communications and marketing.

Virginia Barbour, Ph.D. (Australasian Open Access Strategy Group, Australia) and Brian Nosek, Ph.D. (Center for Open Science, USA) will advise the project team and facilitate contact with relevant organisations.

Details of proposal – vision, aims and influence on open research

Vision

The Open Research movement is growing, but intense competition within academia limits the uptake of open practices by individual researchers. For example, many researchers choose not to share preprints because they fear it will cost them time (e.g., formatting and uploading) or prestigious publications (e.g., some journals don't accept papers that have previously been preprinted). Similarly, many researchers continue to publish in expensive subscription-based or 'hybrid' journals, rather than low-fee Open Access (OA) journals, because these venues are perceived to be more beneficial to one's career. Both cases are examples of a collective action problem, in which individuals could benefit from joint action (e.g., by enabling rapid access to the latest information and reducing publishing costs) but fail to cooperate due to conflicting personal interests. These (and other) open research practices could be instantiated as a cultural norm, however, by organising a critical mass of support for the desired behaviour and acting in a coordinated manner to bring about change ('collective action'). This strategy has proven widely effective in non-academic industries (e.g., trade union strikes) and political situations (e.g., civil protest). Currently, however, no mechanism exists to organise collective action in the research community.

We aim to build a platform that can measure and coordinate community support for various open research practices. Initially, we will host three 'flagship' campaigns: (1) 'Green OA', in which researchers pledge to share a preprint of every research article they submit to a journal for publication, (2) 'Gold OA', in which researchers pledge to publish exclusively in non-hybrid OA journals, and (3) 'Platinum OA', in which researchers pledge to publish exclusively in fee-free OA journals. Researchers will sign onto the platform (using ORCID) and select the following settings for each campaign of interest:

Authorship position/s that the pledge will apply to (first, middle, last)

Anonymity (prior to their pledge being activated)

Level of community support (e.g., 5%, 10%; details below) at which the researcher (and the rest of their pledging cohort) will be contacted and directed to carry out their pledge.

We believe that this platform can help to accelerate the global transition toward OA. The uptake of OA has been particularly slow in the health research community, despite the many benefits OA could have in the search for solutions to the world's biggest health problems (e.g., epidemics). To address this, we will partner with health-related organisations (e.g., AMA) and initiatives (e.g., medRxiv, ASAPbio) and utilise their networks to promote the campaigns (e.g., social media, mailing lists). We will also develop a network of ambassadors to promote the project and engage the community via conference presentations, blog posts, multimedia (e.g., animated videos) and advertising (e.g., social media). In particular, we will target early- and middle-career researcher communities (e.g., Right to Research Coalition) as this sector is the most constrained by cultural inertia but also the largest within academia.

The project will incorporate four major activities:

Beta-test the MVP (found here: http://fok.codi.coop)

Refine the platform

Release the platform

Promote the platform

Influence

Cultural inertia has impeded the growth of the OA movement and will continue to impede the uptake of open research practices well into the future. We believe that the proposed campaigns can help to align personal and societal incentives and spur the global research community into embracing behavioural change. Following the launch of the OA campaigns, we will develop the platform to host user-created campaigns aimed at a wider array of open research behaviours, e.g., pre-registration, loss-of-confidence statements, open data contributions, reporting medical trials (e.g., AllTrials), adopting open standards, or using progressive review platforms (e.g., PREreview).

We believe this project will influence open research practices in the following ways:

Increase awareness of open research practices

Facilitate cooperation between researchers

Empower researchers to drive change

Support 'top-down' policies (e.g., Plan S) with a complementary 'bottom-up' movement

Protect researchers made vulnerable by OA mandates

Generate data on the distribution of support for different open research practices

Pressure publishers and universities to update policies (e.g., preprint and hiring policies)

Produce open source code for calculating community support

Encourage innovation and adoption of new technologies

Details of proposal – evaluation plan

Software development will be delivered using an 'agile' methodology across a series of biweekly 'sprints'. Progress within each sprint will be monitored according to the number and difficulty of tasks delivered, the number of commits to a production code branch (found here: https://github.com/FreeOurKnowledge/platform), and the number of blocked or backlogged issues.

Community engagement will be monitored using 'support metrics' calculated separately for each campaign and research field, according to the following procedure:

1. Calculate the total number of citations generated by all articles in a particular research field (e.g., clinical sciences, psychology or neuroscience, as categorised by Dimensions: https://app.dimensions.ai/), separately for each of the last 10 years.
2. Calculate the subtotal of citations generated by publications on which pledgers were an author in the indicated position (first, middle and/or last). This calculation will be facilitated by requiring researchers to make pledges with their ORCID identifier and asking them to maintain their ORCID record.
3. Calculate the geometric ratio (to account for skewed distributions) of pledger citations to total citations, separately for each year, and average yearly ratios.

We aim to achieve 5% support for one campaign in a health research field within 12 months. We anticipate pledges in psychology to outpace other health fields due to the relatively wide awareness of problems in scholarly publishing (e.g., the 'reproducibility crisis'). To achieve this goal, we will need approximately 7,500 pledges from psychology researchers (assuming a uniform distribution of impact in the pledging cohort and 150,000 psychology researchers globally). Following this, we will implement the pledge activation phase and monitor the pledging cohort for non-compliance (posting any instances on the website). We will also analyse the pledge data along various dimensions of interest (e.g., researcher demographics) and prepare these findings for publication.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title **Structured, granular and attributable recognition of researcher Open Research activities in an Open Ledger.**
Lead Applicant **Mr Richard Wynne**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Not shortlisted
Comment on decision from Wellcome The applicant opted not to share this information

Title

JOGL (Just One Giant Lab): a collective and open lab environment to foster collaborative health research

Lead Applicant

Mr Thomas Landrain

Details of proposal – team members and collaborators

Just One Giant Lab (JOGL, https://jogl.io/), nonprofit initiative based in Paris. JOGL is the first research and innovation laboratory operating as a distributed, open and massive mobilisation platform for collaborative task solving. Founded by a platform designer, an open science activist and social entrepreneur, and a network science researcher, JOGL will lead the development of the proposed platform features.

Thomas Landrain [JOGL project executive lead]

Leo Blondel [JOGL project technical lead]

Marc Santolini [JOGL project scientific lead]

Camille Masselot [Lead on the Co-Immune open research program on vaccination]

The Center for Research and Interdisciplinarity (CRI Paris), an Institute aimed at fostering research at the crossroad between interdisciplinary life and health sciences, basic understanding of learning processes and novel education technology/methodology testing and implementation, and digital sciences. CRI Research Fellows involved are :

Anshu Bhardwaj [Will organize a scientific challenge on anti-microbial resistance on JOGL]

Marc Santolini [Will help organize the scientific challenge on anti-microbial resistance on JOGL]

Sanofi France, the largest pharmaceutical company based in France. A leader in the creation of vaccination solutions.

Diane Brement [Assist in the conduct of the Co-Immune open research program on vaccination]

Details of proposal – vision, aims and influence on open research

Visions

We wish to facilitate the conduct of open research and health innovation projects through a virtual environment that (1) encourages collective intelligence within a community of collaborators and (2) promotes the continuous publication of results and observations within an open and citable laboratory notebook.

Technological issues of the project

The technological challenges are at several levels:

Automatic distribution of actions of scientific interest

Centralization of data flows of scientific interest

Indexation of professional and amateur actors of a scientific issue

Activation of a community of relevant actors on a given task

Creating information flows that are customized within a news feed

Project deliverables

Implement within the JOGL platform a collaborative documentation module (lab notebook) of results and observations associated with the profiles of their authors

Make these open labs note books citable

Test this functionality with the Co-Immune open research program (collaboration between JOGL and Sanofi France) and with the CRI-Paris Fellows' open health research projects (TB Antibiotic resistance, Network medicine...).

