Data management in health research: You're funded but are you ready to be FAIR?

Author
Dr Brendan Palmer
Clinical Research Facility-UCC

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-UCC
HRB Clinical Research Facility University College Cork

sonraí
Irish Data Stewardship Network

EXECUTIVE SUMMARY

The following report seeks to capture the conversations that took place during the "Data Management in Health Research" symposium held in University College Cork on 25th October 2023. Attendees, drawn from a wide range of Research Performing Organisations (RPOs), were invited to participate and share individual experiences of data management approaches to support health-focused research in Ireland. The overarching theme was the conduct of research in an environment that must marry heightened data protection obligations to broader engagement with open research practices being sought by funders.

The event was split into a morning session of round table discussions centred on the key overarching themes of "Are you a Data Steward?", "What do the FAIR Guiding Principles mean to you?" and "Supports and Resources Available/Needed". This was followed by presentations delivered by the Chief Academic Officer of the South/Southwest Hospital Group and representatives from the Health Service Executive, the Health Research Board, Science Foundation Ireland and Public Health Scotland. Each presentation outlined signature approaches from a perspective of advancing toward a more modern, open and engaged health research environment.

There is growing recognition of the advantages that accompany closer engagement with open research practices. Set against a backdrop where the General Data Protection Regulation (GDPR) has significantly strengthened individuals' rights over their data, the handling and use of health research data needs to be managed.

At the intersection of these developments lies provision of project-level data stewardship supports. Data documentation, storage, access, ownership and secondary use requires careful evaluation, planning and attention throughout the research lifecycle. The **FAIR** guiding principles for data management (2016) provides a roadmap to assist researchers identify ways in which the resultant data can be made **F**indable, **A**ccessible, **I**nteroperable and **R**e-usable[1]. Data re-use is now entering the collective vocabulary but without dedicated supports and direction, the resultant data will remain "re-useless".

The fundamental question put to all attendees was simple: "**You're funded, but are you ready to be FAIR?**"

[1] Wilkinson *et al.*, (2016). The FAIR Guiding Principles for scientific data management and stewardship.

`

TABLE OF CONTENTS

Overview of the Irish health research landscape 1

"You're Funded But Are You Ready To Be FAIR" Conversations 2

Recognising the importance of data stewardship 2

Infrastructure 2

Supports 4

Recognition 5

Access to Funding 5

Fail to Prepare, Prepare to Fail 6

Data Stewardship supports available through the CRF-UCC 6

Open Research Case study: Public Health Scotland 7

Presenters: Carole Morris & Terry McLaughlin 7

Reproducible Analytical Pipelines 8

Sonraí – Irish Data Stewardship Network 10

Data stewardship sits at the intersection of initiatives 11

A cautionary tale To Avoid history repeating itself 12

The "you" of 3 months ago is terrible at answering email 13

Acknowledgements 14

`

OVERVIEW OF THE IRISH HEALTH RESEARCH LANDSCAPE

Underutilisation of national health datasets in Ireland has been documented. An OECD report (2021) ranked Ireland bottom in a list of 23 countries scored against eight key elements of dataset availability, maturity and use and second to bottom for linkage[2]. Immature (and often absent) national health data governance and infrastructure(s) have compounded the issue where minimal dataset interoperability/linkage dominate the health data research ecosystem (Image 1).

