researchfish
by interfolio

Studying Science and Innovation Data

Gavin Reddick
25 October 2021

Contents

- Who am I?
- What is Researchfish
- Why collect information on research?
- How can the information be used?
- What information can be leveraged?
- Description of sources of funding information

Who am I? What is my Background?

Gavin Reddick

Quantitative Behaviouralist

Senior Information Analyst at UK Medical Research Council – Evaluation

Chief Analyst at Researchfish

What is Researchfish?

Researchfish is an online platform designed to make it easier for researchers to report to report the outputs, outcomes and impacts of their research.

Why ask for information?

- **Advocacy** for research funding
- **Accountability** to the funders of research
- **Analysis** to understand what works in research and leads to impact
- **Allocation** of future research funding

'If I was a philanthropist giving millions of pounds, I would want to see what is happening with it – and even I get frustrated [about reporting], as a researcher, when you see millions of pounds going to researchers and you wonder what is coming out of it, and you see nothing coming out of it ... And this is always the case. In a lot of things money is given, and no one follows up … Gone are the days when you can do what you like. Now you are publicly funded … we want to account for what you are doing ... is it leading to some benefit to the public?...' *PI Quote - Researchfish: A Forward Look, p 20*

Traditional Evaluation/Reporting?

- Traditionally Impact Evaluation normally gathered information on the outcomes generated via Final Reports
- Large documents containing write up of project
- Reports completed immediately post funding but some outputs/impacts take time to become realised
- Detailed reports in free text produced by PI but systematic analysis problematic – can use text mining techniques but even then “key” information is often missing

Analytical Rigour

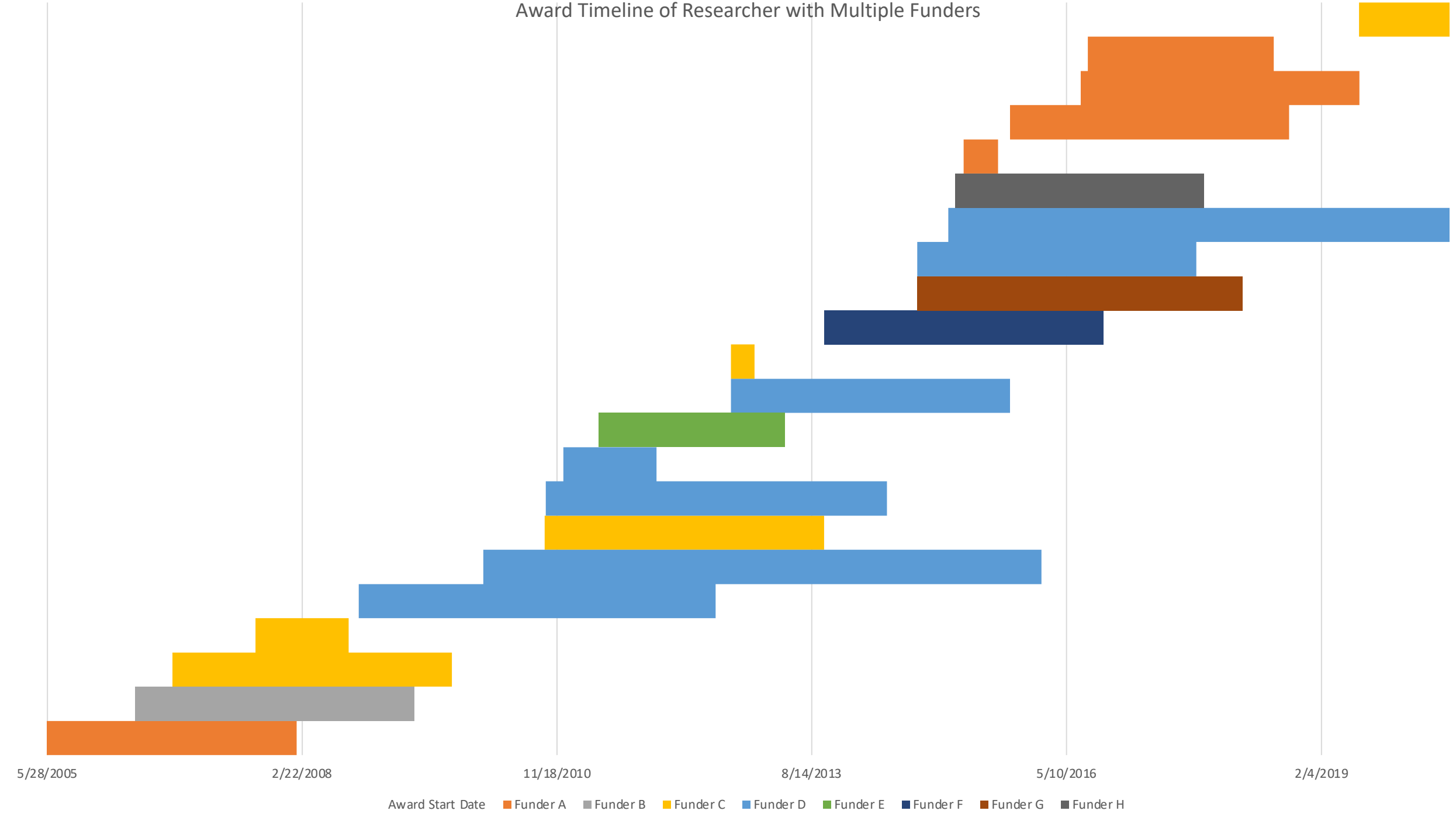
- Funders of research are often very rigorous about the research that they fund but it can be a challenge to apply the same level of rigour to the funding process itself.
- Importance of the “4 As” has driven the demand to evidence claims about the efficacy of funding.
- Research/generation of knowledge is a complex social phenomenon.
- Data ≠ Analysis ≠ Evaluation

