NHS Adult Inpatient Survey 2020

Data cleaning instructions

Coordination Centre for Mixed Methods

Care Quality Commission

Ipsos MORI

1 Data cleaning – an overview

At the end of fieldwork, data needs to be submitted to the Coordination Centre for Mixed Methods (CCMM) in a raw or uncleaned format (for details of this, see the guidance on the NHS surveys website on Entering and Submitting Final Data: http://nhssurveys.org/survey-instructions/entering-and-submitting-final-data/). To ensure that the cleaning process is comparable across all NHS trusts, the CCMM collates and cleans the full dataset of all trusts.

This document provides a description of the cleaning processes that will be followed by the CCMM to clean and standardise the data for the 2020 Adult Inpatient Survey, to allow data users to replicate the cleaning process on raw uncleaned data, and to understand the cleaning processes taken. These instructions focus on the selected answer codes, rather than the free text comments, which are reviewed separately to ensure confidentiality and identify safeguarding concerns.

All data submitted to the CCMM at the end of the survey must be uncleaned.

Definitions

Below is a list of definitions of terms commonly used in this document, as they apply to the 2020 Adult Inpatient Survey:

Raw / uncleaned data: Raw or uncleaned data has been entered from returned questionnaires following the guidance on the NHS surveys website on Entering and Submitting Final Data: http://nhssurveys.org/survey-instructions/entering-and-submitting-final-data/.

Data cleaning: This refers to all editing processes applied to the final collated dataset.

Routing questions: These are items in the questionnaire which instruct respondents to either continue on to the next question or to skip irrelevant questions depending on their response to the routing question. For the 2020 Adult Inpatient Survey, the routing questions in the questionnaire are Q1, Q6, Q12, Q30, Q43, and Q49.

Filtered questions: These are items on the questionnaire that are intended to only be answered respondents who have selected specific answer codes in relevant routing questions. For the 2020 Adult Inpatient Survey, the filtered questions in the questionnaire are Q2, Q7, Q13, Q31 – Q33, Q44, and Q50.

Non-filtered questions: These are items in the questionnaire which should be answered by all respondents, as they are not subject to routing questions. For the 2020 Adult Inpatient Survey, the nonfiltered questions are Q1, Q3 – Q6, Q8 – Q12, Q14 – Q30, Q34 – Q43, Q45 – Q49, and Q51 – Q57.

Sample data: Patient data that is provided by the trust as part of the sampling process. This includes: gender, year of birth, ethnicity, date of admission and discharge, length of stay, treatment function code, admission method code, CCG code, ICD-10 chapter code, NHS site code for both admission and discharge, and information on whether the patient was diagnosed with and/or treated for COVID-19, as it is recorded on the trust's system.

Response data: Data from the completed questionnaire which is provided from the patient. This includes answers to Q1 through Q57.

Out-of-range data: This refers to instances where responses to a question are not permissible. For example, if there are three answer codes for a question, a '4' would be considered "out-of-range". A full list of all "out-of-range" responses for the 2020 Adult Inpatient Survey are listed in Appendix B: Out-of-range data.

Outcome: An outcome code is given to each patient to indicate whether or not they responded to the survey and (where available) their reason for not taking part. This is used to calculate the adjusted response rate for the survey, meaning it is vital all patients are coded appropriately. The coding for outcome is as follows:

Outcome 1: Returned completed questionnaire
Outcome 2: Undelivered / moved house
Outcome 3: Deceased after the start of fieldwork
Outcome 4: Too ill / opt out
Outcome 5: Ineligible
Outcome 6: Unknown
Outcome 7: Deceased before the start of fieldwork

Non-specific responses: This term describes response options that do not provide evaluative information or indicate the question is not applicable to the respondent. Most commonly, these are responses such as "Don't know / can't remember" or "I did not have any questions". A full list of such responses for the 2020 Adult Inpatient Survey can be found in Appendix C: Non-specific responses.

Missing' responses: This term is used to describe data which are not stored as a valid response for a question or variable in a dataset. There can be a number of different types of missing data, with the most common being classed as 'user missing' data. Within the data cleaning process, a number of different missing response codes are used to identify how data for a particular respondent has been handled. These are arbitrary values which are not included in the analysis for question responses, but are used to monitor for questionnaire development. These codes are as follows:

999: this code is used when someone should have answered a question, but didn't or contradicts a response to a later question.

998: this code is used when someone answered a question but shouldn’t have. For example, filtered questions.

