Involvement in the HRCS Report 2014: Frequently Asked Questions

Executive Summary

This document is intended as a lay explanation to help clarify some of the key areas of the HRCS. In particular, this gives some additional detail into the planning for the production of the 2014 report.

This FAQ has been produced by the MRC Project Management Team, and approved by the HRAF.

The Basics

Who are the UKCRC and what is the HRCS?

The UK Clinical Research Collaboration (UKCRC) is a group of representatives from the major public, Industry and charitable funders of UK Health Research. It's role is to co-ordinate strategic approaches to clinical research to benefit patients, researchers and the nation. In 2004, the UKCRC established the bespoke Health Research Classification System (HRCS) to categorise all types of health research across all disease and areas of health. More information is available on the UKCRC website (www.ukcrc.org) and HRCS website (www.hrcsonline.net).

Who are the HRAF?

The UKCRC Health Research Analysis Forum (HRAF) is a group of representatives from the 12 largest funders of public and charitable health research and the Association of Medical Research Charities (AMRC), representing 133 medical research charities. The forum was established in 2009 to take over the maintenance and development of the HRCS. This includes the co-ordination and production of the HRCS Reports.

What is the HRCS Report 2014?

The 2014 report is the third in a set of quinquennial (5-yearly) reports on Health Research funding in the UK. The AMRC also commissioned a report in 2007 entitled "From Donation to Innovation". This covered 29 further charities not included in the earlier UKCRC analysis. More information on HRCS reports are on the website

(www.hrcsonline.net). For the 2014 report the HRAF aims to include data not only from the 12 largest funders, but as many other UK health research funders as is possible.

I am an AMRC member, and I've not been in previous HRCS reports. Why am I being asked to be involved?

Since the last report in 2010, the use of HRCS coding as a more routine procedure for many funders. This includes AMRC, who perform HRCS coding routinely on behalf of its members. As a result, the 2014 report will differ from previous analyses. This year, it is our hope to include the 12 HRAF members, the 29 charities (from the 2007 report), and as many other AMRC members as possible. This would make this report the largest and most complete analysis of the UK health research to date.

Please note if you are an AMRC member not on the HRAF group of funders and have any specific questions not answered by this document, please get in touch with Gemma Luck at AMRC (g.luck@amrc.org.uk).

What does participation in the 2014 report involve?

To be part of the analysis, we will need you to provide details of your awards that were/are active in the calendar year 2014 (1/1/14 to 31/12/14).

If you are an AMRC member and not part of the HRAF: you may have already provided this information to the AMRC in which case the AMRC will ask your permission to share the information with the project. If unsure, let Gemma at AMRC know if you would like to be involved in the project, they can arrange access to your coded award portfolio kept in the AMRC database. AMRC collects this information from members during its single data request each year and can therefore provide access to this on your behalf, mitigating extra work for smaller organisations.

What data fields are required for submission?

The list of data fields required for the analysis is included as a table in Annex 1 at the end of this document. There are some details of these fields within the table, however a full explanation for each field can be found in the accompanying HRCS Report 2014 Submission Guidance.

The requirements list may seem lengthy, but it's important to remember that submission of these details has been part of the HRCS analysis before. In the 2004/05 report, the same fields were required for de-duplication and disambiguation purposes and will be used for this same purpose in the 2014 report.

What about data publicity? Who will be able to see the final data?

As with previous reports, the final dataset (a spreadsheet of each award included in the analysis) will be made publically available via the HRCS website.

In previous years, this data has been relatively limited, featuring only HRCS standardised grant ID, funder, HRCS codes and institution/region. However, in the five

years since the last report was published, more and more data on awards have become publically available. Research Councils make all awards available via Gateway to Research. In addition, many charities also make their own awards available, either via their own websites or through centralised databases such as EuroPubMed. This includes greater levels of detail, including grant codes, titles and abstracts.