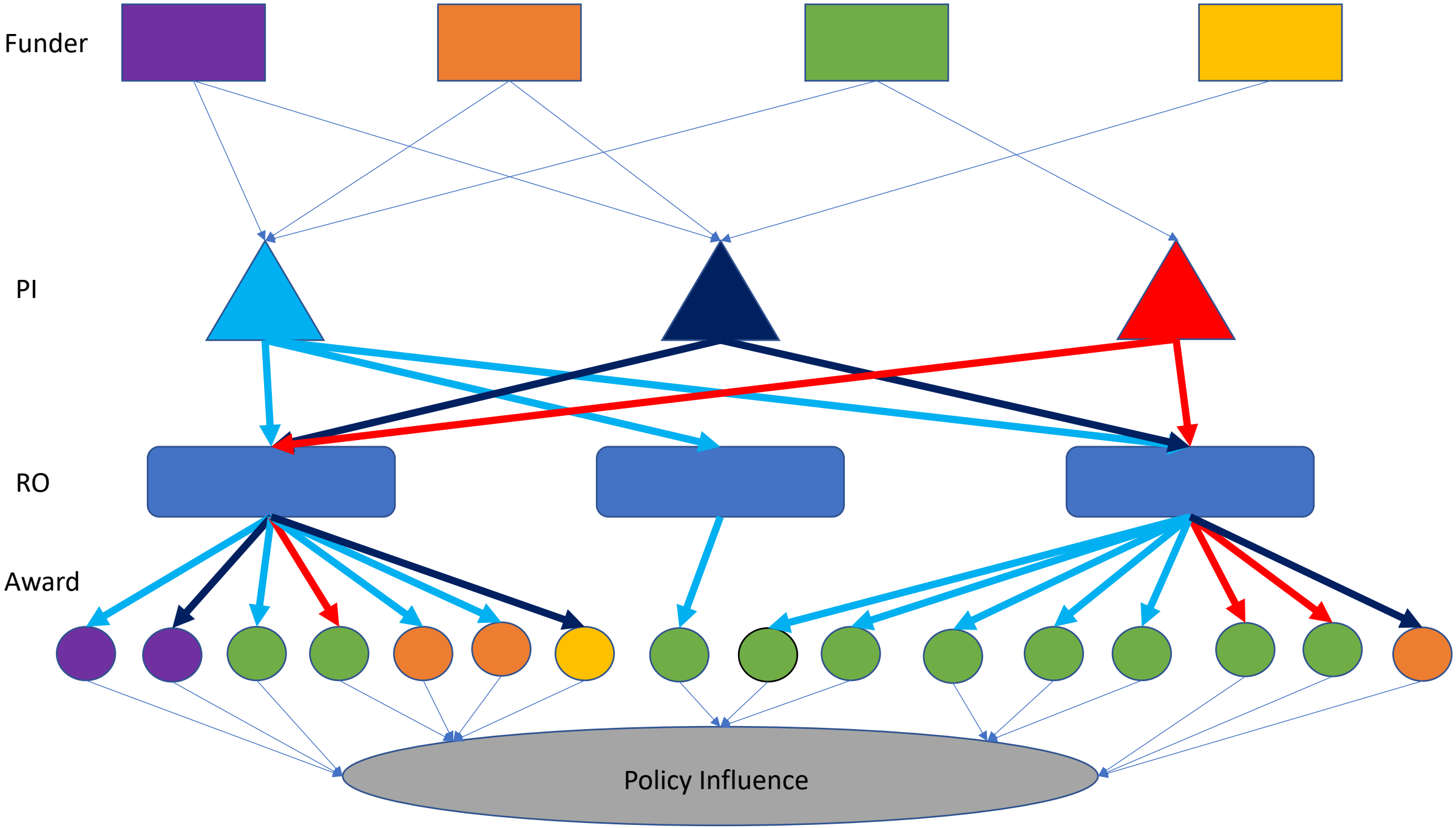
Why is Interconnectedness Important?