997: this code is used when someone incorrectly multi-codes a single code question or provides two incompatible responses to a multi-code question. It is also used if an out-of-range response has been provided for the year of birth question.

996: this code is used to suppress data at trust level when a question has fewer than 30 responses. These responses would also remain suppressed from the overall base at national level.

Approach and rationale

The aim of the cleaning approach is to ensure an optimal balance between data quality and completeness, while still ensuring measurement of participant error to feed into survey development. Where responses are known to be inappropriate or erroneous these are removed, but where possible, participant responses are edited as little as possible.

2 Steps for editing and cleaning the final data

Cleaning filtered questions

Where participants have answered questions they are instructed to skip, it is important to remove these inappropriate responses as they are not relevant to the participant. It is likely that they simply missed the routing instructions and thought they had to provide an answer to the following question. These responses should be recoded to 998 to indicate they were coded incorrectly, and allow for measurement of levels of missed routing to inform questionnaire development.

Where a routing question is missing or has been left blank, the respondent should not be cleaned from the filtered questions, as they may have simply missed or been confused by the routing question.

These instructions should be followed in the order shown in the table below.

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

Table 1. Appropriate cleaning for routing questions in the 2019 Adult Inpatient Survey

Routing question	If this answer code is selected at the routing question, following questions should be recoded if answered	Filtered questions to be recoded as 998 if answered
Q1	2	Q2
Q6	3 or 4	Q7
Q12	6 or 7	Q13
Q30	2	Q31 – Q33
Q43	4	Q44
Q49	16 or 17	Q50

Incompatible answer codes for multi-code questions and multi-coding single code questions

Where participants have answered two incompatible codes in a multi-code question, these should be removed, as it is not possible for both those answers to be correct. For example, at Q5 a participant cannot select that “Noise from staff” prevented them from sleeping at night, but also that “None of these” options prevented them from sleeping.

Table 2: List of multi-code questions and answer codes that can only be single-coded

Multi-code question	Answer codes that cannot be multi-coded
Q5	6
Q14	1 and 5
Q39	5 and 6
Q49	16 and 17
Q51	4

Similarly, where a participant has selected more than one answer code at a single code question, these answer codes are incompatible and need to be cleaned out.

Table 3: List of single-code questions that cannot be multi-coded

Single code questions that should not be multi-coded
Q1 – Q4A
Q6 – Q13
Q15 – Q38
Q40 – Q48
Q50
Q52 – Q57

For this reason, where participants have selected more than answer code at a single code question or selected incompatible answer codes at a multi-code question, these should be recoded as 997.

Age / Year of birth eligibility

When the sample is initially reviewed, checks are conducted to ensure everyone invited is over the age of 16. However, there may be cases where the response data indicates that the respondent is under the age of 16. These will be manually reviewed if the survey response indicates they are 15 years old. Otherwise, when this occurs, respondents will not be considered ineligible for the survey unless the sample year of birth is missing, as it is not possible to determine whether this is caused by an error in the completion of the questionnaire or an error in the sample file. For example, a common error when completing this question is to write the current year, rather than the year of birth, which would imply the participant is less than one year old. For this reason, unless the participant has no year of birth in the sample, participants will be considered eligible, even if they provide a response at Q52 that implies that they are under 16.

Table 4: Eligibility and outcome codes of patients based on sample and response data of age

Original outcome code	Sample data	Response data	Eligibility	Final outcome code
1	YoB ≤ 2004	Q52 > 2004	Eligible	1
1	YoB ≤ 2004	Q52 ≤ 2004	Eligible	1
1	YoB ≤ 2004	Q52 = missing	Eligible	1
1	YoB ≤ 2004	Q52 = out of range	Eligible	1
1	YoB = missing	Q52 > 2004	Ineligible	5
1	YoB = missing	Q52 ≤ 2004	Eligible	1

Out-of-range data

In general, questions should be set as out of range where a value is given for that question that does not correspond to the answers available. For example, if a question only has three response options, a value of 4 would be considered “out-of-range” and should be set to 997

As the survey includes a question where participants are asked to include their year of birth, there is more potential for error, and therefore cleaning needs to be tailored for this question. A common error when completing year of birth questions on forms is for respondents to accidentally write in the current year. In this case, the response to Q52 would be considered an out-of-range response and would therefore be set to missing. For the 2020 Adult Inpatients Survey, out-of-range responses for Q52 are defined as 997. This must only be done after the Age / Year of Birth eligibility cleaning, as described above, has taken place to determine if participants are eligible for the survey.

A full list of out-of-range responses for the 2020 Adult Inpatient Survey is listed in Appendix B: Out-of-range data.

Usability

Sometimes questionnaires are returned with only a very small number of questions completed. As in previous years, for the 2020 Adult Inpatient Survey, questionnaires containing responses to fewer than five questions are considered “unusable”. All

responses are therefore removed and these participants are updated for the purpose of response rates as not having completed the survey (their outcome is set to 2).

For the purposes of this cleaning, each multi-code question is considered one question, so even if a respondent has given multiple responses at a question, that would still only count as them having completed one question. The number of questions responded to is also counted after all cleaning has been conducted, to ensure questionnaires where respondents have given invalid responses to all their questions are also removed.

This should only affect a very limited number of cases and should not have a significant impact on response rates.

Duplicates

Where more than one response is received from the respondent, the data used are selected according to the case that is the most complete (i.e. with the fewest unanswered questions, treating each multi-code question as one question). If there is no difference in completeness, the data used are then selected according to a priority order with online data having precedence. If a duplicate of the same level of completeness within the same mode is identified, the earliest response will be selected.

Age / Year of birth analysis

In a small number of cases, participants may give a different age or gender than is provided in the sample. For example, sample data may identify an individual as being born in 1980 only for the patient to report being born in 1985.

For response rates, these need to be calculated on sample data only, to avoid introducing a bias between what would be able to be updated by participants, and what is left un-changed for non-respondents. Therefore, only sample data should be used to calculate response rates by demographics, or non-response weighting.

However, for analysis, where responses to demographic questions are present, it is assumed these are more likely to be accurate than sample data as respondents are considered best placed to know their own age.

Because questions about demographics tend to produce relatively high item non-response rates, it is not always appropriate to rely on response data alone for analysis. For demographic analysis on groups of cases by age, the CCMM use a combination of the information supplied in the sample data and response data. Where response data is provided, this is given priority, but where it is missing, the data from the sample is included instead. For a very small number of respondents demographic information may be missing in both the sample and response data - in these cases data would be left missing in the new variable.

Missing question responses

Where respondents have left questions blank that they should have answered, each question with no answer code for that participant is recoded as 999.

Question suppressions

Results at both a national and trust level are supressed for questions with fewer than 30 respondents, to avoid responses being identifiable and ensure minimum base sizes for comparability. Questions with fewer than 30 responses are coded as 996.

Non-specific responses

When calculating percentages, in addition to excluding missing responses, the CCMM removes “non-specific response” options from base numbers for percentages. This is to ensure the percentages only relate to those participants able to give an evaluative response to the question, or to participants to whom the question was relevant.

As shown in table 5, using hypothetical data, non-specific response options 4, 5 and 6 have been excluded from the base number when calculating percentages for Q4A. This is because those selecting answer options 4 and 5 said they either did not need to keep in touch with friends and family or that there were no restrictions, so the question on whether they were able to keep in touch despite restrictions was not relevant to them. Those selecting answer option 6 said they did not know or could not remember, so were not able to provide an evaluative response to the question. Therefore, any percentages used based on Adult Inpatient 2020 data would use the percentages in the column on the far right of table 5, excluding the non-specific response options.

Table 5: Example of how percentages are calculated excluding non-specific response options with hypothetical data

Q4A: There were restrictions on visitors in hospital during the coronavirus (COVID-19) pandemic. Were you able to keep in touch with your family and friends during your stay?				
Response options	**Original base numbers**	**Percentage <u>including</u> non-specific response options**	**Base numbers for percentages**	**Percentage <u>excluding</u> non-specific response options**
1. Yes, often	6,000	58.5%	6,000	60.0%
2. Sometimes	2,000	19.5%	2,000	20.0%
3. No, never	2,000	19.5%	2,000	20.0%
4. I did not need to	50	0.5%	-	-
5. There were no restrictions on visitors	75	0.7%	-	-

6. Don’t know / can’t remember	125	1.2%	-	-
Total base	10,250	-	10,000	-

For a full list of non-specific responses in the 2020 Adult Inpatient Survey, please see Appendix C: Non-specific responses.

3 Weighting

Weighting is used to ensure trusts are comparable with one another, standardising for demographic differences, and to take into account non-response, to ensure results are representative of the populations being measured.

National data weights

When calculating national data, weights are used to account for non-response, by weighting the completed population back to the original sample. To do this, the sample is split into strata by gender, age band and route of admission (elective or emergency) from the initial sample. A weight is then calculated to ensure each stratum is the same size in the completed responses as in the initial sample, for each trust. For example, if men aged 16-35 admitted in an emergency made up 10% of the initial sample in one trust, but only 5% of their responses, these respondents would be given a weight of 2, so this group would now be twice the size and make up 10% of responses. This weight is capped at 5 to ensure that no excessive weights are used.

An additional weight is also applied per question to make sure each trust has the same number of weighted respondents. This is done to ensure no trust is over- or under-represented in the national results.

To calculate the national data, the two weights are multiplied together and applied before the data is run, meaning each trust is an equal size and the results reflect the sampled population.

Trust data weights

To calculate trust scores, a weight is used to standardise the trusts by age band, gender and route of admission (elective or emergency). This is to ensure that trusts do not appear to be performing better or worse than one another, simply because they are serving a different population or providing more of a different type of care.

Unlike with the national data weights, the strata are calculated using a combination of sample data and questionnaire data. These strata are then calculated to match the overall population of responses to the survey at a national level, for every trust. Therefore, if 10% of respondents to the survey were women aged 66+ admitted for elective care, then each trust would be weighted to ensure 10% of their responses were from this group. This ensures every trust has a consistent population.

Other data weights

Separate weights are used for site level, medical and surgical analysis. These follow a similar process as the trust weight, but at site level or medical or surgical level within a trust, rather than at overall trust level.

4 Appendix A: Example of cleaning

Table 6 shows hypothetical raw / uncleaned data for five participants, three of whom have responded to the survey. Of the three participants, '0002' and '0005' have correctly followed the routing, but '0003' has answered Q2 when they should not have, as it was not asked of those who said they were admitted in an emergency.

Table 6: Hypothetical data showing correctly and incorrectly followed routing for Q1-Q3

Patient record number	**Outcome**	**Q1**	**Q2**	**Q3**
		Was your most recent overnight hospital stay planned in advance or an emergency?	How did you feel about the length of time you were on the waiting list before your admission to hospital?	How long do you feel you had to wait to get to a bed on a ward after you arrived at the hospital?
D0001…	6			
D0002…	1	1	2	3
D0003…	1	2	4	1
D0004…	4			
D0005…	1	2		4

The cleaning instructions (as shown in Table 7) specify that if response value 2 is selected at Q1, then responses to Q2 should be recoded as 998.

Table 7: Routing cleaning instructions for Q1 and Q2

Routing question	**Response values requiring cleaning**	**Filtered questions to be recoded as 998 if answered**
Q1	2	Q2

Table 8 below shows how the data at Table 6 would look, once it had been cleaned by the CCMM to update ‘0003’’s responses to follow the correct routing.

Table 8: Hypothetical data (as at Table 6) showing correctly cleaned responses for routing for Q1-Q3

Patient record number	**Outcome**	**Q1**	**Q2**	**Q3**
		Was your most recent overnight hospital stay planned in advance or an emergency?	How did you feel about the length of time you were on the waiting list before your admission to hospital?	How long do you feel you had to wait to get to a bed on a ward after you arrived at the hospital?
D0001…	6			
D0002…	1	1	2	3
D0003…	1	2	998	1
D0004…	4			
D0005…	1	2		4

5 Appendix B: Out-of-range data

Variable	Out-of-range data
Birth	≥2005
Gender	<0 3-8 ≥10
Ethnicity	Anything except A-H, J-N, P, R, S or Z
DayOfAdmission DayOfDischarge	≤0 ≥32
MonthOfAdmission MonthOfDischarge	≤0 ≥13
YearOfAdmission	≤2018 ≥2021
YearOfDischarge	≤2019 ≥2021
LengthOfStay	≤0
DayQRec	≤0 ≥32
MonthQRec	≤0 ≥5
YearQRec	≤2020 ≥2022
Q5_1, Q5_2, Q5_3, Q5_4, Q5_5, Q5_6, Q14_1, Q14_2, Q14_3, Q14_4, Q14_5, Q39_1, Q39_2, Q39_3, Q39_4, Q39_5, Q39_6, Q49_1, Q49_2, Q49_3, Q49_4, Q49_5, Q49_6, Q49_7, Q49_8, Q49_9, Q49_10, Q49_11, Q49_12, Q49_13, Q49_14, Q49_15, Q49_16, Q49_17, Q51_1, Q51_2, Q51_3, Q51_4	<0 ≥2
Q30	≤0 ≥3
Q1, Q4, Q16, Q17, Q19, Q20, Q21, Q37, Q38, Q41, Q45, Q47, Q50, Q54,	≤0 ≥4
Q2, Q6, Q9, Q11, