Power Query / Power BI Instructions on recreating HRCS pivoted data from a single line format

Foreword – the purpose of this document

This document is designed to help those interested in recreating the UK Health Research Analysis methodology to assess research award portfolios. This is very much a draft document, but has been helpful for me to perform some additional analyses without the use of the out-dated MS Access conversion method.

While I've tried to keep things simple, the steps do require a reasonable working knowledge of Excel, and does require knowing how to create a pivot table. It also requires PowerQuery for Excel (standard in 2016, available in certain packages back to 2010) or the separate Power BI programme from Microsoft. No doubt other systems could do the same conversions, but these are the processes I've used and am most familiar with.

If you have queries about any of the steps in this document, you can use the contact us option of the HRCS website to get in touch. I would be happy to help you out.

Best wishes,

Jim Carter

Introduction - Award basics

At the core, the UK Health Research Analyses are simple pivot-based assessments of collated award information. The trickiest part is the amalgamation of many thousands of awards from multiple funding sources. However, if you are looking at individual organisations or simply a portfolio of awards, the basic conversion and pivoting is not too complex.

However, to be viable for analysis, this process does require some basic information about your research awards and is designed around that information being in a 'single line' format, i.e. that one row of an Excel spreadsheet contains all the information for a single award.

- **Grant References** – while not essential for analysis, a unique identifier will be the best way to track the additional metadata generated here back to the original awards.
- **Award start and end dates** – which can be used to create the timeframe for your analysis.
- **Award value** – generally we would recommend the total value of the award, be this the final commitment if the award is on-going, the expected expenditure at award end or the total expenditure to date[1].

[1] Note if that using expenditure to date may impact on some of the calculations associated with end date. It may be beneficial to adjust the 'end' date of the award to match the 'expenditure to' date.

- **HRCS coding** – column by column for Health Categories and Research Activities, including separate columns for percentage allocations.

These are the only mandatory data pre-requisites but obviously the more additional award information you have the more in-depth subsequent analysis can be. Given the process converts a single line format spreadsheet into a multi-line pivot-friendly version some information becomes duplicated. This means you will need to make some logical decisions about what type of analysis you want to achieve and how to prepare your data for this.

Optional – Additional filter columns

Depending on the number of and different types of organisations in an analysis, you may wish to apply some filter columns that help you quickly select groups of funders, awards or other data categories. These will be entirely dependent on the number of comparisons you wish to examine. In the 2014 analysis, we inserted the following into the combined dataset:

- **HRAF (Y/N)** – whether the funder was part of previous analyses (04/05 and 09/10)
- **AMRC (Y/N)** – whether the funder was a member of the AMRC
- **HRAFnewRO** – whether funder was existing member of HRAF, new charity funder (48 AMRC members beyond WT, BHF, CRUK & ARUK) or new other public funder (e.g. additional councils, Innovate UK)
- **Charity_Public** – categorises funders by status as Research Council, Charity of Other Public (Government) funder
- **InD2I (Y/N)** – whether medical charity funders also featured in the AMRC-commissioned *Donation to Innovation* analysis (2007).

These are just examples, but all of these columns allowed for easy filtering of the subsequent analysis data. If you have additional details to use as filters, I'd recommend making simple categories where you can as this can help speed along group aggregate comparisons.

RECREATING THE UK HEALTH RESEARCH ANALYSIS 2014 FOR EXCEL POWERQUERY / POWER BI

Step 1 – Create an 'active in 2014' value

Because the UK Health Research Analyses are based on the average commitment of awards in a given year, we need to determine when each award is active and apportion the total commitment accordingly. This is referred to as 'annualised commitment' in the reports. If you're using a different measure of award value (e.g. incurred spend in a given year) you may not have to perform this step. However the benefit is that if you have the three key criteria for each award - the Start Date, End Date and Total Value for the award – available, this methodology can be applied for multiple time periods, allowing for multivariant analysis.

1. First, calculate the overall duration of the award in days:
    a. If your data is formulated as dates, Excel should be able to do this by subtracting the Start Date from the End Date.[2]
    b. We'll call this column the Total Duration.
2. Next create five new columns;
    a. startpre14 - determines if the award starts before 2014
    b. endpost14 - determines if the award ends after 2014
    c. Duration2014 - how many days the award is active in 2014
    d. Value/Duration - calculate a financial per diem value (= Total Value / Total Duration)
    e. 2014Value - uses the latter two columns to calculate the annualised value for 2014.
    f. The formulae used to calculate these are below:

```
[StartPre14]=COUNTIF([START DATE],"<01/01/2014")
```

```
[EndPost14]=COUNTIF([END DATE],">31/12/2014")
```

StartPre	EndPost	Duration2014
0	0	=[DURATION]
0	1	=DATEDIF([START]-1,"01/01/2015","D")
1	0	=DATEDIF("31/12/2013",[END]+1,"D")
1	1	=365

```
Value/Duration=[Total Value]/[Total Duration]
```

```
2014Value=[Duration2014]*[Value/Duration]
```

Obviously, this is for **2014** specifically. If you wanted to do the same for other years, you would need to adjust the COUNTIF and DATEDIF calculation dates to match your revised criteria. This can potentially mean you have a whole series of additional columns in your spreadsheet to cover different periods (I've recently done 2012 to 2016, so annualised values for five different years).

[2] If this doesn't work you may need to do some calculations or conversions to ensure what looks like a date in excel is read as a date. There's lots of different ways to do this, so I won't go into them here, but the important thing is that the formulae won't work if the cells are not true dates in excel!

You'd need to carry all of these through into the next steps if you wanted to analyse multiple years simultaneously.

Step 2 – Calculate N values for number of codes

These columns will calculate the number of health categories and research activity codes that have been applied to each award, then uses these to determine a award code weighting. This makes the assumption that all codes have had the standard equally proportioned. If you have awards not equally apportioned, you will need to use a different methodology.

You need to do this twice, once for health categories and once for research activities, creating one new column for each. Use the following formula:

```
=IF(COUNTA(P2:T2)=0,0,(1/COUNTIF(P2:T2,"*")))
```

Where "P2:T2" is the cell range for the HC **or** RA entries (remember you need to do this once for each). We'll call these columns `N_HC` and `N_RA`. You'll see this essentially gives a decimal percentage value (1, 0.5, 0.333, 0;.25 etc.)

Once you have both values, create a third column for the combined 'number of HC/RA combinations', which we'll call this `N_allcodes` and is a simple calculation: `=N_HC*N_RA`

Finally, we can use the `N_allcodes` value to weight the annualised `2014Value` for the award equally across all HRCS codes. I'll call this the `AllcodesValue` and its another simple calculation: `=N_allcodes*2014Value`