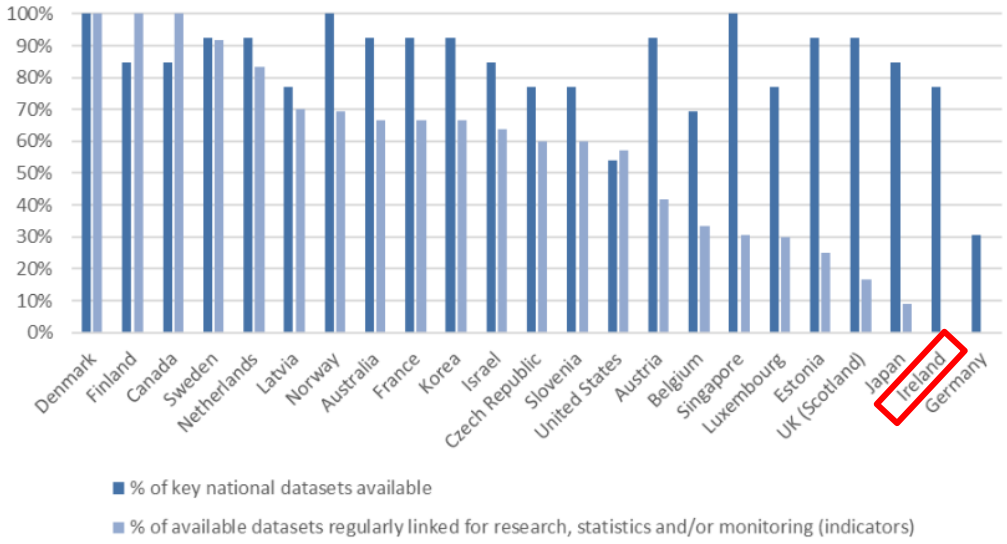

Image 1. OECD Health Working Paper No. 127 (2021) Percentage of key national health datasets available and regularly linked for monitoring and research. The lack of a unique individual health identifier accounts for the absence of dataset linkages in Ireland.

The consequences of a health information system operating with largely unlinked data collections were exposed during the COVID-19 pandemic. Reassuringly, lessons from this period feature prominently in recent policy documents and publications. In May 2022, the Department of Further and Higher Education, Research, Innovation and Science published Impact 2030[3]. In it, recognition of "national-level [data] infrastructure" investment is acknowledged. A legal framework for the development of digital health records is now set out in the Irish Health Information Bill published in 2023[4]. Provisions around enhancing access to patient healthcare data and secondary uses of healthcare data are set out in the document which will start to redress the lack of dataset linkage going forward.

[2] OECD Health Working Paper No. 127 – National Health Data Infrastructure and Governance
[3] Impact 2030: Ireland's Research and Innovation Strategy
[4] Department of Health, Health Information Bill 2023

"YOU'RE FUNDED BUT ARE YOU READY TO BE FAIR" CONVERSATIONS

"Open research, also referred to as open science or open scholarship, is an approach to the scientific process based on open cooperative work, tools, and diffusing knowledge."[5]

RECOGNISING THE IMPORTANCE OF DATA STEWARDSHIP

During the morning session, attendee conversations consistently returned to the same underlying issue: the lack of emphasis on data stewardship at local level within organisations. This was most keenly voiced from attendees based in academia where funder expectations on data management have risen in recent years.

At the beginning of research projects, there's often a lack of clarity regarding the necessary support needed to implement FAIR data guiding principles. This results in ineffective data management throughout the research process. Moreover, there's a noticeable absence of specific data management roles or responsibilities outlined in job descriptions. Instead, data management tasks tend to be dispersed among team members in an *ad hoc* manner.

While national policies increasingly move towards open research initiatives, translating these policies into actions within individual research teams remains the challenge.

INFRASTRUCTURE

Research data integrity and enhanced data quality begin with access to dedicated digital research infrastructure and the presence of specialised skillsets. The intersection between data management and IT expertise is critical in this regard.

The National Open Research Forum (NORF) Landscape Report has identified that, in the Irish research ecosystem, there is little by way of infrastructure for supporting the goals of open research[6]. Those exploratory infrastructure projects that have emerged through initial NORF funding program calls face significant challenges to address long-term sustainability following the conclusion of seed funding. It is not just the technical infrastructure but also the person infrastructure accompanying new projects that must be looked at in tandem.

[5] NORF National Action Plan for Open Research 2022-2030
[6] NORF National Open Research Landscape Report 2021

Participant contribution: "One of the biggest obstacles we encounter is infrastructure and that means that we're very much struggling to engage with FAIR."