1. Reduces the need to ask people for more information – data can be understood from the linkages

2. Increases the consistency and quality of the information supplied if it has been externally validated instead of simply reported

3. Allows for more sophisticated understanding of the data and what it might mean


Award Timeline of Researcher with Multiple Funders
5/28/2005
2/22/2008
11/18/2010
8/14/2013
5/10/2016
2/4/2019
Award Start Date
Funder A
Funder B
Funder C
Funder D
Funder E
Funder F
Funder G
Funder H


Funder
PI
RO
Award
Policy Influence

Publications	Intellectual Property
Collaborations	Medical Products
Further Funding	Artistic and Creative Products
Next Destination	Software and Technical Products
Engagement Activities	Spin Outs
Influence on Policy	Awards and Recognition
Research Tools and Methods	Use of Facilities
Research Databases and Models	Other Outputs

Publications	Intellectual Property
Collaborations	Medical Products
Further Funding	Artistic and Creative Products
Next Destination	Software and Technical Products
Engagement Activities	Spin Outs
Influence on Policy	Awards and Recognition
Research Tools and Methods	Use of Facilities
Research Databases and Models	Other Outputs

Research Funding Data

1. Today there is no single source containing “good” information on research funding.
2. There are a number of global initiatives seeking to improve the available information on research funding, most notably Crossref, but the nature of funding data poses greater challenge than e.g. publications.
3. Publications are public but information on even publicly funded research is not without controversy (funding in different areas, etc.)
4. Funding organisations are generally not in the habit of thinking about funding data as “of external interest” as opposed to internal files – data management.

Gateway to Research

1. Information on UK Research and Innovation funded awards
2. 100,000+ awards and 1.6 million outputs
3. Mainly 2006-present day
4. https://gtr.ukri.org/

5. Information available via API – GTR2
6. Excellent data but limited in coverage to relatively few (though important) funders.
7. Data model used makes it difficult to understand lateral connections via outputs (too many “unique” identifiers for one thing)

Federal Reporter

1. Information on US federally funded awards
2. 1,00,000+ “FY awards” and 1.6 million outputs
3. Mainly 2006-present day
4. https://federalreporter.nih.gov/

5. Information fractured, one award/project/program can have many records across financial years and funding streams, making it difficult to draw conclusions.
6. Large volume of data but difficult to use and difficult to connect outside of specific purposes for which the data is published.

CORDIS

1. Information on European Commission funded awards
2. 1984-present day
3. https://cordis.europa.eu/

4. Information available via API but only partially – most award information available via spreadsheets.
5. Non-unique acronyms are the main way used to refer to these projects which can make it difficult to use to link data externally.

Combining data from multiple sources

1. Importance of understanding the schema of the data
2. Importance of understanding the flows of the data
3. Importance of “soaking and poking” the data

EuropePMC

1. Information on health relevant funded awards
2. 29 funders (mainly UK based)
3. https://europepmc.org/grantfinder

4. Information available via API.
5. Tries to join together funder validated award details and publications.
6. Contains space for PI ORCID but low population.

OpenAire

1. Information on funded awards, mainly European, though also US, Australia
2. 29 funders (mainly UK based)
3. https://www.openaire.eu/

4. Information available via API.
5. Information does not include funded amounts.

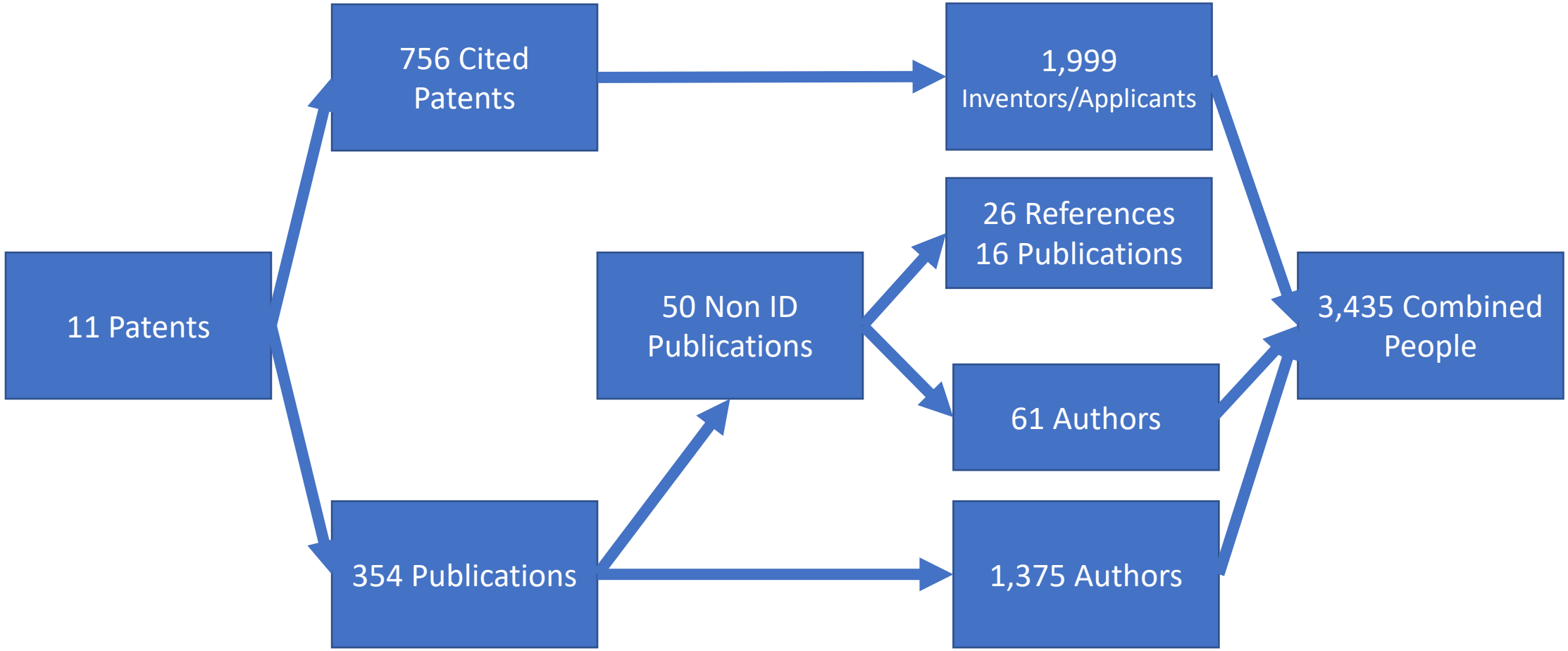
ORCID

1. International non-profit initiative
2. Goal to create unique, persistent identifier of people
3. https://orcid.org/

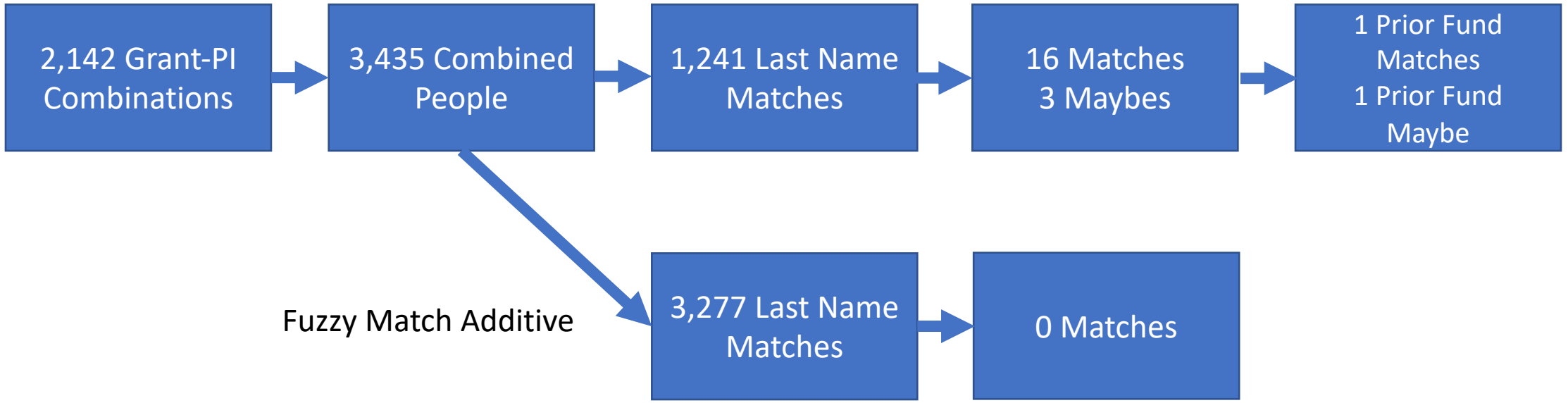
4. Information available via API.
5. Some challenges in “multi-unique”
6. Some challenges in data quality of “works” linked to people.