Step 3 – Unpivoting

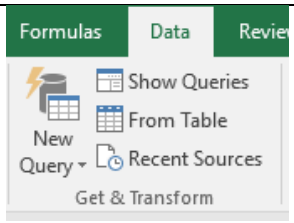
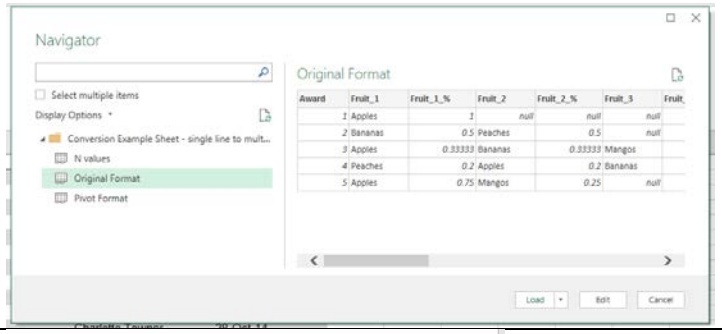
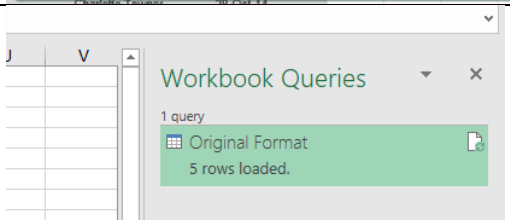
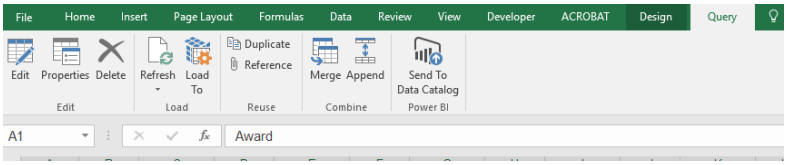
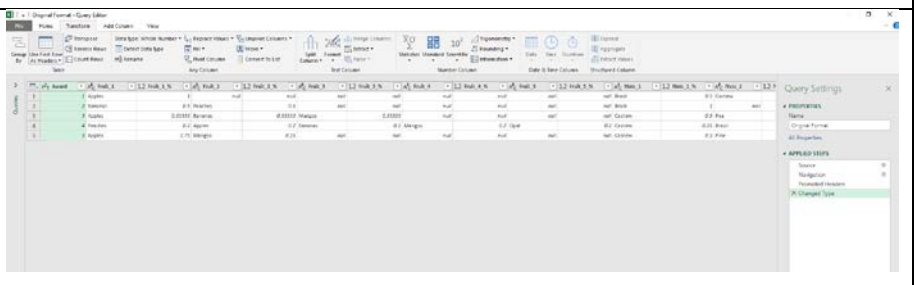
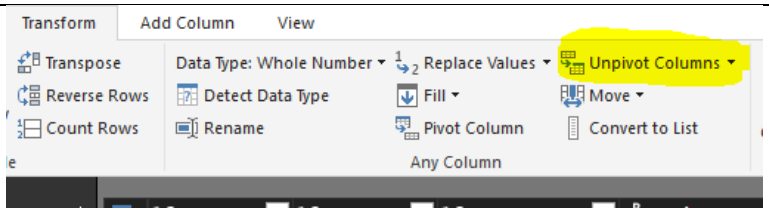
This is where you'll need either Power BI desktop version or access to PowerQuery (which comes as standard in Excel 2016, available as an add-in in earlier Excel versions).

To create the sort of charts in the UKCRC Analyses, the only columns you need to transfer will be:

- the Grant Reference (the unique identifier for your portfolio)
- all the HC and RA entry columns
- the `N_allcodes` column (see above)
- the `2014Value` column (see above)
- the `AllcodesValue` column (see above)

Power Query in Excel

PowerQuery and Power BI work in very similar ways. I am making the assumption that if you have downloaded / have access to Power BI you will be able to follow these Excel specific instructions too. Most of the menus and the same the steps required mostly match. Get in touch if you want to learn more.

Once you have the data ready to go, you select your file by selecting DATA and then NEW QUERY from the ribbon. You'll be asked to locate a file and select a worksheet. Once you have the right one, confirm by hitting the 'Load' button.	 
This should create a new worksheet and convert your data into a table format. You should also see the Query panel appear on the right side of the screen. To edit the Query, you can right click from this panel and select 'edit' or use the new 'Query Tool' entry on the ribbon and the edit button.	 
This will open a new view of your data, the Query Editor, which has a different ribbon across the top and a Query Settings panel on the right hand side.	
Select all of your Health Category columns, then choose 'Transform' tab from the ribbon, and the 'Unpivot Columns' option.	
NOTE this has to be done separately for Health Categories, and then again for Research Activities	

This converts the single line HC1, HC2, HC3, etc. or RA1, RA2 etc. columns into a vertical pivot friendly version, and duplicates all other associated cells (i.e. grant reference, N and Values) – note this can rapid increase the size of your worksheet by adding many rows as the number of codes increases.

Essentially this is now ready for analysis, where you can compare your new unpivoted values against the calculated values you've created.

I'm not going to explain how to make a pivot table in Excel. If you've read this far I'm guessing you will know how to do this!

As a sense check, you should find your sum of the Nallcodes column and the sum of the AllcodesValue column match the total number of data rows and the sum of the Total Value column from your original single line spreadsheet. If it doesn't… well something has gone wrong and it would be difficult for me to say where exactly. Look for obvious errors (e.g. deleted cells, missing data) but beyond that I'm not sure I can help via a document!