The preceding decade has not prepared the Irish academic health research community for the paradigm shift that applied to health research data following the introduction of GDPR and the Health Research Regulations. The emergence of short-term research roles prioritised project level supports at the expense of central resource provision and infrastructure expansion. Presently, many research projects are islands unto themselves; the work undertaken may be discipline specific, the study population may have a rare condition, the study sites may be widely dispersed, the expertise available is largely dependent on the skillset of incoming new hires. For smaller projects, the hiring, budget, ethics and administrative activities may all rest with a single individual.

Image 2. Word cloud capturing attendee responses to the prompt "What supports and resources does effective data stewardship require".

Among attendees, routes towards providing effective data stewardship supports to projects emerged under the three main headings of (1) infrastructure, (2) centralised support and (3) leadership. Sub-headings of institutional policy, organisational structure, the position of the library and senior staff advocacy were also evident.

In the absence of intra-institution coordination, implementation of Open Research is determined by the individual research team and the success of any research group/team approaches become reliant on the existing skill set among members of that group/team.

SUPPORTS

The FAIR guiding principles for scientific data management and stewardship have been widely incorporated into funding calls to advance approaches to data management engagement. However, the distinction between quantitative data and qualitative data has not been formerly communicated to the research community. Many existing supports map more closely to "scientific" quantitative data. Operationalisation of qualitative open research principles requires attention. The skillset/approach to enact the FAIR guiding principles depend on the primary data types and further direction is required in this regard.

FAIR-centred data stewardship activities depend on the purpose and the person/team. The existing supports provided by data stewards in the audience were described as being dependent on the local context and the needs of the unit they operate in. There was broad agreement in the room that approaching data stewardship should not stop with the primary dataset. Additional outputs from the project such as protocols, methodologies etc. add value to the wider research ecosystem if developed for downstream dissemination to external audiences.

There remains a high degree of uncertainty around making data available post-project. Professor Kalpana Shankar (UCD) spoke on the challenges researchers face when navigating open science and intellectual property rights in the digital health ecosystem. The question(s) surrounding how the research ecosystem balances open research mandates with commericalisation remains unclear and there is little by way of guidance or policy in place currently.

Participant contribution: "Commercialisation of research is very attractive to the university, but the open research route is more of a community effort."

There is still a very informal nature around accessing supports. Most university libraries have established a research data service but there is an opportunity for Clinical Research Facilities/Centres to partner with library staff and deliver tailored supports matched to the health research environment within which they operate. This would allow for [patient-focused] local initiatives to take shape through the medium of CRF/Cs and help release the siloed knowledge that currently exists.

Participant contribution: "People in the same organisation do not know that they have access to the same supports."

RECOGNITION

"Data steward" can describe a variety of roles and responsibilities yet these functions are largely undefined within active research teams. Among symposium attendees, 37% self-identified as the "data-person" of their research group. A further 26% recognised another member of the research team has having that responsibly.

Just **3%** of attendees present did not see a data stewardship element in their role.

Participant contribution: "We see it [data management] in our daily workloads but it isn't being recognised by management."

ACCESS TO FUNDING

Funders recognise the inherent value of the data collected from the research they support. Researchers and their host institutions need to respond to maintain eligibility to apply for prestigious awards as the opportunity to submit an application can reliant on access to local infrastructure.

Participant contribution: "The university is looking to increase the number of awards from European sources but the supports to even begin approaching data management requirements just aren't there."

Effective from January 2023, the US National Institute of Health Data Management and Sharing (DMS) policy expects investigators and institutions to (i) plan and budget for the managing and sharing of data, (ii) submit a DMS plan for review when applying for funding and (iii) comply with the approved DMS plan[7].

Similarly in Europe, Research Data Management is mandatory for any Horizon Europe project generating or reusing research data. It is a key part of Horizon Europe's open science requirements. Key minimum requirements include (i) preparation of a Data Management Plan and keep it updated throughout the course of the project, (ii) deposition of data in a trusted repository and provide open access to it and (iii) provision of information (via the same repository) about any research output or any other tools and instruments needed to re-use or validate the data[8].