From a data analysis perspective, there are distinct benefits to having a greater amount of data publically available, as it allows for further analysis of awards. For example, a sub group analysis of mental health funding specifically for 'depression' can be achieved using similar keyword/datamining techniques used in auto-coding (see the sections on UberResearch's involvement in the project for more on this topic). As a result a publically accessible dataset would allow the HRCS report to become an even more valuable resource than it already is.

Therefore in keeping with most participating funders' support for open accessibility of data, and the potential post-analysis benefits of doing so, our aim is to make final data, including grant codes, titles and abstracts, publically available.

Please note that this does not include every data field. Annex 1 includes a column to show which fields will be included in the publically available final version.

There will be an option for specific exceptions where funders have reasons to exclude details – e.g. where awards identify personal details (e.g. salaries), contain confidential details (e.g. patent details) or there are concerns about the safety of researchers (e.g. highly detailed in vivo experiments).

If you have concerns regarding data publicity please contact the HRCS Project Team. If you are a non-HRAF AMRC member, please direct your initial enquiries via Gemma at AMRC.

What is the timeline for the project?
We are currently preparing to receive the first and largest part of the dataset from funders, termed the 'work in progress' (WIP) dataset. This captures all the currently available awards. The 'final' dataset needs to be provided by the 13th of March 2015, and will contain the remaining awards not included in the WIP dataset (e.g. awards activated later in 2014). Data analysis will continue March-April, with the draft report going to the UKCRC board at the beginning of May 2015. Our aim is to have a final published version available in June/July 2015.

The Coding

How is the award coding being completed?

In some cases, funders have systems in place for coding their awards periodically. However, as we are requesting full award data for the complete 2014 calendar year (01/01/14 to 31/12/14) we may need to 'code as we go' for the more recent awards. Our estimate for total number of awards that would be eligible from all funders is between 12,000 to 15,000 awards.

How will up to 15,000 awards be coded?

The good news is that a lot of this coding has already been done. Because funders with the majority of awards (and AMRC) code routinely, we are not in the same situation as previous reports where all awards have had to be coded from scratch. From our survey of HRAF members, approximately 75% of awards are already coded. Of the remaining 25%, around half will be new awards still being processed (e.g. from late 2014 funding calls).

How do we submit data?

Data submission guidelines that accompany this FAQ are available on the HRCS website, http://www.hrcsonline.net/pages/data.

Data submissions for HRAF members are achieved via a standardised excel spreadsheet, emailed to the project management team. AMRC members who would like their data included in the analysis but are not part of the HRAF will have data submitted collectively via AMRC, not directly to the project.

What is second pass or quality control (QC) coding?

In addition to basic 'first pass' coding, some quality control (QC) will be carried out to assure an overall level of coding consistency for the project.

Our target for the 2014 report is to "second pass" code 40% of awards included in the analysis. Our aim is to do this second coding 'blinded', meaning the person carrying out the second pass coding is both independent (i.e. not the same person who first pass coded) and unaware of what the previous coding was. Once this is complete, we can compare first pass coding with second pass QC coding, and resolve any discrepancies.

Does someone else second pass coding my awards mean I'm being 'tested' in HRCS?

No. Our primary goal for second pass coding is to ensure the final data is as consistent as possible. It is not intended as a critique or examination of individual coders.

No coding approach will be entirely objective, nor will guidelines for using it be completely unambiguous. The application of the HRCS to award abstracts is open to individual interpretations. One of the key secondary aims of the project is to review and update the HRCS. Therefore finding areas where there is variation in opinion between

coders is one way in which we can improve guidance and ensure future coding becomes more accurate in the future.

What if my funder already has coding check / QC procedures?

Again, many funders will already have their own procedures for checking coding accuracy. It is not our intention for coders to repeat work. Therefore so as long as there is a clear procedure for doing so, we will happily accept already QC'd coding, which will go towards to the 40% second pass coded aim.