Target audiences

We use this tool mainly for the following categories of people:

Researchers

Health professionals

Students

Patients

Amateurs

Long term goals
The JOGL project addresses the issues of resource utilization, redundancy and cooperation in the world of research. The JOGL project makes it possible to reduce the management costs associated with the complexity of massive digital assemblies by offering a tool for observing and facilitating large-scale collective intelligence phenomena.
go beyond quarrels and questions of ego, individual recognition, interpersonal or organizational competition, the need for psychological security, and more generally any psychological obstacle to the collective manifestation of intelligences;
overcome material obstacles to the instantaneous, rapid, fluid, relevant, scalable transmission of information, knowledge, ideas or innovations between people, whether they are in very small startup teams or gathered in structures of several hundred thousand people;
allow members of the same multitude to be guided and informed in real time of the people with whom it seems relevant to exchange immediately to allow this sharing of intelligence.
The main technological challenge of the JOGL project is to create a database containing results that are incorruptibly certified. This makes it possible to ensure the certification of any micro-contribution: such and such a user has produced such and such a result on such and such a date, and this in an inviolable way. In order to accomplish this task, it is necessary to combine three existing technologies:
modification tracking (git type)
peer-to-peer file sharing
certification technology
The technological challenge is to interlock these three technologies automatically and frictionlessly.
The expected impacts are:
Improvement of large-scale academic health research.
Fluidification of scientific collaboration between health research laboratories
Increasing the effectiveness of large-scale scientific collaborations
Increased permeability between the academic and private health sectors (patient communities, amateurs and NGOs).
The sharing of common resources and results allows for greater transparency and reproducibility of scientific productions (positive and negative results)

Details of proposal – evaluation plan

Through this project, we aim to implement within the JOGL platform the ability for authors to document their results and observations in a collaborative, citable, commentable manner. The monitoring of the success of this implementation will be conducted through mixed methods, both at the quantitative level of platform usage analytics, and at the level of questionnaires aimed at focus groups. Our key monitoring features will be:
Number of users. We aim to reach 300+ users and 20+ projects by the end of 2019, and 3000+ users and 200+ projects by the end of 2020
Frequency of usage of documentation feature. We aim to have a few (5+) highly active projects with daily use by end 2019.
Nature of the content: is the feature used to document in-depth, detailed work processes? This will be monitored through content size (expectations 5 sentences or more by post).
Number of citations of the items. Are they internal to the platform? External? We aim to have a few posts (20+) with high citation levels (50+), including external citations (10+)
Engagement in terms of number of comments: are commenters out of the first circle of contacts of the author (designed serendipity)? Do they have a different background (interdisciplinarity)?
Usability and utility of the developed feature: questionnaire to users after 3 months of usage.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal with a strong evaluation plan. The proposal would have benefited from more detail on how uptake would be encouraged, and concerns were raised about the feasibility of the project.

Title

Improving the Accessibility of Evidence Using Evidence Maps

Lead Applicant

Mrs Jessica Meeker

Details of proposal – team members and collaborators

Dr Louise Clark (Monitoring, Evaluation and Learning Manager, IDS) will lead the project, providing methodological guidance, strategic oversight and key input into the monitoring and evaluation.
Alan Stanley (Digital Knowledge Manager, IDS) will lead on the digitisation of the map and apply a user centred approach to the development of the map and the uptake.
Jessica Meeker (Knowledge Officer, IDS) will provide research support, reviewing the literature, assessing the quality of evidence, applying the mapping methodology and research uptake.
Alice Webb (Digital Knowledge Coordinator, IDS) will support the digitisation of the map and research uptake.
Simon Colmer (Developer, Independent consultant) will lead on the development and build of the web-based application for the map.

Details of proposal – vision, aims and influence on open research

Vision: To significantly contribute to improved understanding of ways to improve the health of children in LMIC countries through better infant and young child feeding practices. We will do this by increasing access and availability of quality assured research through the development of an evidence gap map.
Approach: This proposal seeks to develop an innovative but intuitive tool that will promote evidence use by making research findings more readily accessible to a wide range of users. Evidence maps are an emerging tool for evidence-informed decision-making and strategic prioritisation. By visualising existing evidence and rating its quality, it enables users to determine how much confidence they can have in an individual study and identify where the gaps in knowledge are. We plan to apply this methodology to a new body of data that has been derived from a recent IDS-led scoping review, focusing on gathering the latest evidence for behaviour change communication strategies within social protection programmes to improve infant and young child feeding.
Map design methodology: We propose to deliver the map as a web-based application that draws directly on the underlying data and that can be simply updated when new studies are added. The visual presentation of the map will be based on the 'traditional' gap map design used by 3ie, IRC and others, but we will further develop the basic map usability to better suit our target audiences.
Target audiences: The evidence map would be relevant for those working within research, policy and practice in health and nutrition, social protection and early childhood development. The map would enable researchers and practitioners to see a visual representation of the data in a way which makes it easy for the user to identify relevant evidence (or the lack of it), understand the strength of the body of evidence available, and quickly access the available knowledge base of documents.
In addition to sector strengthening, the map would also be used to build the understanding and capacity of researchers and practitioners. We will develop a targeted plan for dissemination which will include the map being used as part of a toolkit that participants receive when attending the IDS Centre for Social Protection short course. This well-established course provides a unique opportunity for policymakers, practitioners and researchers to broaden their knowledge base and gain critical insight into the most recent thinking about social protection.
The map would also encourage greater evidence-based decision making, providing those working in programme design and implementation with a better understanding of 'what works' and encourage decisions to be based on the latest evidence.
Software development and licensing: The evidence map will be a web-based application using a server-side scripting language, PHP, which can be embedded into any website or intranet.

We will ensure sustainability by using existing content management systems and will create a simple WordPress website as the home for the map.
To minimise maintenance requirements, we will use a very simple architecture whereby the application draws on data contained in a spreadsheet to generate the map. This could be a dynamic relationship (whereby the map is updated automatically when changes are made to the underlying data).
IDS is committed to Open Source software development, so any code produced would be made available for re-use under an Open License and accessible via GitHub with accompanying documentation.
How does this promote open research? This would be a technical and content driven project aimed to develop new ways to share and build a bank of evidence. This project builds on IDS's commitment to Open Access publishing and seeks to broaden perspectives by incorporating evidence from a wide range of sources. The programme will support increased access and availability of research, but it is also hoped that through the development of this open source evidence map, users will identify the gaps, new evidence will be developed, and the body of evidence will grow over time.

Details of proposal – evaluation plan

We will draw on the high-quality evaluative methods and infrastructure that the IDS community can provide. Suggested outcomes and indicators could include:
Outcome 1: Increased availability and accessibility of the profiled evidence.
Indicators and metrics: Increased distribution of evidence to target audiences – downloads of research outputs.
Outcome 2: Policy and practitioner communities are better informed.
Indicators and metrics: Number of policymakers/ practitioners engaged with, number of testimonies providing examples that audiences value the evidence.
Outcome 3: Academic thinking is built – new research areas identified.
Indicators and metrics: Interviews with social protection course attendees demonstrates the value of the evidence and new opportunities for new research or collaborations.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an interesting proposal with clear methodological approach. However, the level of innovation as well as the potential impact was felt to be limited.

Title

Brainpower – The Irish Brain Injury Research Network

Lead Applicant

Ms Grainne McGettrick

Details of proposal – team members and collaborators

Grainne McGettrick, Policy and Research Manager, Acquired Brain Injury Ireland will lead on the establishment of a brain injury research network. This network will be disseminating research from both home and abroad. A wide group of experts will be part of this collaboration.

These experts who will work on this research project include front line staff such as Professor Mark Delargy, Consultant in Rehabilitation Medicine from Irelands National Rehabilitation Hospital, both Dr. Phil O'Halloran, Neurosurgeon and Ciara O'Rourke, Clinical Nurse Specialist (Traumatic Brain Injury) from Beaumont Hospital, Dublin.