Optional – RA Groups

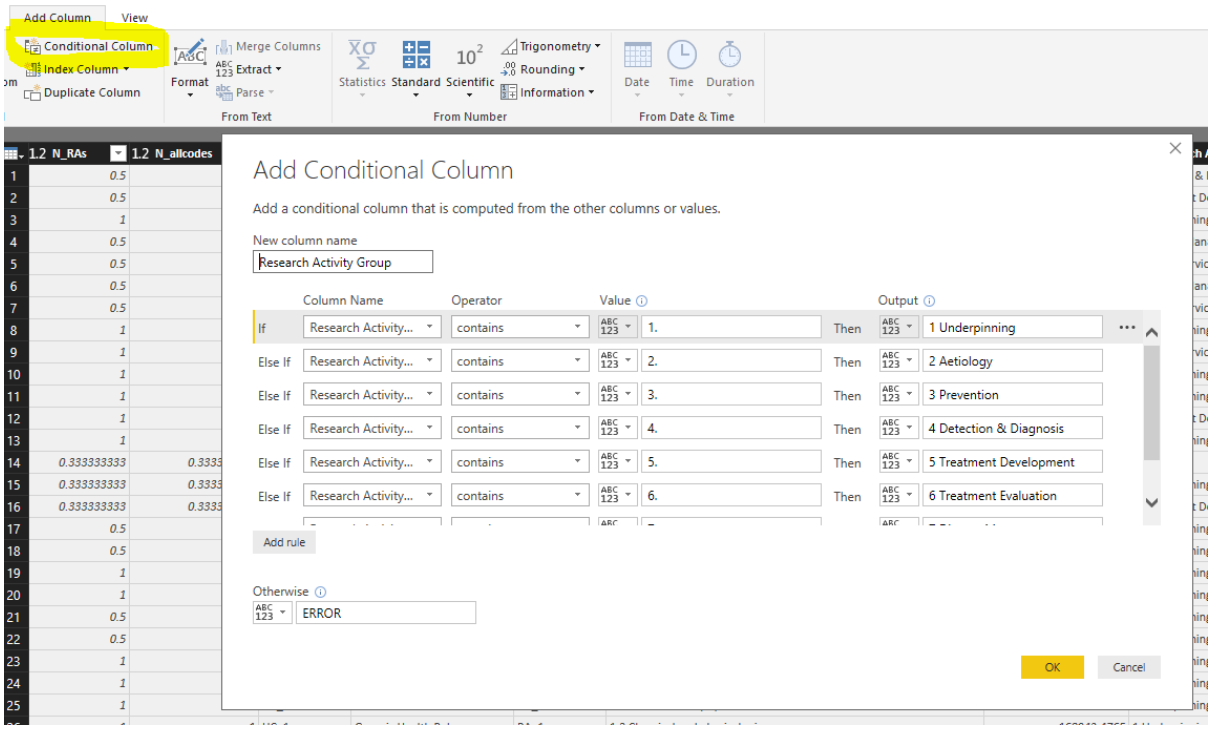
The data for HRCS coding is usually at the sub-code level (i.e. 1.2, 3.4, 8.1 etc.). However much of the broad scope analysis just examines the upper group level codes, i.e. 1-8. If you don't have these saved separately already, you can use the following methods to get what you want:

In Excel – use the complicated formula below in a new column, replacing M2 with whichever cell contains your sub-code:

```
=IFS(ISNUMBER(SEARCH("1.",M2)),"1 Underpinning",ISNUMBER(SEARCH("2.",M2)),"2
Aetiology",ISNUMBER(SEARCH("3.",M2)),"3 Prevention",ISNUMBER(SEARCH("4.",M2)),"4
Detection & Diagnosis",ISNUMBER(SEARCH("5.",M2)),"5 Treatment
Development",ISNUMBER(SEARCH("6.",M2)),"6 Treatment
Evaluation",ISNUMBER(SEARCH("7.",M2)),"7 Disease
Management",ISNUMBER(SEARCH("8.",M2)),"8 Health Services")
```

Then just fill down to complete this for all your rows of data.

In Power Query / Power BI – use the conditional column option, then the 'contains' rules as in the screenshot below:



Note: the ‘contains’ operator is only available if the Research Activity Code column is text, so you may have to convert this first. However I have found that converting the 1.1, 1.2 values to text can sometimes cause an error where it appears as 1.099999999 or something similarly bizarre. I’m not sure exactly why, maybe an issue with the original cell format, but it does mean if you wanted to do analysis at the RAC level you’d need to then go back and clean these up… which if there’s lots is VERY annoying.