For those that do not, we will be asking them to participate in a communal second pass coding process. See "how does the communal coding process work?" for more details.

Who will be doing the first and second pass coding?

There are many people working for different funders who already use the HRCS and experienced in applying HRCS coding. In addition, part of the aim of the project is to provide additional training sessions to provide new coders with the knowledge and skills to apply HRCS.

We will therefore be asking all coders, those already trained and those newly trained, to assist with the completion of the data collection process.

What extra workload is involved?

Principally this will be doing what coders already do as part of their normal work, making sure the awards are all fully coded. However it is true that some additional work will be needed to help meet our 40% second pass coding rate.

This may seem daunting at first, but it is important to remember the more who participate, the easier it becomes for all. If every coder we know of contributed equally, each person would only need to help with between 100 to 200 awards. This will mean a few hours of 'extra' work, but spread at your own pace between now and the end of the data collection period (13th of March 2015).

Please note that it is our aim to ensure that coders carrying out second pass coding will do so from their own funder's awards. However, due to the limited number of coders, we may sometimes have to ask those involved to second pass code awards from other funders.

The QC Coding, UberResearch and Data Publicity

How does the communal coding process work?

As previously mentioned, for the majority of funders the first pass coding will be largely complete. We would expect this process to continue to ensure the majority of data we receive is already first pass coded (or will be in advance of final submission deadlines). How this is achieved will vary from funder to funder, and we would not wish to interfere with this process.

The second pass coding is handled differently:
- All data will be submitted to the MRC project management team in exactly the same way as previous reports.
- Of the total awards submitted, a randomised selection totalling 40% will be tagged for second pass coding.
- If any awards have already been second pass coded, as long as we have been informed, we can count those towards the 40% and avoid repeating coding work.
- To ensure blinded coding, awards will be split into work packages for individual coders to process.
- Wherever possible, the second pass coder will be different to the first pass coder.
- The second pass coder will be asked to complete their assigned awards before the end of data collection; the 13th of March 2015.
- Results of second pass coding will be compared to first pass coding, and any discrepancies resolved before the final coding is submitted as part of the analysis.

There are several important notes concerning this process:
- Wherever possible, we will ensure that coders carrying out second pass coding will do so from their own funder's awards.
- However, because of the communal nature of the second pass coding 'pool', the limited number of coders available and the need to keep coding independent, we may have to assign awards to coders from different funders.

Who are UberResearch and how are they involved in the process?

UberResearch are an independent company who develop decision support systems for science funding organisations. Their primary focus is allowing funders to examine their portfolios, compare their funding to others, and facilitate finding personnel for recruitment and/or peer review purposes.

The HRAF as a whole, and as individual organisations, have been in talks with UberResearch for some time over the development of an automated coding system for predicting HRCS Health Categories. The algorithm developed by UberResearch has now be refined to provide an approximately 90% accurate prediction rate.

As a result of this, we have initiated a specific collaboration with UberResearch to facilitate the second pass coding process. UberResearch have developed, without commitment from funders to subscribe to any services, a coder interface where we can upload communal data, segregate assigned work packages to individual coders, and monitor the progress of the coding. UberResearch are providing this interface to the

project at no cost to funders for the purpose of completing the coding exercise for the 2014 analysis.

The benefits to the coder:

- A log in to the UberResearch interface, from which coding can be carried out at your own pace.
- Each award will have the predicted Health Categories. This should speed up the process of coding each award individually as you chose to accept, modify or completely edit the predicted categories.
- Please note that Research Activity coding will still have to be done manually for each award, within the UberResearch system.
- Access to the UberResearch site will also allow coders (and the funders they represent) to use the Dimensions system free of charge for the duration of the project.

The benefit to the Project:

- We have a centralised system for distributing and monitoring the whole second pass coding process.
- This would otherwise be organised manually via the distribution of spreadsheets, which has an inherent risk of duplications and/or errors.
- By having a central source from which individual work packages have been distributed, we can monitor each coder's progress award by award, allowing us to spot potential problems early and help when needed.
- Combined, this greatly enhances the management of the process, which should prevent the delays and confusions that have hampered previous reports.

The benefits to UberResearch:

- By having multiple users checking and correcting predicted codes, UberResearch gain the information they need to further improve the algorithms used to predict the Health Categories, making the whole process better for future use.
- By providing the access free to all participants, UberResearch get potential customers aware of the system as a whole, and its capabilities. The system will remain free for all users for the duration of the project.

Please note that neither the MRC nor the HRAF specifically endorse UberResearch. It will be up to individual funders at the end of the project if they wish to continue to use UberResearch systems, and the subscription costs that access entails.

I'm concerned about data publicity. UberResearch requires more data, including Principle Investigator (PI) name for their database. Why is this, and will this be made publically available too?

Submission of these details has been part of the HRCS analysis before. In the 2004/05 report, the same fields were required for de-duplication and disambiguation purposes and will be used for this same purpose in the 2014 report.

UberResearch submissions require this data for the same 'data cleaning' reasons, and to allow our HRCS data to be correctly uploaded to their system. In addition, as UberResearch's work involves helping funders find the right researchers, this is information is vital for all awards that migrate to the main Dimensions database.

There are three important caveats to this process:
- Data will only be migrated across if the funders who own the data agree. As such, each funder can submit a full portfolio to the system for second pass coding, but ask that it be removed at the end of the project.
- All the information in the main UberResearch system comes from publically accessible systems (Gateway to Research, EuroPubMed, etc.) and as such most funders will find that this information is already in the public domain. However these data may not be as up to date as those submitted for the HRCS report.
- The data transferred to the Dimensions database will only be available to UberResearch's subscribers, i.e. other research funders. This information is not available to the public as a whole.

PI name will not be part of the dataset made public via the HRCS website. Therefore only funders who agree to the migration to the Dimensions system will see their PI details being made (semi-)public.

It is our hope that, given the caveats above, all funders will agree to provide a complete portfolio including those fields required for upload to UberResearch. Further details on data publicity, including UberResearch's data sharing agreement, are available in the accompanying Data Publicity Statement.

Contact Details

If you have any questions regarding the HRCS 2014 Project, please contact Dr. Jim Carter at the MRC via james.carter@headoffice.mrc.ac.uk.

If you are an AMRC member, you can also contact Gemma Luck (g.luck@amrc.org.uk) for queries regarding the role of the AMRC in co-ordinating data collection/submission from their membership.

Annex 1 – HRCS 2014 Data Fields List/Dictionary

Excel Column	Field Name	Purpose	Notes/ Advice	Data type	Required?	Public in HRCS Datset?
A	FundingOrganisation	Full name of partner organisation	Fixed value for each record	Text	Yes	Yes
B	FunderAcronym	The acronym by with the funder is known.		Text	Yes	Yes
C	OrganisationReference (aka grant/award code/ID)	Internal ID used by partner organisation	Unique value for each record.	Text	Yes	Yes
D	PITitle	Title of the award lead investigator	Dr, Professor, Mr, Mrs etc.	Text	No	No
E	PIFirstName	First name of the award lead investigator		Text	Yes	No
F	PIMiddleName	Middle name of the award lead investigator		Text	Yes	No
G	PISurname	Last name of the award lead investigator		Text	Yes	No
H	PIInstitution	Host institution of the award lead investigator	Full institution name (not abbreviation)	Text	Yes	Yes
I	PIAddressLine1 (aka department)	First address line of award lead investigator	For most researcher s this will be their department	Text	Yes	No
J	PIAddressLine2	Second address line of award lead investigator		Text	No	No
K	PIAddressLine3	Third address line of award lead investigator		Text	No	No
L	PICity	City of award lead investigator		Text	Yes	Yes
M	PIPostcode	Postcode of award lead investigator		Text	Yes	No
N	PICountry	Country of award lead investigator	Will be excluded from analysis if award is funded outside UK	Text	Yes	No
O	PIEmail	Email address of the award lead investigator		Text	Yes	No
P	FundingMechanism	Type of award made	Name of research programme or funding scheme used by partner organisation e.g. fellowship, project, programme, unit, institute	Text	No	Yes (if available)
Q	FundingStream	Funding stream which supports the award made	Name of board, reviewing panel or funding stream under which the award was made	Text	No	Yes (if available)