Another brain injury NGO (Headway) will work alongside Acquired Brain Injury Ireland. Their experts include Sonya Gallagher, Head of Rehabilitation Services and Richard Sables, Head of Information and Support Services.

Finally, we have gathered academic experts from two of the top universities in Ireland. Professor Anthony Staines (population health) and Dr. Catherine Corrigan (Nursing studies) Dublin City University with Assistant Professor Dominic Trepel, Health Economist, Global Brain Health Institute, Trinity College Dublin.

For this project we have gathered the top experts in Brain injury across Ireland to ensure the best research results.

Details of proposal – vision, aims and influence on open research

Vision

To grow the brain injury research community in Ireland to promote greater awareness and engagement in brain injury research in all its guises.

Aims

1. To increase the availability of longitudinal data, and access to that data.
2. To disseminate research from both home and abroad, within the network and on behalf of the network. This could include at key national and international events, such as those run by the International Brain Injury Association and the American 3. Congress of Rehabilitation Medicine.

To ensure that the views of brain injury survivors remain absolutely central to research.

To encourage thematic research that aligns with what's most needed.

To foster collaboration on brain injury research that spans different disciplines and communities. This collaboration could extend to international projects and research consortia.

Target Audience

Clinicians, allied health professionals, researchers, academics, students, families and people with brain injury.

Activities of the Brain Power Network:

1. Data: An online portal of brain injury research, including links to other existing portals, such as the NRH Research Register.
2. Dissemination: A periodic e-newsletter, that spotlights upcoming, ongoing and recently published research and events. The highlight of the calendar year will be an annual ABI Ireland - hosted seminar.
3. Lived Experience Input: A network steering committee that regularly seeks input from people living with a brain injury so that they can remain central to decisions taking in respect to the network.
4. Thematic Research: A number of working groups dedicated to various 'themes' of brain injury research. These groups will be open for members to join, can host meetings, trainings, provide resources, and can help in the development of briefs for prospective researchers.
5. Collaboration: A member directory, and workshops / events to build rapport, cooperation and culture.

Proposal influence on open research practices

The Brain Power Network would increase the visibility of brain injury research and create a platform for dissemination for all the members. Public and patient engagement (PPI) would also be a key feature. The Network would promote all forms of brain injury research. It would create accessible content for a wide range of audiences and ensure it is openly available in a range of formats. It will be multi-disciplinary in nature and will promote cross fertilisation of ideas and inter-disciplinary engagement on project proposals, funding opportunities and knowledge transfer projects.

Details of proposal – evaluation plan

Monitoring and evaluation:

The Brain Power Network will embed monitoring and evaluation as a key part of its work plan.

It will monitor and evaluate using qualitative and quantitative techniques including:

No of members (Target = 80-100 individuals, 20 institutions/organisations)

No of events (3 regional based events and one large national event))

No of people participating in events (N=250)

Website analytics(unique users N=300) and data on usage of web portal (active users 80-100)

Social media engagement (Utilising online analytic tools)

Members will be asked to evaluate qualitatively their experiences of the network at 6 month and 12 month intervals. Each event will utilise a feedback form to ensure that there is continuous improvement and refinement of the activity to meet the needs of the members.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an ambitious proposal to grow the brain injury research community in Ireland. The potential impact of this proposal to transform health research through openness was considered limited.

Title

Virtual Reality Body Mapping: A Chronic Pain Experience Visualisation Tool for Patients, Clinicians and Researchers

Lead Applicant

Ms Sarah Ticho

Details of proposal – team members and collaborators

Sarah Ticho is an an experienced professional in the field of VR, arts and health. Founder of Hatsumi she has consulted and worked with organisations including Nesta, Immerse UK and Stanford University. She is the Product Owner and will oversee project management, ensuring the research maintains strong methodological underpinnings.

Keisuke Suzuki, Sackler Centre for Consciousness Science (SCCS), University of Sussex. Keisuke specialises in Cognitive Neuroscience, and uses Virtual Reality and Machine Learning techniques to study human consciousness. Keisuke will oversee research and clinical rigour in the project.

Edward Silverton is cofounder of Mnemoscene and the lead developer of the Universal Viewer, a leading open source viewer for cultural heritage content, which has been adopted by Duke University as part of their Morphosource platform to display 3D biometric data. Edward will lead software engineering for the project.

Nicolas Slack is an experienced software engineer, specialising in Optimisation and 3D graphics. Executive Technical Consultant with Metasonics, VR Craftworks and Mnemoscene, and was formerly a Graphic software developer and guest lecturer at the University of Sussex. Nick will support on software engineering.

Clare Plumley is an artist, and educator researching chronic pain and arts in health. Clare will provide insight into the lived experience of chronic pain and advise on drawing tool capability

Details of proposal – vision, aims and influence on open research

43% of the UK population have or will experience chronic pain, with 30.8 million working days lost annually to such health complications. The NHS England has identified harnessing technology and expanding access to digital services as a way to help reduce costs, and aims to go paperless by 2020.

This funding will enable us to build a prototype body mapping experience in WebVR, to enable people with chronic pain to illustrate subjective lived experience using 3D drawing tools. This project builds on the success of a previous prototype developed in Unity, existing at the intersection of participatory art, research and healthcare. An extension of an established participatory arts health research method known as “body mapping”, (DeJager et al, 2016), the process involves tracing around a person's body to create a life-sized outline, which is filled in during a creative and reflective process, producing an image revealing new forms of insight into often indescribable, immeasurable, and subjective forms of lived experience.

The team plans to leverage “Aleph”, an open source 3D volumetric viewer and biometric data delivery toolset, recently developed as part of the morphosource.org project in a collaboration between Duke University, Penn State University, and Mnemoscene Ltd. The architecture of these components is unique in both supported file formats and the approach to annotation using Mozilla’s open source A-frame WebVR framework. This will offer a significant improvement over our existing Unity solution in that our application will be accessible by anyone with a web browser, thus removing the overhead of distributing to multiple app stores, and drawings created by users will be persisted to a library-grade digital asset management system.

We will use a standardised pain survey - the McGill Pain Questionnaire - as the foundation of our system. A collection of 3D drawing tools will be available to annotate directly onto a lifesize avatar which scales to match the users’ height in VR. The captured 3D drawing data will support comparative quantitative and qualitative analysis (location, degree of pain conveyed by brightness/opacity, type of pain conveyed by colour/texture).

The final outcome will be a proof of concept website that can be accessed via WebVR. A repository of illustrations will be created in a one day user testing workshop, constituting an

anonymised formal dataset that will allow for further subsequent analysis. By having users map their pain onto a common 3D template, we can begin to quantitatively compare their subjective experiences (e.g., by comparing the relative areas indicated as intense pain, or coloured with blue versus red).

In the future, this data will contribute to a greater body of research in collaboration with The University of Sussex, Sackler Centre of Consciousness. The NHS Recovery College have expressed interest in a partnership to create body mapping workshops with their chronic pain user group. Extracting data from user-annotations of community-accessible 3D models can be seen as potentially transformative for bioinformatics that rely on pairing large phenotypic datasets with genomic data. To date, no scientific resource has developed tools for translating 3D annotations into machine-readable data. This would be the first attempt to do so in a widely supported, interoperable framework. We intend to continue collaborating and consulting with MorphoSource as we develop this project to ensure it fits their use cases, and are confident that the platform we build will be robust and applicable to a broad range of disciplines.

Our ambition is to generate greater awareness of the benefits of Body Mapping in a therapeutic context and to build a foundation from which we can further develop and design other 'palettes' of lived experience with different groups e.g. anxiety, depression.

See a video of the existing early proof of concept here: https://youtu.be/gxJ8yhrO_T4

Details of proposal – evaluation plan

All development activity will take place over 8 weeks, working to two-week sprints. The team will hold meetings at the beginning and end of each sprint, involving all project stakeholders to plan, review and feedback on progress. Working to an Agile method, these structured meetings will ensure that arising issues can be effectively communicated and managed. The majority of activity will take place at The Fusebox, a VR co-working space in Brighton where the development team are based.