[7] NIH Data Management and Sharing Policy
[8] European Commission – Open Research Europe

FAIL TO PREPARE, PREPARE TO FAIL

Another recurring theme from the morning conversations was the need to plan early, but equally the need to include staff drawn from various backgrounds. Patient facing staff may identify issues with the data collection process. Study statisticians might identify opportunities to improve within-project data quality that will support downstream analyses. Data management takes the form of many tributaries feeding into the broader body of a research project.

Participant contribution: "Ultimately, unshared data contributes to research waste."

Finances and time are increasingly scare as a body of work enters the final stages of the project. The ability to deliver on time and meet end-of-project targets is facilitated by an integrated approach to data stewardship that has engaged with stakeholders throughout the research lifecycle. Alongside the data files, parallel processes to build metadata and refine study documentation all have roles that lead aim to maximise shareable outputs.

Ronald Fisher, Statistician (1890-1962): "To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination. He can perhaps say what the experiment died of."

DATA STEWARDSHIP SUPPORTS AVAILABLE THROUGH THE CRF-UCC

Within the CRF-UCC, the Statistics, Data & Analysis Unit (SDAU) has sought to apply this principle. Overtime, the maturation of the service has been driven by inputs from CRF staff (Director, Principal Statistician, Data Manager etc.) and through the development of partnerships with colleagues locally (Library, Research Office, Data Protection Office).

The SDAU seeks engagement with research teams at the earliest opportunity ensuring that (i) guidance is provided to inform grant applications, (ii) tailored budget lines are included to protect within-project activities, (iii) delivery of study objectives are aligned to the Statistical Analysis Plan, (iv) structured electronic data capture is deployed throughout the project and (v) the resultant coordinated approach links to open research ambitions (Image 3).

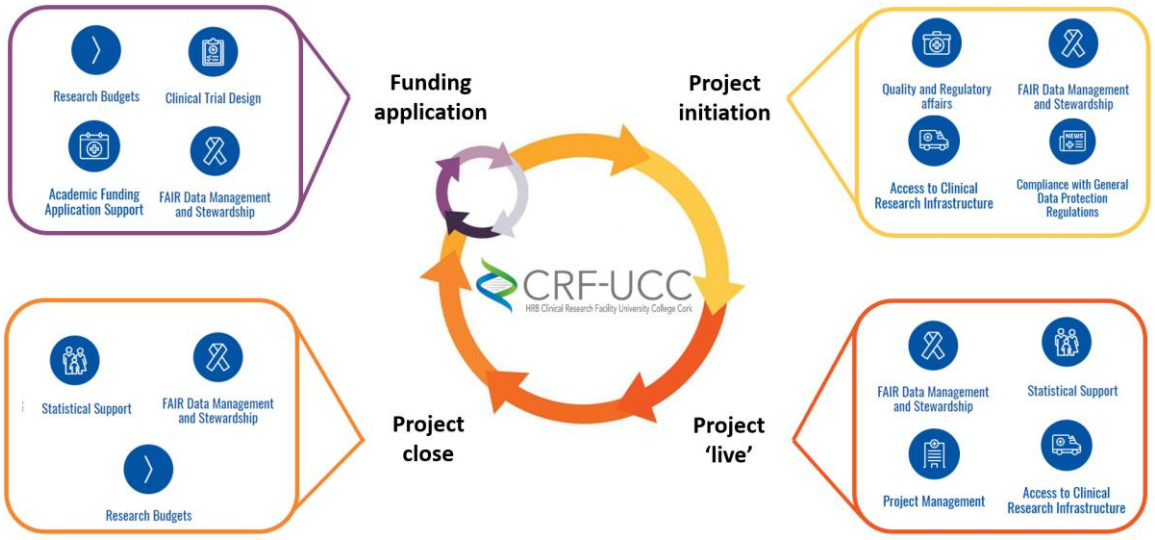

Image 3. Research lifecycle engagement with the Clinical Research Facility-UCC

The SDAU structure/service offering has been led by local UCC experiences. Similar activities/services exist in various forms and stages of development within other CRF/Cs. Ultimately, the CRF/Cs nationally all have a core directive to support academic-led patient focused research. There is an opportunity therefore to better coordinate the successes of this established infrastructure network to foster growth and stimulate knowledge transfer.