Excel Column	Field Name	Purpose	Notes/ Advice	Data type	Required?	Public in HRCS Datset?
R	StartDate	Award funding start date	Date when award spending commences. Preferred format is dd/mm/yyyy.	Date	Yes	Yes
S	EndDate	Award funding end date	Date when award is completed NB StartDate + Duration = EndDate. Preferred format is dd/mm/yyyy.	Date	Yes	Yes
T	Duration	Duration of awarded funding in months	NB StartDate + Duration = EndDate	Integer	Yes	Yes
U	TotalAward	Total funding for duration of award		Currency	TotalAward or AnnualAward	Yes
V	AnnualAward	Amount awarded per annum		Currency	TotalAward or AnnualAward	Yes
W	AwardTitle1	Title of the award or abstract	Full title	Text	Yes	(Yes)
X	AwardAbstract2	Scientific abstract of the award	Usually 200-300 words.	Note / Memo	Yes	(Yes)
Y	Keywords	Partner specific keyword descriptions	e.g.- MeSH keywords	Text	No	Yes
Z	IndirectAward	To segregate out those awards not easily classified/ coded.	Options are: *<leaveblank> *Infrastructure *Personal *Missing/ Incomplete *Non-Health Research	Limited Text	Yes	Yes
AA	CoderComment	Include any additional explanatory text here	Leave blank if no issues with award or coding.	Text	Yes	No
AB	CoderName	To track coding process, and ensure, where possible, any QC coding is independent (i.e. sent to a different coder)	Leave blank if coder is unknown or cannot be traced to specific person.	Text	Yes	No
AC	QCCodingApplied	If any QC 'second pass' coding has already taken place, indicate it here so we avoid unwarranted 'third pass' coding.	Leave blank if unknown or no QC coding has been carried out	Text (Yes/No)	Yes	No

---

[1] Award Title and Abstract need to be made public to allow sub-analysis of the dataset, e.g. by keyword search. However, if these fields contain sensitive information, such as details of in vivo experiments, patent details or copyrighted materials, we will ensure the data is anonymised.

[2] Scientific abstracts only. Please note that lay abstracts will not be suitable for coding purposes unless there is no scientific abstract available.

Excel Column	Field Name	Purpose	Notes/ Advice	Data type	Required?	Public in HRCS Datset?
AD	AssignedUberUser	Admin only: awards selected for QC second pass coding via UberResearch system need to be defined as part of the spreadsheet before importing.	n/a, for admin purposes only.	UserNa me	Admin only	No
AE	RA_1				Yes	Yes
AF	RA_1%				Yes	Yes
AG	RA_2				Yes	Yes
AH	RA_2%				Yes	Yes
AI	RA_3		For large awards only, see HRCS guidance		Yes	Yes
AJ	RA_3%				Yes	Yes
AK	RA_4		For large awards only, see HRCS guidance		Yes	Yes
AL	RA_4%				Yes	Yes
AM	HC_1				Yes	Yes
AN	HC_1%				Yes	Yes
AO	HC_2				Yes	Yes
AP	HC_2%				Yes	Yes
AQ	HC_3				Yes	Yes
AR	HC_3%				Yes	Yes
AS	HC_4				Yes	Yes
AT	HC_4%				Yes	Yes
AU	HC_5				Yes	Yes
AV	HC_5%				Yes	Yes