Sprint 1

- Write User Stories.
- Backlog elaboration and refinement.
- Set up project management system and development environment.
- Fix outstanding minor bugs on the existing prototype.

Sprint 2

- Develop screen capture mechanism for use in a website gallery.
- Develop freehand drawing capability.

Sprint 3

- Develop database and data preservation/retrieval system.
- Develop menu system to enable effective drawing tool selection.

Sprint 4

- Integrate the creative treatment with artist.
- Workshop with NHS Recovery College.
- Final review, evaluation and planning next steps.

By the end of the development period, the team intends to have successfully developed a functional 3D annotation and archival system will be ready to trial during a workshop with the NHS Recovery College, with up to 12 service users. We will gather feedback from the user group on functionality and efficacy of the system, and document their findings as a project deliverable. Suggestions for amendments will be noted for future developments. With permission from participants, an exhibition of their work will be developed to translate knowledge on the experience of chronic pain. This will be disseminated via the teams networks, websites, social media channels and newsletters.

The success of the programme will also be determined by its interoperability and value to a wider pool of researchers and developers. Through our ongoing collaboration with Morphosource, we will gain ongoing feedback and advice.

Decision Not shortlisted
Comment on decision from Wellcome This was an interesting proposal to create an open virtual reality platform which would allow illustration of lived experience of chronic pain. However the potential of the proposed activities to transform health research was unclear and so the proposal w

Title

Practical steps to improve metadata quality and data sharing in circadian research

Lead Applicant

Prof Andrew Millar

Details of proposal – team members and collaborators

Prof Andrew J. Millar FRS FRSE will lead the project. He leads the Open Research strand of our Wellcome ISSF3 funding and senior user representative to the University's Research Data Service. Dr Tomasz Zielinski (School of Biological Sciences) is a senior research software developer, lead architect of BioDare2, and will supervise the research software engineer funded by this award. Dr Niki Vermeulen (Science, Technology and Innovation Studies) will supervise the evaluation of user attitudes. She specialises in science and innovation policy and the social organisation of research.

Details of proposal – vision, aims and influence on open research

(i) Vision. Making data FAIR and Open still encounters social barriers, in the form of perceived costs and perceived risks. The BioDare2 resource has overcome the perceived cost of time taken to share data but some researchers still perceive a competitive risk from sharing data. Quality data sharing, with quality metadata, requires a change in research culture to create endogenous motivation for sharing rather than an external mandate. We will implement and evaluate low-cost, transferable methods to improve metadata quality and attitudes to sharing, in BioDare2's established, international community of users in circadian biology. This community crosses many biomedical research fields, giving us broad potential influence.

Target audience. Daily, circadian rhythms are a fundamental property of cellular regulation, acknowledged by a 2017 Nobel Prize. Circadian research touches many fields of biomedicine, and our users work from parasitology, to immunology, metabolism and mental health. Circadian timeseries over several days are costly to obtain, and mathematical analysis is required to measure the timing features in the data. Analysis of and access to timeseries data are therefore critical, across the diverse, rhythm research community.

BioDare2 (https://biodare2.ed.ac.uk) is the only public, online repository for circadian timeseries. BioDare2 provides fast timing analysis, summary statistics and attractive data visualizations in a modern web interface, all to ensure the best user experience. Access to BioDare2 is provided on condition that data will be made public. Thus Open data sharing is a "side effect" of using our analysis tools and is not perceived as an additional burden. On the contrary, BioDare2 increases research productivity as its analysis methods are faster and easier to use than any alternative. This approach has successfully attracted data from users, gaining over 340 000 data timeseries in the two years since inception, with continuing exponential growth.

Proposed Activities. Successful data re-use depends on good-quality metadata. We have empirical evidence that enforcing strict metadata standards can not only impede the uptake of our service but also triggers undesired behaviours, for example duplicating previous experimental descriptions or entering random characters as metadata. Human data curation is a possible solution but has not been achievable with current funding, a common problem for community resources.

With over 500 active user sessions per month, we can perform controlled experiments on the effectiveness of alternative methods to influence user behaviour and promote reliable data sharing. BioDare2 is therefore uniquely positioned to allow practical, quantitative evaluation of these approaches:

(1) Automatic metrics of metadata quality will be derived from online text analysis, completion of fields, and similarity to existing records. These provide immediate feedback to users, during data upload.

(2) Sharing badge scheme: the user and their dataset will receive badges depending on the user's openness and the quality of their metadata, based on automatic metrics (1) or human "triage" (3) Badges could convey both positive and negative messages.

(3). Human “triage” to rank metadata quality: users will be asked to score the quality of metadata, comparing fabricated descriptions and real content. Peer rankings contribute to many online communities, such as the successful, question-review process in the software resource Stackoverflow. Rankings can help to form, and to educate users about, the community’s views of good practice.
(4) Functional rewards for richer metadata or earlier data release, where users gain access to new visualization or analysis methods. This approach can estimate the user’s acceptable level of openness or metadata effort, in terms of the level of benefits required to engender each behaviour.
(5) Questionnaires during the login process will evaluate user experiences and attitudes towards Open research at the start and end of the project, and can also provide information and some training.
(ii) Influencing Open practices. These measures directly improve metadata quality and Openness within BioDare2, reaching users across multiple biomedical fields. Moreover, our peer ranking process could lead towards community-based data curation in future.
The proposed evaluation of various online techniques for changing research culture is unprecedented to our knowledge. The results will provide invaluable input to the Open strategies of community resources that face similar funding pressure, far beyond the circadian community.

Details of proposal – evaluation plan

Evaluation is a major component of the proposal. Few, if any, biomedical resources provide data on metadata quality, so the basis for comparison to other resources is currently limited. Our results will be publicly released, to facilitate such comparisons in future. Our targets are:
Automated evaluation of metadata quality for all datasets.
The quality of metadata will be measured using automatic metrics developed in Activity (1), validated by manual scoring of a sample of datasets. Automated feedback to users (Activity 2) and peer evaluation (Activity 3) will encourage culture change towards better metadata.
Improved metadata quality on average, as defined by these metrics.
In particular, we will eradicate the lowest-quality (meaningless or duplicated) descriptions of experimental data.
Improved attitudes towards Open research; reduced perception of risks associated with Openness.
User attitudes will be measures by questionnaires before and after the intervention, designed based on our experience of comparable studies. We will compare the attitude change in our users with contemporary literature evaluations of the biomedical research community. Collecting broad baseline data is beyond the scope of this focussed proposal.
Improved Openness in practice, targeting voluntary early release of 20% of datasets (note that all data become Open by default).
Metrics of early release, and dis-aggregated measures to identify which groups have engaged, will be automatically calculated by the resource. We expect increased re-use of the data in the medium term; we will collate successful examples but we do not expect sufficient data to evaluate robustly within a 12-month project.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Enabling clinical decision support tools through the Virtual Metabolic Human knowledge base

Lead Applicant

Prof Ines Thiele

Details of proposal – team members and collaborators

Prof. Ines Thiele, Molecular Systems Physiology Group, National University of Ireland, Galway(NUIG). Ines will be the principal investigator andsupervise all activities related to this proposal.The research group of Ines has developed the Virtual Metabolic Human knowledge base (VMH, www.vmh.life)for the systems biology community. Her research focuses on building comprehensive computational models of human and gut microbial metabolism, which are at the core of the VMH, and oninvestigatinghow diet influences human health, with emphasis on Parkinson's disease(PD).

Dr Cyrille Thinnes, Molecular Systems Physiology Group, NUIG.Cyrille will establish and implement the user engagement strategy. He will overview the monitoring and dissemination, and supervise the organisation of workshops, tutorials, and communications.

Dr Almut Heinken, Dr Johannes Hertel, and Dr Dimitri Ravcheev, Molecular Systems Physiology Group, NUIG. Almut, Johannes, and Dimitri will collect and generatethe novel VMH content.