OPEN RESEARCH CASE STUDY: PUBLIC HEALTH SCOTLAND

PRESENTERS: CAROLE MORRIS & TERRY MCLAUGHLIN

Public Health Scotland (PHS) was formed in April 2020 and is sponsored by the Scottish and local government. It has ~1,300 staff, 500 of which work in the data and statistics field.

PHS controls a wide variety of data sets that cover the Scottish population from (pre-)'cradle to grave' (Image 4). It employs standardised data collection protocols and the majority of collections are digitised. Most data sets offer 100% coverage and can be linked together using the "CHI number" which is unique to each individual. Everyone who touches the health service in Scotland, whether it's registering with a GP upon entry into the country or when you're born in Scotland, is given this CHI number. Some data sets are dynamic, such as the prescribing payments system. Many data sets go back decades, giving rise to a powerful source of longitudinal, rich, linked data.

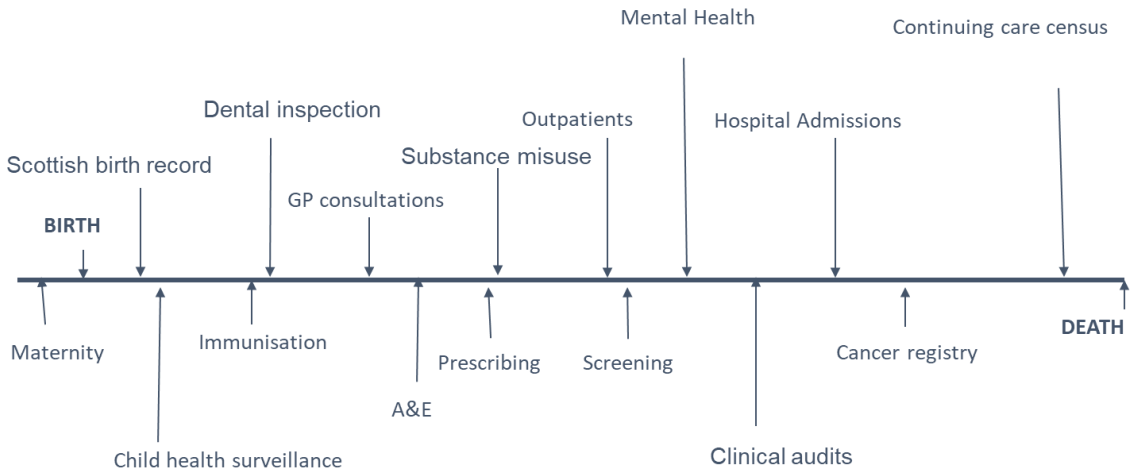

Image 4. A simplified overview of the PHS data catalogue.

Collaboration with academia is key where local safe havens exist as collaborations between the NHS universities that offer secure and confidential storage or processing of digital data. In 2020, Research Data Scotland was established to serve as a 'front door' to develop enhanced research data access by working with data controllers to streamline data access.

In PHS, the eDATA Research and Innovation Service (eDRIS) was established to create a single point of entry for the research team to access the data collections held. Services such as identifying data availability, details on data content, relationship development with external data controllers and data access permissions are offered. If a research team has data that can be re-used, eDRIS can work with the project team to build the metadata and include it in the data catalogue. This process includes a data linkage procedure ensuring that data exiting the service is not identifiable.

REPRODUCIBLE ANALYTICAL PIPELINES

Reporting workflows within PHS traditionally relied on primarily manual tasks for the production of official statistics and publications.

Data from a data store would be exported into the statistical software suite, SPSS. These files went on to be formatted in Microsoft Excel for publication. The publications would be built using Microsoft Word where numbers and charts were copied and pasted manually from Excel. Errors were found by reading through the document and mistakes, when found, were corrected by repeating the entire process. It was complex, error prone and labour intensive. The Hospital Standardised Mortality Ratio (HSMR) was one such report. HSMR is a measure of mortality adjusted to take account of some of the factors known to affect the underlying risk of death. The report is based on all acute inpatient and day case patients admitted to all specialities in hospital (excluding obstetrics and psychiatry). The primary outputs were:

- 40 page report released quarterly
- 2 page report summary
- 7 Excel workbooks containing a mixture of tables and charts
- **7 days** to produce a single report from start to finish

To remedy this error prone and labour intensive process reproducible analytical pipelines (RAPs), based on the R programming language and the RStudio Integrated Development Environment, were developed (Image 5). The methodology combines the principles of reproducible research with data science, tools and best practice. A number of immediate benefits were apparent; (i) no (or few manual steps), (ii) it became an auditable and reproducible process enabled by version control and (iii) resulted in high quality and validated output supported by peer review.

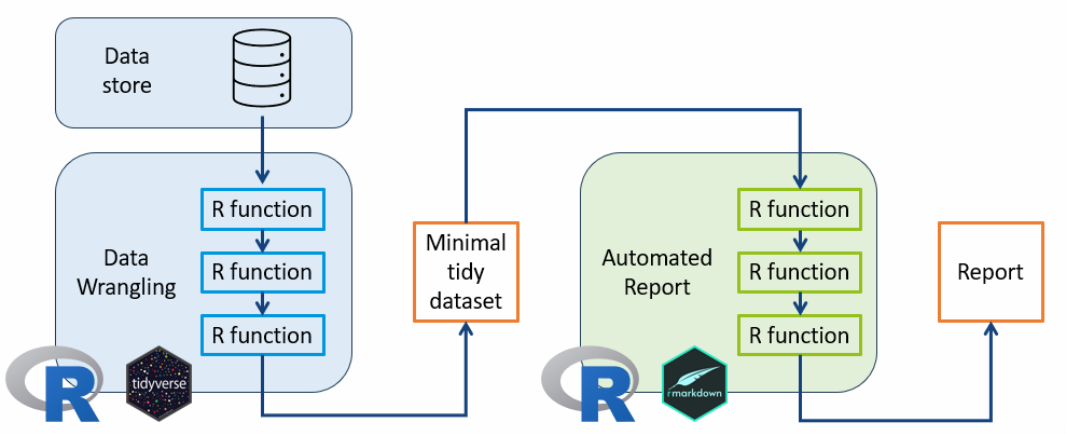

Image 5. Overview of Reproducible Analytical Pipeline workflows.

One additional output of the transition to RAP was the production of an R package containing 17 functions to carry out all the tasks that are repeated each quarterly publication cycle. All of the code is published and if you had access to the data, you would be in a position to reproduce the report. The production of the HSMR publication is now automated from start to finish and continues to be published quarterly:

- a 45-page publication report[9]
- a 2-page report summary[10]
- a web page summary and R Shiny dashboard[11]
- 3 Excel workbooks
- a data release on an open data platform[12]
- **50 minutes** to produce a single report from start to finish

[9] HSMR report - 8th August 2023
[10] HSMR report summary - 8th August 2023
[11] HSMR web page summary -April 2022 to March 2023
[12] HSMR open data platform

To future-proof initiatives in this space, there is ongoing investment in IT infrastructure. The RAP environment is now hosted on Microsoft Azure cloud computing and takes advantage of Kubernetes technology. This allows for an environment that can automatically scale up or scale down based on resource demands over the course of a day.

PHS have worked closely with NHS IT partners and external IT consultants to expand the offering and are working to integrate the RAP workflow more widely into reporting processes. Alongside technical infrastructure, there have been initiatives to support staff training and software development. Going forward, PHS are developing further internet-facing tools to share outputs with external customers and stakeholders, such as Scottish Government, Health & Social Care Partnerships and Third sector.

SONRAÍ – IRISH DATA STEWARDSHIP NETWORK

The establishment of NORF in 2017 was the initial spark to ignite Irish open research engagement and adoption.

NORF is funded by the Department of Further and Higher Education, Research and Innovation and Science through the Higher Education Authority. NORF seeks to (i) develop and propose national actions to address the challenges of changing the Irish research system and (ii) to oversee and guide implementation of the National Action Plan for Open Research. One mechanism it has at its disposal to do it is through the award of research funds that meet these aims.

Among the first round of NORF awardees was a project to establish an Irish Data Stewardship Network. Led by Dr Aoife Coffey, UCC Library, the group termed Sonraí (Irish meaning data), came into being in 2023[13]. Sonraí aims to enable the development of data stewardship skills across the national research landscape by (i) raising the profile of data stewards, (ii) obtaining greater recognition of the need for data stewardship and (iii) professionalisation of the role of data steward.

Participant contribution: "A fundamental problem in health research is that all PIs are already sitting on the data collected and don't see themselves in a stewardship role over that same data."

Sonraí is a grassroots response that recognises the changes that have taken hold since 2018 and seeks to provide a framework to structure data stewardship supports around. The

[13] Sonraí - Irish Data Stewardship Network

coalescing of funder policies on research data management has provided the impetus for data management descriptions to appear in job advertisements. The development, training and advocacy for staff occupying these positions is in its infancy. Sonraí has to potential to distil the fragmented pieces of data stewardship support puzzle and help fill the resource/support void that exists (c.f. Image 2).

Image 6. Sonraí introduction delivered by Dr Aoife Coffey, UCC library and project lead.

An opportunity to create synergies with existing supports such as those provided centrally through libraries or CRF/Cs, as previously described, now exists. In the immediate term, Sonraí can act as the linchpin of an emerging relationship.

However, initiatives such as Sonraí need to be sustained post-award. A culture of data stewardship needs to be fostered and allowed to grow. It is here that RPOs and research funders need to evaluate how open research can be more holistically supported. Discrete funding awards are sufficient to open the door, but in order to keep it open, a mechanism to finance continued development of an offering needs to be defined.

DATA STEWARDSHIP SITS AT THE INTERSECTION OF INITIATIVES

Health research is conducted in a rapidly changing and increasingly complex environment. The demands and pressures on academic-led clinical research today exceed that of 20 years ago. Legislative, administrative, fiscal and societal impacts all feed into the process.

Central government, research funders and RPOs need to recognise deficits in infrastructure, support and training to ensure that active researchers can divest themselves of roles not

directly aligned to their expertise and place trust and guidance in the hands of the institution where they are employed. An adapted "Theory of change" model offers a route toward deploying a hierarchy of initiatives that support adoption of embedding open research practices in the Irish health research setting (Image 7).

Image 7. Approaching open research in the Irish health research setting based on the "Theory of Change" model. Credit: Brian Nosek, Centre for Open Science.

Repeatedly, over the course of the symposium, a positive view of two-way engagement between the funder and the researcher was expressed. Researchers, identifying weaknesses in a process should have (through a forum such as Sonraí) an opportunity to feed this information back to the funder. Equally, funder reviews of policy should be circulated to the research community for wider reach.

A CAUTIONARY TALE TO AVOID HISTORY REPEATING ITSELF

During the COVID-19 pandemic, governments around the world faced difficult choices. All decisions would ultimately be based on the underlying data collected from the case numbers, hospitalisations and deaths associated with the SARS-CoV-2 infections. In the United States of America, such information was collected, organised and reported at State level whereas national decision-making occurs at Federal level.

During the initial stages of the pandemic, news organisations sought to obtain the underlying statistics to assist with their reporting of the situation. One such organisation,

The Atlantic, assembled a team of volunteers led by an editor and data scientist to collect, clean and aggregate daily reported COVID-19 cases. Over time, this largely volunteer effort not only informed *The Atlantic* writings, but figures and graphs emanating from their work began to appear on official government updates. This led to a startling realisation among *The Atlantic* staff…

"Before March 2020, the country had no shortage of pandemic-preparation plans. Many stressed the importance of data-driven decision making. Yet these plans largely assumed that detailed and reliable data would simply… exist."

In March 2021, the Covid Tracking Project concluded and *The Atlantic* published an article outlining their experiences[14,15].

THE "YOU" OF 3 MONTHS AGO IS TERRIBLE AT ANSWERING EMAIL

Patient outcomes are better in research active health settings. This is seen in Cork University Maternity Hospital where infant survival by week of gestation has improved from 14% to 63% at 23 weeks over the space of a decade, a situation that clinical staff attribute directly to the research culture across departments (Image 8).

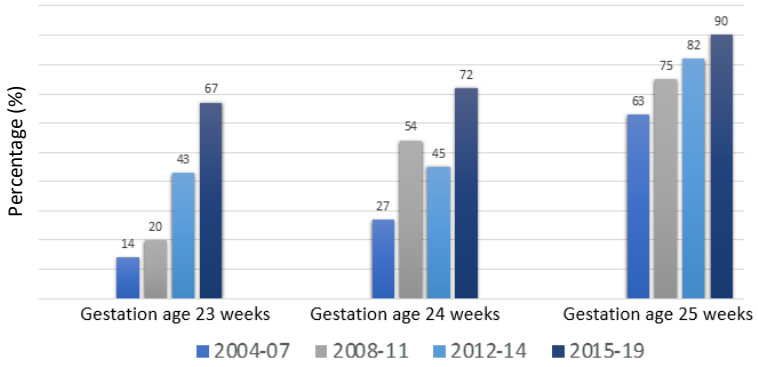

Image 8. Survival by week of gestation at Cork University Maternity Hospital 2004-2020. Credit: Prof Gene Dempsy, Consultant Neonatologist in the Cork University Maternity Hospital and Clinical Professor of Paediatrics, University College Cork

[14] The Covid Tracking Project
[15] The Atlantic – Why the Pandemic Experts Failed

Data quality and research integrity are both beneficiaries of adopting open research practices. Improved data quality directly enhances the accuracy of research findings, providing decision-makers with more reliable insights to inform their actions. Dissemination of findings in an 'open' and transparent way allows for independent verification and validation of said findings.

If ambitions for data accessibility, transparency and reproducibility are to be realised in Ireland, we need to begin building in a steady, sustained commitment to support data stewardship.

For research funding to maximise cost effectiveness, access to training and skills development at local level need to match disciplinary-specific requirements. Centrally provisioned core infrastructure(s) to support research data and relevant expert supports need to be made available. Leadership and alignment of initiatives at national levels are now needed to cement local activities.

ACKNOWLEDGEMENTS

The organisers wish to thank all attendees to this symposium. Special thanks to the invited speakers for contributing their expertise and time to fuel the conversations that followed.

Professor Ivan Perry, Interim Director, CRF-UCC

Dr Aoife Coffey, Research Data Coordinator, UCC Library and Sonraí project lead

Professor Kalpana Shankar, Information and Communication Studies, UCD

Professor Helen Whelton, Head of the College of Medicine and Health, UCC and Chief Academic Officer to the HSE SSWHG

Dr Maria Quinlan, General Manager, National Research & Development, HSE

Dr Sharon Kappala, Health Research Board

Oonagh Ward, Head of Research and Innovation Infrastructures, Health Research Board

Jenny Clarkin, Grants Compliance Manager, Science Foundation Ireland

Edie Davis, Science Foundation Ireland

Carole Morris, Head of Data and Modelling Services, Public Health Scotland

Terry McLaughlin, Principal Information Analyst, Public Health Scotland