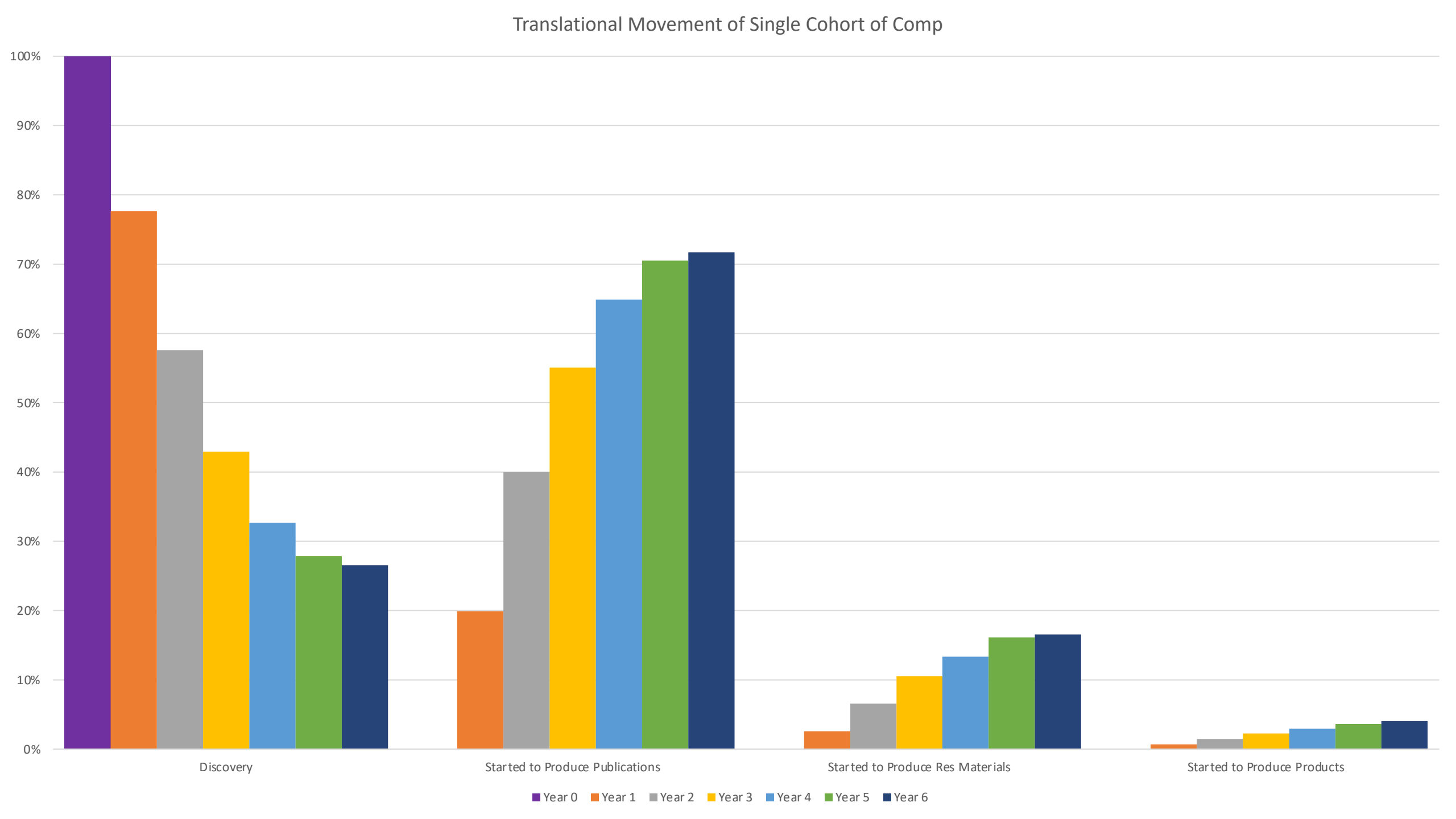
Projects using data

Therapies and Identified Patents



Grantee Matching




Translational Movement of Single Cohort of Comp
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Discovery
Started to Produce Publications
Started to Produce Res Materials
Started to Produce Products
Year 0
Year 1
Year 2
Year 3
Year 4
Year 5
Year 6

Basic model: duration and funding value

```
Call:
glm(formula = is_ref ~ duration_bi + value_binned, family = binomial(link = "probit"),
    data = ref_model_df, maxit = 100)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.3070  -0.2506  -0.2434   2.6696

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9056816  0.0123525 -154.28   <2e-16 ***
duration_bi1  0.2210407  0.0191395   11.55   <2e-16 ***
value_binned  0.0126574  0.0007999   15.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21165  on 63522  degrees of freedom
Residual deviance: 20493  on 63520  degrees of freedom
AIC: 20499

Number of Fisher Scoring iterations: 79

   McFadden McFaddenAdj    CoxSnell        Tjur
 0.03175503  0.03147154  0.01052468  0.01735982

         Predicted 0 Predicted 1 Total
Actual 0       60981          28 61009
Actual 1        2505           9  2514
Total          63486          37 63523
```

Independent variables:

1) Duration (4 Binary)
2) Funding value (Bins of 100k)