Prof. Ronan Fleming, Systems Biochemistry Group, Leiden University. As the coordinator of the Horizon 2020 funded project "Systems Medicine of Mitochondrial Parkinson's Disease (SysMedPD)"consortium, Ronan will provide access tothisconsortium, whichincludes prominent Parkinson's disease clinicians and experts (http://sysmedpd.eu/participants/). Ronan is an expert in computationaland mathematicalmodellingappliedto Parkinson's disease.

Details of proposal – vision, aims and influence on open research

Ourvision is to transformthe Virtual Metabolic Human knowledge base (VMH, www.vmh.life) into the primary online resource forhealthcare professionals,therebyenabling future developments ofclinical decision supportsystems.Specifically, we aim to 1. expand the VMH's target audience to physicians, with emphasison neurologists, 2. lower the barriers of entry for adoptioninto the clinical routine, and 3. foster user retention.

The freely available VMH knowledge base is an interdisciplinary, comprehensive, open source online resource developed using manually curated information from the scientific literature.The VMH provide detaileddescriptions, via dedicated entry pages, for5,607 unique metabolites, 19,313 unique reactions, 3,695 human genes, 255 Mendelian diseases, 818 gut microbes, and 8,790 food items.Each of these entries iscross-referencedallowing for complexqueries (e.g., which reactions are associated with a particular disease, and arethesereactionsalso in gut microbes).The VMH connects its entities to 57 other web resources making, including disease and clinical trial resources,it an ideal starting point for biomedical research.The metabolic networksunderlying the VMH are accessible, e.g., viaa road map-like interfacethroughwhich experimental data can be visualised. Also, a Leighsyndrome-specific gene-to-phenotype map permits to identify reported phenotypes with known genetic mutations.

To achieve our overarching goal of transforming the VMH into a tool actively used by the clinical and healthcare (research) community, we aim to

1. Proactively engage with the clinicalresearch communityat NUIGand SysMedPDto identify crucialclinical and medical informationand web resourcesto be addedtothe VMH to make it a powerful tool for clinical decision support ultimately.Therefore, via email and social media, we will contact collaborating neurologists, medical students, and research clinicians to create an active working group. We will ensure to collect information in sucha modularmanner that further expansions beyond PDcanbe easily done after the end of the proposed project.

2. Lower the barrier of entry fortheVMH adoption by healthcare professionalsthrough:

(a) Development of a novel VMH interface designdedicated to healthcare professionals by surveying their needsand trackinginterfaceinteractions. Particularly,relevantcontent will be

highlighted for 'one-click' interactions for most efficient access. The design efforts will be supported through the consultancy of a weband softwaredesigner.
(b) Developments of a set offree online tutorialsspecifically targeted towards PD healthcare users with dedicated sample applications.
(c) Organization ofworkshops to provide personalised guidance for the VMH applications.
(d) Leverage ofuser experience information (interface interaction, feedback) to iteratively facilitate user interaction, and therefore continuously lower the barriers of entry for healthcare community adoption.
3. Foster user retention, whichisclosely related to user adoptionand engagement. Therefore, we will
(a) Feature community membersand users, e.g., viavideo interviews highlighting thought leaders and use cases.
(b) Expandour social media presence for content dissemination(articles, pictures, video) to the healthcare community.
(c)Create personal userspaceswithinthe VMHenvironment byenabling, e.g., saving of individual information,such assearches, medical data, and configurations for connectingto computational analysis tools, such as the open-access Constraint-Based Reconstruction and Analysis (COBRA) Toolbox, led by Prof. Ronan Fleming.
We chose PDas a demonstration and implementation case for the VMH transition to highlightto clinical research communitythe increasing evidence thatPD also hasstrong metabolic and microbial componentscontributingtodisease aetiology and treatment response.
The VMH continues its strong commitment to open research practices and strives to lead by example. The VMH is freely accessible through the online interface(www.vmh.life),and all data can be freely downloadedvia the website and through an applicationprogramming interface (API).The VMH wasvisited more than 90k times since March 2013 and has a growing group of returning users.Theaccompanying scientific publications are open access, and preprints are postedon bioRciv. Educational material, including workshops and videos, are provided freeon a dedicated YouTubechannel.
By expanding and diversifying the user base from the systems biology to the healthcare community, we will promote open science via the network effect. This community diversification shall be facilitated by the considerationthat the VMH is relevant to anaudience interested in the biomedical sciences.

Details of proposal – evaluation plan

We are committed to monitor and quantify the VMH user experience to the best of our means, using.
1.Website analytics, e.g., Google Analytics,to analyse and segment our user base including, e.g., accession numbers, provenance, workflow, and engagement. The VMH was visited by >20,000 visitors in 7 months, from >100 countries. For this project, we would like to double the accession rate at least.
2.Online user feedback, e.g., Doorbell,asanintegrated user feedback function. Wetarget an at least doubled rate of feedback, predominantly during workshops.
3.Social media analytics, e.g., YouTube Studio, and similar integrated social media analytics tools, to enable the quantification of user engagement. Whileour social media posts are visible (in the hundreds per month), user engagement (sharing, commenting), is relatively low (tens per month). We targetat least double the number of views and engagements.
4.Feedback from workshops, e.g., using Mentimeter,asan online polling platform for real-time interactive audience participation. We will aim to collect constructive feedback by comparing short surveys immediately before/after our workshops, to better understand the needs of our target users, to improvefuture engagements.
5.Expertinteraction, e.g., through interviews.We will interview specific target users to better cater to their needs. We shall interview select practising physicians, nurses, and nutritionists, to

improve the VMH user interface foraddressingtheir use cases. We will aim to interview at least tenrepresentatives during the initial stages of this proposed project. By design, all VMH platformmodifications are implemented in real time, i.e. the usersexperience new functionalities as soon as our VMH team has approved them. In combination with the integrated feedback functions, we strive for the most efficient creation-feedback iterations to present users with the most recent advancements as fast as possible
Decision Not shortlisted
Comment on decision from Wellcome This proposal sought to expand an existing online resource with good commitment to openness. However it was felt that the proposal would have benefited from some understanding of the needs of healthcare professionals already, and so the level of demand wa

Title **Recovering lost research: creating an open index of conference papers and presentations**
Lead Applicant **Prof James Thomas**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Shortlisted, not funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title **TyphiNET – Unlocking public health lab data on travel-associated typhoid for sentinel surveillance**
Lead Applicant **Prof Kathryn Holt**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Scienceverse: Towards a Grammar of Science

Lead Applicant

Prof Lisa DeBruine

Details of proposal – team members and collaborators

All team members will contribute to the development of the grammar.

Lisa DeBruine (University of Glasgow, UK) will coordinate the development of the tools and database, working closely with the postdoctoral researcher. She has extensive experience with online tool and database development (e.g., ERC-funded development of webmorph.org).

Daniël Lakens (Eindhoven University of Technology, NL) will coordinate the creation of tutorial materials and best practice documents. He has extensive experience with communication of best practices in methods (e.g., >27K students enrolled in Improving your Statistical Inferences).

Gjalt-Jorn Peters (Open Universiteit, NL) will coordinate initial testing of Scienceverse in Registered Reports at Health Psychology Bulletin, where he is an editor.

Details of proposal – vision, aims and influence on open research

Vision: The increasingly digital workflow in science has made it possible to share almost all aspects of the research cycle, from pre-registered analysis plans and study materials to the data and analysis code that produce the reported results. Although the growing availability of research output is a
positive development, most of this digital information is in a format that makes it difficult to find, access, and reuse. A major barrier is the lack of a framework to concisely describe every component of research in a machine-readable format: A grammar of science.

Details of proposal – evaluation plan

A grammar is a formal system of rules that allow users to generate lawful statements. The goal of a grammar of science is to allow users to generate rich, standardized metadata describing experiments, materials, data, code, and any other research components that scholars want to share. Such
standardization would facilitate reproducibility, cumulative science (e.g., meta-analysis) and reuse (e.g., finding datasets with specific measures). While many projects focus on making data FAIR, Scienceverse aims to make every aspect of research findable, accessible, interoperable and reusable.

Aims: Developing a Grammar of Science, combined with a shared lexicon (e.g., standardized ways to reference manipulations, measures, and variables) aims to facilitate open research practices for researchers and journals. It is intended to mitigate several well-known problems that follow from the lack of organization of research output.

First, it has been shown that even when data and code are shared, computational reproducibility is low (Hardwicke et al., 2018, Obels et al., 2019). Scienceverse improves computational reproducibility by providing a framework that explicitly links hypotheses, materials, data, and code. Scienceverse archive files can store any aspect of research in a systematic way, allowing, for example, automatic evaluation of results against machine-readable specifications of statistical hypotheses. Automated
reproducibility allows journals to compare pre-registered hypotheses with the conclusions in the final manuscript. Scienceverse helps researchers to specify which analyses would confirm or falsify predictions in a structured and unambiguous manner. Journals can automatically check these predictions for the final submission, which will prevent problems with undeclared deviations from the protocol – a known problem in pre-registered studies.

Second, Scienceverse aims to make shared outputs easier to find and re-use. Good meta-data are essential to find research output, but there have been few attempts in health psychology, or social sciences in general, to summarize the structure of those aspects of the empirical endeavour that need to be findable. Scienceverse aims to create a well-structured grammar that provides a complete description of these components of the research cycle, including hypotheses, materials,

methods, study design, measured variables, codebooks, analyses, and conclusions. Referenced against
discipline-specific lexicons, this allows researchers to retrieve any information from archive files. For example, researchers can search for studies that use similar manipulations and retrieve relevant information about the effects these manipulations produce. This information can be used when choosing manipulations for future studies, to design well-powered experiments, or to easily perform meta-analyses. Given specific inclusion criteria, Scienceverse makes it possible to automatically update
meta-analyses and share these with the scientific community.
Our target audience is researchers who want to increase the impact of their outputs by making them more findable and reusable, researchers who want to find and reuse others' outputs, journal editors who want increased clarity about study design, and contributors to methods tools who want to
increase interoperability of research components. Team members are holding a "hackathon" with potential users at the Society for the Improvement of Psychological Science (osf.io/c52yh) and presenting related ideas at the European Health Psychology Society (osf.io/ndxha).
Activities: In this project, we plan to 1) develop a grammar of scientific research in the social sciences, 2) create a tool to guide the creation of archive files describing studies using this grammar, 3) create an online database where archive files can be uploaded and searched, and 4) create and disseminate tutorials with concrete examples for health psychology.
Compared to current practices, where data is increasingly shared, but in a format that makes it practically unusable except with great effort, we hope Scienceverse will move open science forward by providing a structured framework to organize and find research output, which should make all shared research components more findable and reusable.

Decision

Not shortlisted

Comment on decision from Wellcome

This was an innovative proposal from a strong team. However, concerns were raised about feasibility and how a critical mass of users would be reached.

Title

Open-source software for separating low and high arousal states in EEG records.

Lead Applicant

Prof Peter Howell

Details of proposal – team members and collaborators

Guangting Mai, PhD student who developed initial package used in Mai, Schoof and Howell (2019).

Eryk Walczak ex PhD student and continues as research fellow in the group (data scientist and researcher). Tim Schoof, Jyrki Tuomainen – end users. Stuart Rosen, neuroscientist and highly experienced open-source software developer (uses Matlab and R). All have affiliations at UCL (Walczak is employed by the Bank of England).

Details of proposal – vision, aims and influence on open research

Aims: Electroencephalography (EEG) is widely used in research. The EEG signals are low in amplitude, hence procedures average over many trials leading to long test sessions. Consequently, arousal state varies within and across participants. Neural and behavioural performance differs between arousal states (Mai, Schoof & Howell, 2019) where arousal state was determined using a computational procedure based on occurrence of sleep spindles and time-frequency characteristics in EEG-epochs. We intend developing open source software in this project to make these procedures generally available. These open sources (R and Python versions) allow researchers to acquire accurate timing information and time-frequency characteristics of different arousal states for a wide range of applications.

Target audiences: This project aims to provide open source code and software for determining arousal state for researchers and clinicians in the EEG field. This should prove useful for EEG research in general and for studies in aging and in neonates in particular.

Activities: A custom-written Matlab program was used in Mai et al.'s analyses. The current project will develop software packages based on open source languages (R and Python) which are dominant languages in data science (https://insights.stackoverflow.com/survey/2019). R and Python allow software to be developed that is portable across operating systems. Literate programming, like programming notebooks, will be used to distribute code along with sample output and text describing what given lines of code achieve. Other open-source software packages in neuroscience rely on proprietary programming language (e.g. Matlab) whereas R and Python work well across operating systems and are free. Versions in both languages (R and Python) are included to make the code usable by researchers who use either language.

Initial loading of EEG files is hardware-specific (recordings are saved in different formats). Available open source software represents files obtained from different hardware in equivalent formats: e.g. eegUtils for R (https://github.com/craddm/eegUtils) and MNE for Python (https://martinos.org/mne/stable/index.html). The package for determining arousal state will be built on eegUtils and MNE which will perform the initial loading of EEG files. MNE has been established for some time and was developed by a groups of researchers whereas eegUtils is fairly new and was developed by a single researcher (https://staff.lincoln.ac.uk/mcraddock).

Vignettes (code examples with output and text comments; see https://github.com/craddm/eegUtils for a good example) and blog posts demonstrating illustrative examples will be used to demonstrate how to use the software. This is practiced in a limited way in commercial scientific products.

We will use a GPL licence which is a copyleft license that enforces usage of the same licence in any derivative works.

The open source package will be distributed through the official repositories (The Comprehensive R Archive Network, CRAN, for R, and pip for Python). These repositories only admit well-tested packages. Members of the team have experience preparing packages for CRAN. GitHub or GitLab are hosting services for code that will be used for effective development and dissemination of the package. These services were chosen as they have good standards, are easy to use and reliable.

https://ropensci.org/packages/ will be used as a not-for-profit code review (peer review for programs) and the package will be published in the Journal of Open Source Software: https://joss.theoj.org/ or Behavior Research Methods that permit tracking of citations.
Impact on field and more broadly:
The software will expedite re-analysis of data from many studies that have not controlled for arousal. Often such studies instruct participants to either stay awake or fall asleep, but do not monitor this. Such data can be analysed for sleep spindle occurrence and then used to test for any effect of arousal state. Important future applications are in aging research as arousal state fluctuates between age groups, when comparing EEG recordings of neonates (Carles Escera, University of Barcelona, personal communication) and arousal is a fundamental topic of study in neuroscientific research. Future versions of the software will be developed for real-time online monitoring of arousal states for use in broader neuroscience research (e.g. in closed-loop brain-computer interfaces, neurofeedback and interventions that deliver information in real-time during selected arousal states).
Mai, G., Schoof, T. & Howell, P. (2019). Modulation of phase-locked neural responses to speech during different arousal states is age-dependent. Neuroimage, 189, 734-744

Details of proposal – evaluation plan

Software success will be assessed by the number of downloads which can be automatically tracked for CRAN (R package repository) and pip (Python). The number of downloads are shown directly in the code repository. For example, https://github.com/alastairrushworth/inspectdf uses a package inspected for visualisation. The automatic badges (used in an R package) indicate quality of a project to software engineers. Potential badges are:
- passing automatic tests of continuous integration:
https://en.wikipedia.org/wiki/Continuous_integration
- percentage of code covered by automatic tests (to pick up errors)
- CRAN to show that the package is available in the official repository, and to provide package version
- download statistics from the official code repository (CRAN)
- passing all checks that allow software package to be on CRAN

These tests will be incorporated into our program structure using widely available metatools (e.g. devtools or testthat). The pipeline is more standardised than in proprietary languages because the entire environment is open source and clear guidelines and best practices are incorporated.
Other metrics that will be employed are number of watchers (people interested in development), stars (just like Facebook 'likes'), forks (copies made to develop own version), or contributors (people who make changes and push them to the code repository). These metrics are used in code repositories like GitHub.
A DOI will be obtained for the package and we will submit an article to an appropriate journal (e.g. Behavioural Research Methods).

Decision

Not shortlisted

Comment on decision from Wellcome

This was a good quaility application from a strong team. However, the level of impact was felt to be limited, and evidence of wider utility was lacking.

Title

Randomization-based time-locking of study pre-registration

Lead Applicant

Prof Roy Mukamel

Details of proposal – team members and collaborators

Matan Mazor, Wellcome Centre for Human Neuroimaging, University College London.
Noam Mazor, Blavatnik School of Computer Science, Tel Aviv University.
Roy Mukamel, School of Psychological Sciences and Sagol School of Neuroscience, Tel Aviv University

Matan, Noam and Roy will be responsible for setting up the online pre-registration time-locking tool, and will work with available pre-registration platforms towards incorporating this tool into their existing pipelines.

Matan will be responsible for coordinating the production of the animated visual guide.

Details of proposal – vision, aims and influence on open research

HARKING, or Hypothesizing After Results are Known, introduces circularity into the process of scientific inference and by that compromises the reproducibility of scientific findings. In recent years, pre-registration of study plans and hypotheses has grown in popularity in the life and cognitive sciences, as a way to clearly separate prior predictions from post-hoc conjectures, and as an objective means for researchers to demonstrate that indeed their hypothesis were made in advance. However, commonly used study registration platforms such as the open science framework can only provide a time stamp for registration rather than objective evidence that registration indeed preceded study commencement. We refer to this missing feature of review-free pre-registration as time-locking, and note that this feature is crucial for pre-registration to be used as an objective measure of hypothesis-driven research. Without time-locking, pre-registration is just registration.

We recently developed a simple cryptographic tool (pre-RNG) that provides an objective marker of time-locking by making random components of the experimental design contingent upon the contents of the study plans and hypothesis [1]. This tool is intended to help those researchers who want to have objective proof for the integrity of their pre-registration, but also don't want to take the Registered Reports route which introduces an additional peer-review stage before data acquisition [2], for example because they prefer not to commit to any journal in advance, or because they prefer to keep their research plans private before completing their analysis. Our tool is applicable to all research disciplines, but is specific to experimental designs that include random or pseudo-random features (such as the order and timing of experimental events), and that measure some continuous variable over space or time (such as brain activity or eye movements). This is the case in many studies in the life sciences, and specifically in neuroscience and neuroimaging, which is where we come from. Importantly, this tool is very easy to use, and does not necessitate the involvement of any external inspector, reviewer, or journal.

To date, pre-RNG is available as an offline application, and is also available online as part of the JATOS online experiment platform [3]. The current proposal is to implement pre-RNG as an independent web-based application, and integrate it to be part of review-free online pre-registration platforms, and primarily the widely used open-science framework (OSF), such that researchers will be given the option to time-lock their registration as part of the pre-registration process. This will be supported by (a) an online implementation of the time-locking algorithm, and (b) a user-friendly animated visual guide to the process or pre-registration time-locking, with a step-by-step illustration of the process. The OSF development team has expressed their interest in this project.

By making this tool more accessible to researchers, this project will sharpen the important distinction between hypothesis-testing and exploratory phases of the scientific process.

[1] Mazor, M., Mazor, N., & Mukamel, R. (2018). A novel tool for time-locking study plans to results. European Journal of Neuroscience.

[2] Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. Cortex, 49(3), 609-610.
[3] http://blog.jatos.org/Proof-Of-Preregistration-Via-Cryptographic-Hash/

Details of proposal – evaluation plan

The aim of this project is to disseminate the use of registration time-locking, and by that (a) facilitate the separation of confirmatory and exploratory results, and (b) provide an objective marker for such separation. We will set as our primary goal to make the use of our scheme as simple and intuitive as possible. This goal will determine our two metrics for success:
Adoption rate of our scheme. We aim to have 10% of new OSF pre-registrations incorporating our time-locking mechanism by February 2020. This means that by February 2021, we will aim to have 1000 time-locked OSF registrations.
Publication of research reports that were pre-registered using the pre-RNG scheme. We expect these projects to be published starting from 2021.
As support for the feasibility of our goals, we note that our scheme has been adopted by the JATOS online studies framework and by the Wellcome Centre for Human Neuroimaging at UCL, which has recently adopted our scheme and is offering it to its users as an optional form of pre-registration.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.

Title

Epirecipes: a cookbook of epidemiological models

Lead Applicant

Prof Simon Frost

Details of proposal – team members and collaborators

Prof. Simon Frost, University of Cambridge/The Alan Turing Institute (From October 2019: London School of Hygiene and Tropical Medicine & Microsoft Research)
Overall lead for project; initial developer of software platform.
Dr. Rosalind Eggo, London School of Hygiene and Tropical Medicine
Co-Investigator; co-mentor of software engineer and advisor on mathematical modelling.

Details of proposal – vision, aims and influence on open research

Background Mathematical modelling of infectious diseases has become an important scientific field, yielding insights into how pathogens spread and evolve while also becoming an essential tool for informing public health strategies. A rapid response to public health emergencies is assisted if there are 'off the shelf' models that can be quickly modified to the situation, and public health decisions are likely to be more robust if a variety of different models are used. There is also a need to train researchers, in order to help them understand the strengths and weaknesses of these models, which are used to guide government policy, as well as to expose students to a growing area of research.

Aims While there are several textbooks on modelling infectious diseases, there is a dearth of examples in the form of runnable computer code that could be used for modelling and data fitting. A collection of epidemiological models would represent an important resource in expanding expertise, as well as providing a reproducible platform that would allow researchers to examine the breadth of models in the published literature and to share models. Epirecipes (http://epirecip.es) is an online 'cookbook' of models of infectious disease intended to fill this gap.

Target audiences Epirecipes is a resource for both teaching and research, with multiple audiences:
Students: as a collection of models implemented in multiple computer languages, Epirecipes offers an excellent resource to familiarise oneself with the epidemiological modelling field.
Researchers: Epirecipes offers an online, reproducible framework that allows models to be run interactively, and so can serve as a 'live' repository of published models. As the repository grows, gaps in the field in terms of modelling needs may be identified.
Trainers: Epirecipes has already included several chapters from Ottar Bjornstad's recent 'Epidemics' book that allows one to run through the worked examples without installing any software other than a web browser.

Activities Epirecipes is already a working prototype, but needs further support in order to improve both the platform and the content.
Epirecipes is a 'Jupyter book', comprised of a number of Jupyter notebooks. Other platforms (e.g. R Markdown) are also widely used, and currently are converted using a semi-automated pipeline. We would like to improve accessibility by accepting a wider range of model formats.
Outside of hackathons, models can be added by anyone that has a GitHub account, but this requires a certain degree of familiarity with the process of dealing with git repositories. We would like to simplify the submission process to provide a simple upload interface along with some basic checks on the formatting, that the models run as desired, etc..
The free, public-facing Epirecipes site uses a free hosting service to launch models, which limits the time and resources that can be used to run a model, and is not suited to saving modified models. Previously, we used a large server for a 3 day 'hackathon', but this is not cost efficient in the long term. We would like to set up a 'one-click' mechanism to set up a server for Epirecipes that would greatly simplify deployment in a teaching setting.
There is still a considerable backlog of models to be implemented, both from the first Epirecipes meeting held in October 2018, as well as those identified in an online 'shopping list' of models (https://github.com/epirecipes/shopping-list/issues). We would like further resources to curate

these models for the cookbook, as well as incorporate additional examples via requests to the authors of published papers.
Influencing open research practices Mathematical models are often regarded as 'reproducible', as many are in the form of sets of differential equations that can be unambiguously interpreted. However, in practice, such models are not easily reproducible. Sometimes there are simple omissions or typographic errors in parameter values; sometimes exotic numerical schemes are required in order to simulate or fit a model; and increasingly, models are complex simulations, where the only unambiguous description is the computer code itself. By providing a platform with a variety of models in a variety of languages, where models can be launched with a single button click, we hope to encourage modellers to deposit their code, allowing it to be run, modified, etc..

Details of proposal – evaluation plan

The code for the platform is hosted on GitHub (http://github.com/epirecipes), and as a 'hard' outcome, there are a number of criteria that can easily be monitored and used to evaluate the success of the project, including the number of examples included, the number of versions of the model in different computer languages, the number of platforms supported, and the scalability of the platform. There are currently over 30 different models in the platform, with a comparable number still to be curated and added, with 1-11 different implementations of each model. Through the work of a full-time software engineer, plus a second hackathon, our goal over the course of the project would be to add at least double the number of models, and ensure that each has at least two implementations. R and Python were the most popular computer languages used in the first hackathon, with Julia running a close third.

Decision

Not shortlisted

Comment on decision from Wellcome

This was felt to be an interesting and feasible idea. However, it was not clear how extensive the user-base would be and the level of innovation proposed was considered limited.

Title

Open Research Knowledge Graph for the Web Editor Software Community

Lead Applicant

Simon Worthington

Details of proposal – team members and collaborators

Simon Worthington, Open Science Lab, TIB – German National Library of Science and Technology, Hannover, Germany. Board member FORCE11. Project coordinator, responsible for project management; ADA microservice implementation coordination and delivery. Expertise in FOSS publishing software.

Johannes Wilm, Fidus Writer (LUND INFO AB), Sweden. Founder, director, and software developer of Fidus Writer. Responsible for Fidus Writer plugin software development. Overseen previous semantic markup additions to Fidus Writer.

Maria-Esther Vidal, Scientific Data Management, TIB – German National Library of Science and Technology. Facilitate contact with medical research groups from the following research Horizon 2020 projects — iASiS, and BigMedilytics. Additionally providing domain expertise in ORKG and knowledge graph technology.

Marcus Stocker, Co-Lead – Open Research Knowledge Graph, TIB – German National Library of Science and Technology. Coordinate ORKG components of the project. Providing domain expertise in ORKG and knowledge graph technologies and provide liaison to ORKG research team.

Johannes Amorosa, Developer, Endocode AG, Berlin, Germany. Research advisor on microservices and cloud virtualisation.

Details of proposal – vision, aims and influence on open research

The vision of the project is to make medical research papers machine interpretable in Alzheimer's research papers and beyond for other medical fields by introducing the use of Open Research Knowledge Graph (ORKG) technology to the community of FOSS scholarly web editor software developers. We will engage this community by creating an exemplar FOSS pilot implementation of ORKG in the Fidus Writer scholarly web editor, with high quality documentation, and providing a ORKG

microservice for other editor platforms to access. ORKG works by makes a granular RDF semantic representation of a research paper which combines general and optional domain specific ontologies. General components in a paper may include: methods, experiments, and charts, etc. Domain specific ontologies, as for example in this project for Alzheimer's informed by the iASiS (a medical big data project) unified schema. The effect is to transform a paper into a set of data objects for use in the open scholarly data ecology to increasing visibility and reuse of research.

We are targeting the 'authoring of papers' part of the scholarly communications workflow for adding ORKG semantic enrichment. The experience of the Open Science Lab at TIB is that researchers are incentivized to become Open Scientists when there are benefits for them personally and that this needs to be demonstrated by example. This strategic aim of incentivizing researchers is then supported by two subordinate aims. Firstly, presenting a UI in Fidus Writer that is easy to use for researchers to manually tag articles. Secondly, creating a model for how experts in other fields can construct and share domain specific groups of ontologies in the ORKG.

There are two target audiences for the project: firstly, editor software development teams; secondly, domain experts who are responsible for making knowledge graphs. In both of these target audiences cases we are using FOSS methods of open consultation to do two things, to disseminate the ORKG technology, and for agile feedback.

We have access to the web editor scholarly publishing community in place through the extensive international work of TIBs R&D in scholarly communications. In medical research publishing field we have the expertise of Dr. Maria-Esther Vidal's team who are part of the iASiS big data project including Alzheimer's, and BigMedilytics projects — both EU Horizon 2020 funded.

The activities of the project are focused on making prototype FOSS components that the community can use. This will take place on top of three existing FOSS technologies; Fidus Writer;

ORKG development which is supported by the European Research Council; and ADA a research project from the Open Science Lab for publishing microservices.
Activities will include: agile software development of a pilot FOSS Fidus Writer plugin for manual and automatic ORKG enrichment; a pilot microservice API for ORKG enrichment based on the ADA infrastructure, the microservice is a key activity for the purpose of domain experts being able to distribute ontology configurations for their disciplines; documentation for FOSS, and medical ontology communities: workshops with the web editor community, and with medical iASiS and BigMedilytics communities; and publishing research papers and presenting at conferences.
ORKG is of benefit to research in the areas for 'research data management' in the management and analytics of research. The long-term vision of ORKG is to move away from largely document-based principles of scholarly communication with workflows and measurement being based on the 'document unit' and instead move the focus on its constituent units — discourse capture, provenance, or concept drift, etc. — so that research contributions can be better traced and the interconnections of research contributions are made more explicit and transparent in data analysis or search engines for scholarly communications.
Practices of authoring, editing, and publisher manual enrichment will need to change to add knowledge graph data. This will be in the area of workloads of manual tagging, but also more fundamentally in the importance attached to certain types of information to describe research, for example: methods, concepts, or experiment types, etc.
The practices of ontology maintainers, of RDM, and of open knowledge graphs and software practices in indexing and search systems will also be affected in Open Science.

Details of proposal – evaluation plan

The project will be monitored in the following ways: Using Agile methodologies to make early software releases: using task trackers, open communications channels, and having development guidance in place such as a roadmap. Agree clear deliverables in close consultation with any software development suppliers inside and outside of the partnership. Start the calls for workshops as early as possible and ensure a well scheduled set of steps are carried out in time to ensure attendance for two workshops. Have appropriate feedback forms for all activities. Have open support and bug ticket systems.
The success of activities will be monitored as follows: For deliverables of software and documentation ensure clear milestones are made and have contingency procedures in place to compensate for any shortcomings that might arise. Monitor usage analytics on website, Twitter, GitLab, and GitHub. Produce internal monthly reports and public project final report.
Targets: 2x software releases; number of workshop attendees – 8 for each of the 2 workshops; Fidus Writer plugin use (Kubernetes easy cloud install) – 20 approx.; ORKG microservice API use – 20 approx.; enquiries into use of system by software development teams and by ontology maintainers – 40 approx.; numbers of engaged web editor software teams 10. Research publishing and conferences: produce 2 software papers, 2 research papers, and 2 conferences attended.
Risks – current risks to the project have been ranked as high, medium, or low. Steps have been taken to address high risks and others. High: UI issues making manual markup by researchers too difficult; Technical contractors having delays in delivering or being unable to deliver. Medium: Accessing development groups within the 12 month time period of the grant. Low: Having medical and development attendees come to workshops; Technical barriers posed by architecture choices by other web editor development groups.

Decision

Not shortlisted

Comment on decision from Wellcome

This proposal aimed to address an important issue, however concerns were raised about uptake among authors and therefore the potential impact of the proposed activities.

Title **A search-engine inspired platform enabling open sharing and access of both data and results from all genomic studies**
Lead Applicant **The applicant opted not to share this information**
Details of proposal – team members and collaborators The applicant opted not to share this information
Details of proposal – vision, aims and influence on open research The applicant opted not to share this information
Details of proposal – evaluation plan The applicant opted not to share this information
Decision Funded
Comment on decision from Wellcome The invited full application resulting from this shortlisted concept note is available in a separate file, alongside review comments on that version of the proposal.