GUIDANCE ON DATA CLEANING FOR THE NHS COMMUNITY MENTAL HEALTH SURVEY 2014

THE CO-ORDINATION CENTRE FOR THE NHS SURVEYS PROGRAMME

picker
Institute Europe

Making patients' views count

Last updated: September 2014

Contacts

The Patient Survey Co-ordination Centre
Picker Institute Europe
Buxton Court
3 West Way
Oxford
OX2 0BJ

Tel: 01865 208127
Fax: 01865 208101
E-mail: mentalhealth.data@pickereurope.ac.uk
Website: www.nhssurveys.org

Key personnel

Chris Graham (Director of Research & Policy)

Hannah Atherton
Caroline Killpack
Susie Stevenson
Mark Waters

Updates

Before you use this document, check that you have the latest version, as there might be some small amendments from time to time (the date of the last update is on the front page). In the very unlikely event that there are any major changes, we will e-mail all trust contacts and contractors directly to inform them of the change.

This document is available from the Co-ordination Centre website at:

www.NHSSurveys.org

Questions and comments

If you have any questions or concerns regarding this document, or if you have any specific queries regarding the submission of data, please contact the Co-ordination Centre:

By e-mail: Mentalhealth.data@PickerEurope.ac.uk

By phone: 01865 208127

Contents

1 Mental Health Survey 2014 . data cleaning

1.1 Introduction

Once fieldwork for the 2014 Community Mental Health survey has been completed, participating trusts and contractors are required to submit data to the Co-ordination Centre in a **raw ('uncleaned')** format. Following this, data for all trusts in the survey is collated and the full dataset is cleaned together to ensure that cleaning taking place on data for each trust is comparable.

This document provides a description and specification of the processes that is used by the Co-ordination Centre to clean and standardise data submitted by contractors and trusts as part of the 2014 Community Mental Health survey. By following the guidance contained in this document, it should be possible to recreate this cleaning process.

If you have any comments or queries regarding this document please contact the Co-ordination Centre on **01865 208127**, or e-mail us at **mentalhealth.data@pickereurope.ac.uk**.

1.2 Definitions

Definitions of terms commonly used in this document, as they apply to the Community Mental Health survey are as follows:

Raw/uncleaned data: ±Rawqor ±uncleanedqdata is data that has been entered verbatim from completed questionnaires without any editing taking place to remove contradictory or inappropriate responses. All responses ticked on the questionnaire should be included in the data entry spreadsheet[1] (see Section 2: Submitting raw (±uncleanedq) data, for detailed guidance on submitting raw data). The requirement for raw/uncleaned data does **not**, however, preclude the checking of data for errors resulting from problems with data entry or similar. Ensuring high data quality is paramount and errors resulting from data entry problems can and should be corrected by checking against the appropriate completed questionnaire.

Data cleaning: The Co-ordination Centre uses the term ±data cleaningqto refer to all editing processes applied to survey data once the survey has been completed and the data has been entered and collated.

Routing questions: These are items on the questionnaire which instruct respondents either to continue on to the next question or to skip past irrelevant questions depending on their response to the routing question. For the 2014 Community Mental Health survey, the routing questions in the questionnaire are **Q8, Q12, Q15, Q18, Q21, Q22, Q24, Q26, Q28, Q30.**

Please note that although Question 1 is technically a routing question, any respondents who tick option 7 ('I have never seen anyone from NHS mental health services') should be recoded as '5' (ineligible).

---

[1] Except where: a) multiple responses have been ticked - set these to missing (the **exceptions** to this are for the ±tick all that applyqquestion **Q9** where respondents may tick more than one response option); b) year of birth has been entered in incorrect format - if someones *intended* response is unambiguous from the questionnaire, then enter this.

Filtered questions: Items on the questionnaire that are not intended to be answered by all respondents are referred to as 'filtered' questions. Whether individual respondents are expected to answer filtered questions depends on their responses to preceding routing questions. For the 2014 Community Mental Health survey, the filtered questions in the questionnaire are **Q9- Q11, Q13- Q14, Q16- Q17, Q19- Q20, Q23, Q25-Q29, Q31.**

Non-filtered questions: these are items in the questionnaire which are not subject to any filtering and which should therefore be answered by all respondents (except those who ticked response option 7 at Q1). For the 2014 Community Mental Health survey, the non-filtered questions are **Q1- Q9**, **Q12, Q15, Q18, Q21, Q24, Q30, Q32-Q48.**

Out-of-range data: This refers to instances where data within a variable has a value that is not permissible. For categorical data . as in the case of the majority of variables in this survey . this would mean a value not allowed in the data, for example, a value of '3' being entered in a variable with only two response categories (1 or 2). For scalar data . eg year of birth . data is considered to be out-of-range if it specifies a value that is not possible (for instance, year of birth as 983 or 2983). Out-of-range responses entered into the dataset **should not** be automatically (eg, algorithmically) removed prior to submitting the data to the Co-ordination Centre (see Section 2: Submitting raw ('uncleaned') data.

Non-specific response: This is a loose term for response options that can be considered as not being applicable to the respondent in terms of directly answering the specific question to which they are linked. Most commonly, these are responses such as 'don't know/can't remember', which indicate a failure to recall the issue in question. Likewise, responses that indicate the question is not applicable to the respondent are considered 'non-specific'. for example, responses such as 'I did not need any support' or 'I do not have any caring responsibilites'. A full list of such responses for the 2014 Community Mental Health survey can be found in Appendix B: Non-specific responses.

2 Submitting raw (±uncleaned) data

For the 2014 Community Mental Health Service Users survey, trusts and contractors are required to submit raw (±uncleaned) data to the Co-ordination Centre. For clarification, raw data is created as follows:

i)   All responses should be entered into the dataset, regardless of whether or not the respondent was meant to respond to the question (eg, where people answer questions that they have been directed to skip past, these responses should still be entered)

ii) Where a respondent has ticked more than one response category on a question, this should be set to ±missing in the data. The **exception** to this is the ±multiple response question, **Q9**, where respondents may tick more than one response option (i.e. ±tick all that apply)

iii) Where a respondent has crossed out a response, this should not be entered in the data. Where a respondent has crossed out a response and instead ticked a second response option, this second choice should be entered into the data

iv) Where a respondent has given their response inconsistently with the formatting of the questionnaire but where their intended response is nonetheless unambiguous on inspection of the completed questionnaire, then the respondents *intended* response should be entered. For example, where a patient has written their *date* of birth in the boxes for **Q46**, but written their *year* of birth in at the side of this, then the respondents year of birth should be entered.

v)   For the year of birth question, unrealistic responses should still be entered *except* following **iv)** above. For example, if a respondent enters ±2014 in the year of birth box, this should still be entered unless the respondent has unambiguously indicated their actual year of birth to the side

vi) Once the data has been entered, no responses should be removed or changed in any way *except* where responses are known to have been entered incorrectly or where inspection of the questionnaire indicates that the ssomeones intended response has not been captured. This includes ±out-of-range responses, which must **not** be *automatically* removed from the dataset. Responses in the dataset should only be changed before submission to the Co-ordination Centre where they are found to have been entered inconsistently with the respondents intended response.

3 Editing/cleaning data after submission

3.1 Approach and rationale

The aim of the Co-ordination Centre in cleaning the data submitted to us is to ensure an optimal balance between data quality and completeness. Thus, we seek to remove responses that are known to be erroneous or inappropriate but do this in a relatively permissive way so as to enable as many responses as possible to contribute to the overall survey results.

3.2 Filters

Some of the questions included in the survey are only relevant to a subset of respondents, and in these cases filter instructions on the questionnaire are used to route respondents past questions that are not applicable to them. For example, people who tick "No" to **Q12** ("Have you agreed with someone from NHS mental health services what care you will receive?") are instructed to skip all further questions on planning care (**Q13-Q14**).

It is necessary to clean the data to remove inappropriate responses where filter instructions have been incorrectly followed. In such cases, participants' responses to questions that were not relevant to them are deleted from the dataset. Responses are only deleted where respondents have answered 'filtered' questions despite ticking an earlier response on a routing question instructing them to skip these questions (eg a respondent ticking "No" to **Q12** but then answering the two questions about planning care as in the example above). Responses to 'filtered' questions are not removed where the response to the routing question is missing.   For example, **Q19-Q20** are filtered by the response to **Q18** (e.g. if **Q18=1**), but if a respondent does not answer **Q18**, or if the **Q18** response is missing for any reason, then responses to **Q19-Q20** should not be removed.

Figure 1 (overleaf) shows a summary of all routing questions, and the filtered questions they relate to, that are included in the 2014 Community Mental Health survey. Please note that these instructions should be followed in the order provided to be consistent with the procedures applied by the Co-ordination Centre.

Figure 1 - List of routing/filtering instructions

ROUTING QUESTION		RESPONSE VALUES		FILTERED QUESTIONS
if	**Q1** =	7	*Recode as outcome=5 (ineligible)*	
if	**Q8** =	2 OR 3	*then delete responses to:*	**Q9 - Q11**
if	**Q12** =	3	*then delete responses to:*	**Q13 – Q14**
if	**Q15** =	2 OR 3	*then delete responses to:*	**Q16- Q17**
if	**Q18** =	2 OR 3 OR 4 OR 5	*then delete responses to:*	**Q19- Q20**
if	**Q21** =	2 OR 3	*then delete responses to:*	**Q22- Q23**
if	**Q22** =	2 OR 3	*then delete responses to:*	**Q23**
if	**Q24** =	2	*then delete responses to:*	**Q25- Q29**
if	**Q26** =	2	*then delete responses to:*	**Q27**
If	**Q28** =	2 OR 3	*then delete responses to:*	**Q29**
if	**Q30** =	2 OR 3 OR 4 OR 5	*then delete responses to:*	**Q31**

Please note that these instructions should be followed in the order shown above.

A worked example of the cleaning process for removing unexpected responses to filtered questions is included in Appendix A: Example of cleaning.

3.3 Dealing with multiple response questions

For most questions, each column corresponds to one survey question. However, there is one exception to this rule. For the multiple response question **Q9**, each response option is treated as a separate question.

Example

Q9. Is this person in charge of organising your care and services… ..

   ₁ ☑ A CPN (Community Psychiatric Nurse)

   ₂ ☐ A psychotherapist / counsellor

   ₃ ☐ A social worker

   ₄ ☐ A psychiatrist

   ₅ ☑ A mental health support worker

   ₆ ☐ A GP

   ₇ ☐ Another type of NHS health or social care worker

   ₈ ☐ Don't know

Responses to each part of this question are coded: **1 if the box is ticked**
0 if the box is not ticked[2]

Q9 takes up eight columns in the data file, labelled as follows:

Column headings	Q9_1	Q9_2	Q9_3	Q9_4	Q9_5	Q9_6	Q9_7	Q9_8
Codings for this example	1	0	0	0	1	0	0	0

---

[2] Please note: if a respondent does not answer any part of a multiple response question, (ie does not tick any of the response options) then it should be left blank or coded as a full stop (.)

3.4 Demographics

Basic demographic information is included in the sample section of the data, but the ±About You‡ section at the end of the questionnaire also asks respondents to provide this information. In a minority of cases, the information provided from the sample frame and by the respondents does not correspond on age and gender . for example, the sample may identify an individual as male only for them to report being female (e.g. **Q45=2**).

Because of this, and because questions about demographics tend to produce relatively high item non-response rates, it is not appropriate to rely on either source of data alone.

Where responses to demographic questions are present, it is assumed these are more likely to be accurate than sample frame information (since it is assumed that respondents are best placed to know their own gender and age)[3]. Where responses to demographic questions are missing, however, sample data are used in their place.

For demographic analysis on groups of cases, then, it is necessary to use some combination of the information supplied in the sample frame and by the respondents. To do this, we first copy all valid responses to survey demographic questions into a new variable. Where responses are missing we then copy in the relevant sample information (note that for a very small number of respondents demographic information may be missing in both the sample and response sections; in such cases data must necessarily be left missing in the new variable)[4].

Certain demographic variables require special consideration during data cleaning:

Age (Q46)

A common error when completing year of birth questions on forms is for respondents to accidentally write in the current year . thus responses to **Q46** of ±2014‡ will be set to missing during cleaning. Out-of-range responses will also be set to missing[5]. For the 2014 Mental Health survey, out-of range responses are defined as **Q46≤1996 OR Q46≥2014**.

---

[3] Respondents will *not* be considered ineligible for the survey solely on the basis of their response to the year of birth question indicating that they are under the age of eighteen. This is because of the difficulty of inferring the source of errors when year of birth from sample and response sections are mismatched . in other words, we cannot be certain whether this mismatch occurs *[see overleaf]* due to an error in the sample file, an error in the patient‡ completion of the questionnaire form, or an error in data entry.

[4] While a combination of sample and response information is used for demographic analysis on groups of cases, **only the sample information is used to calculate response rates by demographic groups**. Unlike subgroup analysis of variables which is only carried out for respondents, response rates are calculated using all patients in the sample. Because response rates vary between groups with differing demographic variables, using response and sample data to calculate response rates would create a systematic source of bias in that we are only able to amend information for the *respondents*. Therefore, only the sample information should be used to calculate response rates by demographic groups.

[5] The majority of out-of range responses present in data relating to year of birth questions result from errors in data entry (for example, not keying one of the digits . so ±1983‡ may become 983, 183, 193, or 198). In such cases it is important that the responses be checked against the completed questionnaire forms, and data corrected if necessary, **prior** to submission of data to the Co-ordination Centre.

3.5 Usability and eligibility

Sometimes questionnaires are returned with only a very small number of questions completed. For the Mental Health survey, questionnaires containing fewer than five responses are considered 'unusable'. we will delete all responses pertaining to such cases, and outcome codes of 1 ('returned useable questionnaire') relating to these cases will be changed to 6 ('questionnaire not returned'). The number of responses per questionnaire will be counted after all other cleaning, and that all responses remaining at this stage will be counted (including responses to the demographic questions)[6]. This should affect only a very limited number of cases, and so should not have a significant impact on response rates.

Outcome codes for respondents will also be changed if the respondents are believed to be under the age of eighteen and therefore ineligible for participation. Since the sample files for the survey are checked by the Co-ordination Centre prior to mailing, this is unlikely to affect more than a handful of cases throughout the survey, as people coded as being aged under 18 will be identified and removed from the sample before the start of the survey.   Sample members will not, however, be removed from the sample if data on their year of birth is missing from the sampling frame. If sample information on a respondent's year of birth is missing, though, and their response to **Q46** indicates that they are under 18 (specifically, if **Q46>1996**) then the outcome code for that person should be recoded from 1 ('returned completed questionnaire') to 5 ('ineligible for participation in the survey'). This should only be done where sample information is missing. If sample information indicates a patient is aged 18 or over, but this is contradicted by the rresponse, then the respondents survey outcome should remain as 1. This is to avoid removing legitimate responses because of an overly conservative approach to assessing eligibility; in other words, where the respondents age is uncertain (because sample and response information contradict each other, and in different instances either of these may be accurate or inaccurate) the benefit of the doubt is given in any assessment of eligibility.

3.6 Missing responses

It is useful to be able to see the numbers of respondents who have missed each question for whatever reason. Responses are considered to be missing when a respondent is expected to answer a question but no response is present. For non-filtered questions, responses are expected from all respondents . thus any instance of missing data constitutes a missing response. For filtered questions, only respondents who have answered a previous routing question instructing them to go on to that filtered question or set of filtered questions are expected to give answers. Where respondents to the survey have missed a routing question, they are not expected to answer subsequent 'filtered' questions; thus only where respondents were explicitly instructed to answer filtered questions should such blank cells be coded as missing responses.

The Co-ordination Centre codes missing responses in the data with the value 999[7]. For results to be consistent with those produced by the Co-ordination Centre, missing responses should be presented but should not be included in the base number of respondents for percentages.

---

[6] The multiple choice question Q9 is only counted once. So if, for example, Q9_1 and Q9_4 are ticked, this would count as only one response for the purpose of determining if a questionnaire is usable.

[7] This is an arbitrary value chosen because it is 'out-of-range' for all other questions on the survey.

3.7 Non-specific responses

As well as excluding missing responses from results, the Co-ordination Centre also removes non-specific responses from base numbers for percentages. The rationale for this is to facilitate easy comparison between institutions by presenting only results from those respondents who felt able to give an evaluative response to questions. For a full listing of ±non-specificqresponses in the 2014 Community Mental Health survey, please see Appendix B: Non-specific responses.

4 Appendix A: Example of cleaning

4.1 Incorrectly followed routing

Figure 2: Example 'raw'/'uncleaned' data

Record	Outcome	Q15	Q16	Q17
Patient Record Number	Outcome of sending questionnaire (N)	In the last 12 months have you had a formal meeting with someone from NHS mental health services to discuss how your care is working?	Were you involved as much as you wanted to be in discussing how your care is working?	Did you feel that decisions were made together by you and the person you saw during this discussion?
001	6			
002	1	2	1	2
003	1	1	1	1
004	4			
005	1	3	3	3
006	6			
007	1	1	2	2
008	1	1	3	2

Figure 2 shows hypothetical raw/uncleaned data for eight sample members, five of whom have responded to the survey. It can be seen from this data that some of the respondents have followed filter instructions from routing questions incorrectly:

> Respondent '002' has reported that they have not had a care review meeting in the last 12 months (**Q15=2**) and Respondent '005' has reported that they cannot remember if they have had a care review meeting (**Q15=3**), but have both responded to filtered questions (**Q16 and Q17**).

Following the cleaning instructions above will remove these inappropriate responses. Firstly, the filter instructions specify that:

if	**Q15**	=	2 or 3	*then delete responses to:*	**Q16-Q17**

In accordance with this, all responses for **Q16** and **Q17** must be removed in cases where the respondent has ticked **Q15=2 or 3** (i.e. not had a formal review in the last 12 months). This will lead to two responses being removed for both respondent '002' and '005' (**Q16 and Q17**).

Figure 3 (below) shows how the data would look following cleaning by the co-ordination centre to remove responses to filtered questions that should have been skipped . cells where responses have been removed are shaded.

Figure 3: Data from Figure 3 following cleaning

Record	Outcome	Q15	Q16	Q17
Patient Record Number	Outcome of sending questionnaire (N)	In the last 12 months have you had a formal meeting with someone from NHS mental health services to discuss how your care is working?	Were you involved as much as you wanted to be in discussing how your care is working?	Did you feel that decisions were made together by you and the person you saw during this discussion?
001	6			
002	1	2	.	.
003	1	1	1	1
004	4			
005	1	3	.	.
006	6			
007	1	1	2	2
008	1	1	3	2

5 Appendix B: Non-specific responses

The following table lists all "non-specific responses" included in the 2014 Mental Health survey. Numbers in the final column indicate the response options that should be considered non-specific. Where the "non-specific responses" column contains only a dash, the relevant question has no such response options. The data presented in the table below is for all questions, whether they are scored or not. The non specific responses given below are defined for all survey outputs (trust level data and national level reporting).

No.	Question	Non-specific responses
Q1	When was the last time you saw someone from the NHS mental health services?	6,7
Q2	Overall, how long have you been in contact with NHS mental health services?	6
Q3	In the last 12 months, do you feel you have seen NHS mental health services **often enough** for your needs?	5
Q4	How easy was the actual **journey** to see this person or people?	5
Q5	Did the person or people you saw **listen carefully** to you?	4
Q6	Were you given **enough time** to discuss your needs and treatment?	4
Q7	Did the person or people you saw **understand** how your mental health needs affect **other areas of your life**?	4
Q8	Have you been told **who is in charge** of organising your care and services? (This person may be anyone providing your care, and may be called a "care coordinator" or "lead professional")	3
Q9	Is the person in charge of organising your care and services… .	8
Q10	Do you know how to contact this person if you have a concern about your care?	3
Q11	How well does this person organise the care and services you need?	-
Q12	Have you agreed with someone from **NHS mental health services** what care you will receive?	-
Q13	Were you involved as much as you wanted to be in agreeing what care you will receive?	4,5
Q14	Does this agreement on what care you will receive take your personal circumstances into account?	4
Q15	In the last 12 months have you had a formal meeting with someone from NHS mental health services to discuss how your care is working?	3
Q16	Were you involved as much as you wanted to be in discussing how your care is working?	4,5
Q17	Did you feel that decisions were made **together** by you and the person you saw during this discussion?	4,5
Q18	**In the last 12 months**, have the people you see for your care or services changed?	5
Q19	What impact has this had on the **care** you receive?	4
Q20	Did you know who was in charge of organising your care while this change was taking place?	3
Q21	Do you know who to contact out of office hours if you have a crisis?	3
Q22	**In the last 12 months,** have you tried to contact this person or team because your condition was getting worse?	3
Q23	When you tried to contact them, did you get the help you needed?	-
Q24	In the last 12 months, have you been receiving any **medicines** for your mental health needs?	-

No.	Question	Non-specific responses
Q25	Were you **involved** as much as you wanted to be in decisions about which medicines you receive?	4,5
Q26	In the last 12 months, have you been prescribed any **new medicines** for your mental health needs?	-
Q27	The last time you had a new medicine prescribed for your mental health needs, were you given **information** about it in a way that you were able to understand?	-
Q28	Have you been receiving any medicines for your mental health needs for 12 months or longer?	3
Q29	In the last 12 months, has an NHS mental health worker checked with you about how you are getting on with your medicines? (That is, have your medicines been reviewed?)	3
Q30	In the last 12 months, have you received any **treatments or therapies** for your mental health needs that do not involve medicines?	4,5
Q31	Were you involved as much as you wanted to be in deciding what treatments or therapies to use?	4,5
Q32	In the last 12 months, did NHS mental health services give you any **help or advice** with **finding support** for **physical health needs** (this might be an injury, a disability, or a condition such as diabetes, epilepsy, etc)?	4,5,6
Q33	In the last 12 months, did NHS mental health services give you any **help or advice** with **finding support** for **financial advice or benefits**?	4,5
Q34	In the last 12 months, did NHS mental health services give you any **help or advice** with **finding support** for **finding** or **keeping work**?	4,5,6
Q35	In the last 12 months, did NHS mental health services give you any **help or advice** with **finding support** for **finding or keeping accomodation**?	4,5
Q36	Has someone from NHS mental health services supported you in taking part in an activity locally?	4
Q37	Have NHS mental health services involved a member of your family or someone else close to you as much as you would like?	5,6,7
Q38	Have you been given **information** by NHS mental health services about getting support from people who have experience of the same mental health needs as you?	4
Q39	Do the people you see through NHS mental health services understand **what is important to you** in your life?	-
Q40	Do the people you see through NHS mental health services **help you** with what is important to you?	-
Q41	I Do the people you see through NHS mental health services help you **feel hopeful** about the things that are important to you?	-
Q42	Overallō .	-
Q43	Overall in the last 12 months, did you feel that you were treated with **respect and dignity** by NHS mental health services?	-
Q44	Who was the main person or people that filled in this questionnaire?	-
Q45	Are you male or female?	-
Q46	What was your year of birth?	-
Q47	What is your religion?	-
Q48	Which of the following best describes how you think of yourself?	-
Q49	What is your ethnic group? (Tick ONE only)	-