Advanced model: 9 predictors

```
Call:
glm(formula = is_ref ~ duration_cat + value_cat + pub_cat + poli_cat +
    ip_bi + spin_bi + col_bi + finished + ff_value_cat, family = binomial(link = "probit"),
    data = ref_model_df, maxit = 100)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9318  -0.2628  -0.1898  -0.0143   4.3741

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.08847    0.15745 -25.967  < 2e-16 ***
duration_cat1  0.16533    0.07807   2.118 0.034191 *
duration_cat2  0.17833    0.07663   2.327 0.019953 *
duration_cat3  0.28042    0.07913   3.544 0.000394 ***
value_cat1     0.28514    0.11049   2.581 0.009861 **
value_cat2     0.22172    0.05354   4.141 3.46e-05 ***
value_cat3     0.21455    0.04413   4.862 1.16e-06 ***
value_cat4     0.43935    0.06015   7.304 2.79e-13 ***
value_cat5     0.49718    0.18230   2.727 0.006386 **
pub_cat1       1.61308    0.13490  11.957  < 2e-16 ***
pub_cat2       2.03681    0.13721  14.845  < 2e-16 ***
pub_cat3       2.42529    0.13820  17.549  < 2e-16 ***
pub_cat4       3.20591    0.13997  22.905  < 2e-16 ***
poli_cat1      0.24889    0.04709   5.286 1.25e-07 ***
poli_cat2      0.40038    0.04729   8.467  < 2e-16 ***
ip_bi1         0.04859    0.05489   0.885 0.376019
spin_bi1       0.30912    0.08338   3.707 0.000209 ***
col_bi1        0.09476    0.02454   3.861 0.000113 ***
finished1      0.24948    0.02874   8.679  < 2e-16 ***
ff_value_cat1 -0.30021    0.11128  -2.698 0.006982 **
ff_value_cat2 -0.02445    0.05436  -0.450 0.652932
ff_value_cat3  0.02009    0.03283   0.612 0.540638
ff_value_cat4  0.07670    0.04986   1.538 0.123986
ff_value_cat5  0.13461    0.08535   1.577 0.114770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21165  on 63522  degrees of freedom
Residual deviance: 15239  on 63499  degrees of freedom
AIC: 15287

Number of Fisher Scoring iterations: 10
```

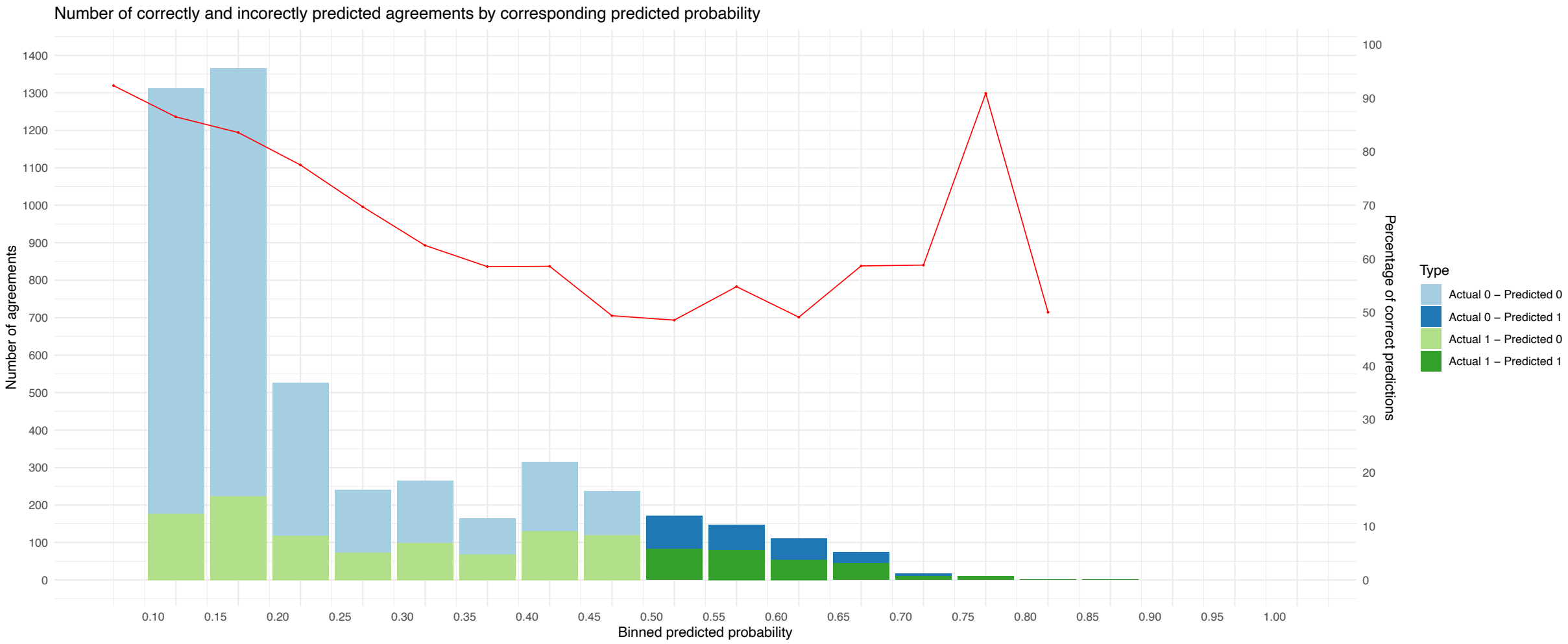
Independent variables:

1) Duration (4 categories)
2) Funding value(6 categories)
3) Number of publications (5 categories)
4) Number of policy influences (3 categories)
5) Number of IP registrations (Binary)
6) Number of spinouts (Binary)
7) Number of collaborations (Binary)
8) Finished or not (Binary)
9) Further funding value (6 categories)

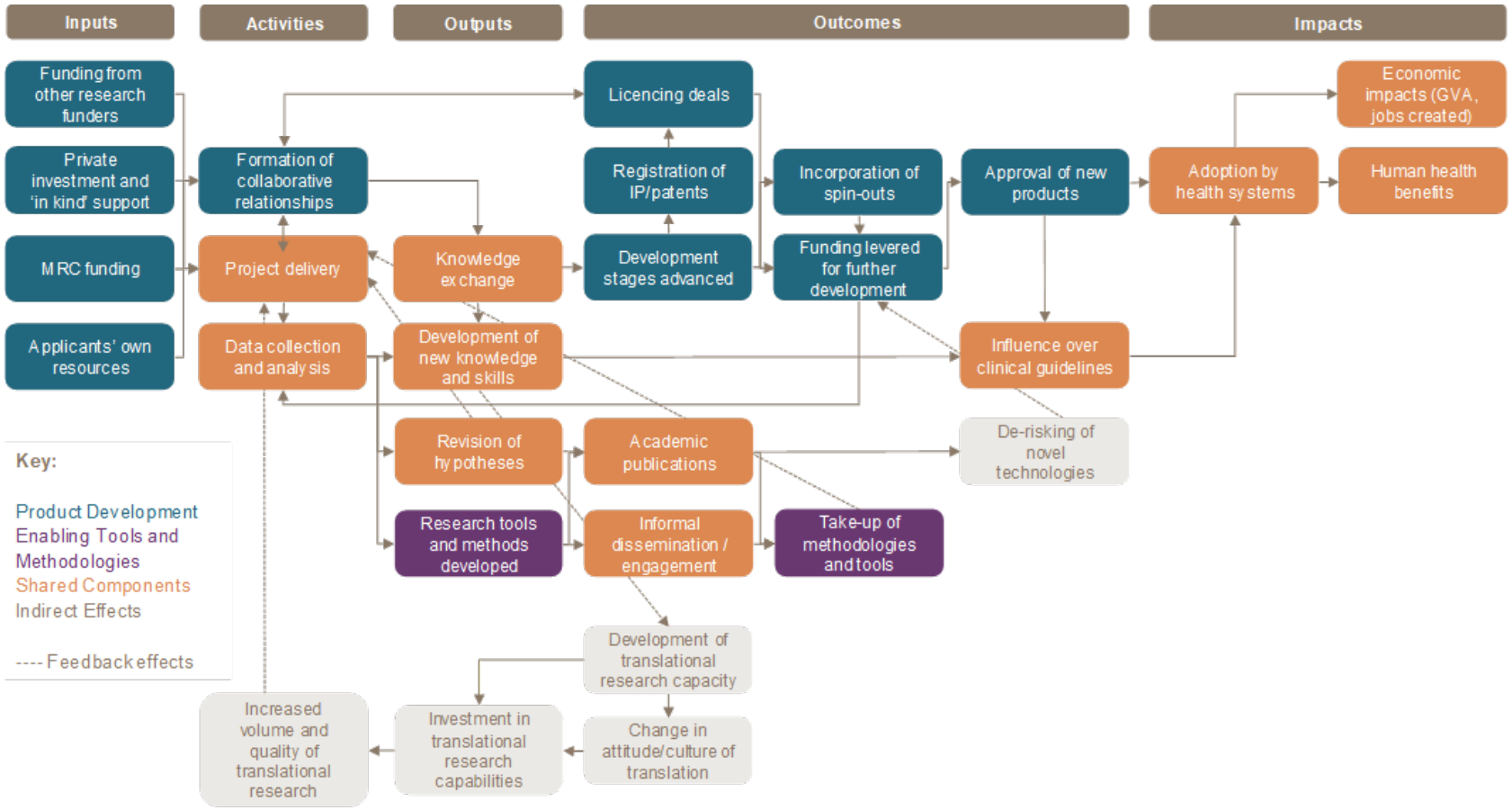
```
     McFadden McFaddenAdj    CoxSnell        Tjur
    0.2799841   0.2777162   0.0890688   0.1684516

C
          Predicted 0 Predicted 1 Total
Actual 0        60758         251 61009
Actual 1         2228         286  2514
Total           62986         537 63523
```

Predicted probabilities



Logical framework for evaluation



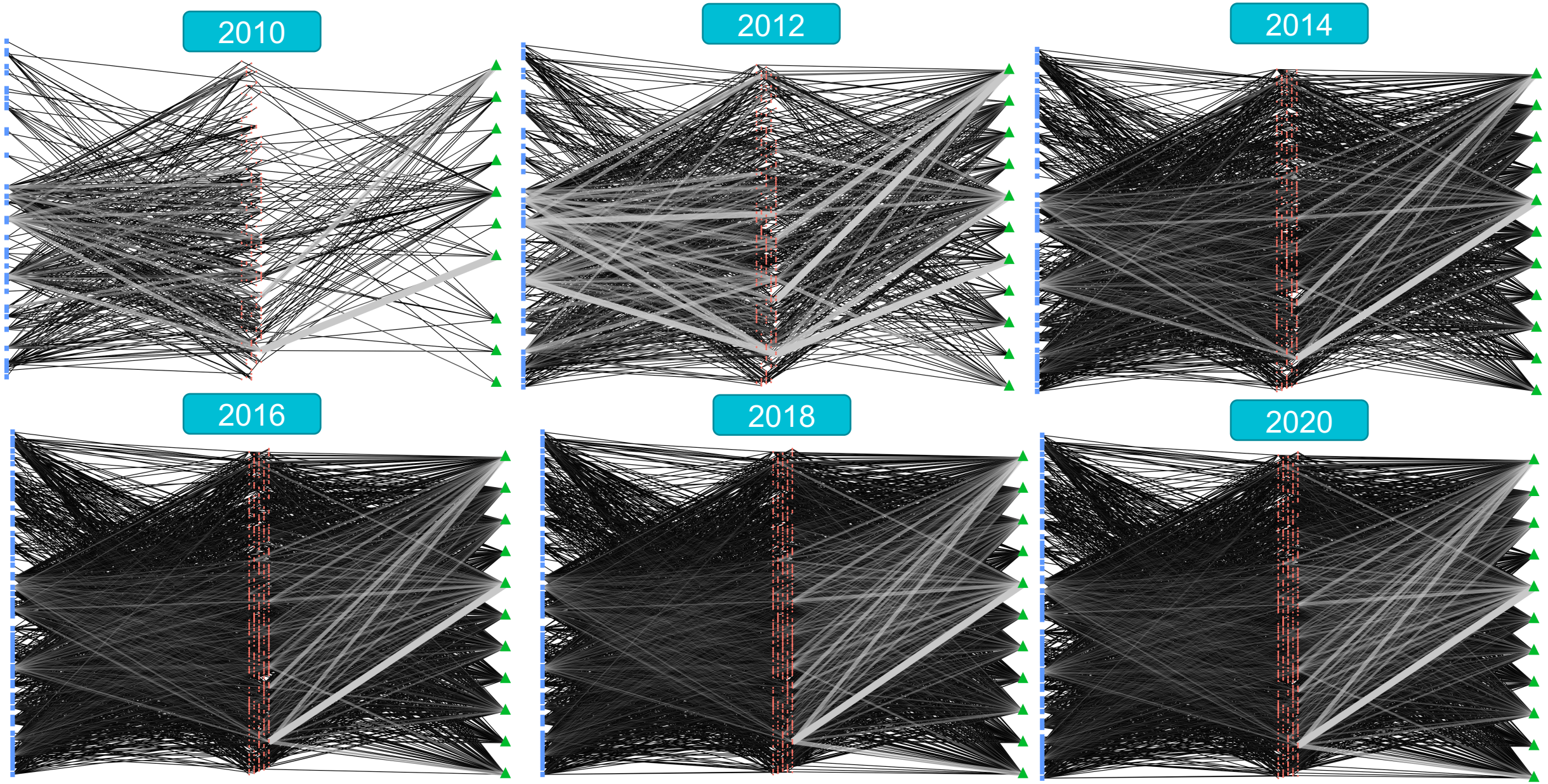
Source: Ipsos MORI and Technopolis

- To guide the evaluation, a theory of change was developed, which set out the full set of expected causal processes from MRC funded research to eventual impact that might be seen across the whole portfolio, while recognising that within single funding schemes a smaller set of pathways and processes might be relevant
- This framework highlighted likely differences in emphasis between the focused and enabling mechanisms within the directed translational portfolio.

Medical Research Council

Example of Logic Model taken from Ian Viney – Bridging the Gap: 10 Years of MRC Translational Research, 2020-02-24

N2. Behaviors over time



Questions?

Contact:

Gavin.Reddick@interfolio.com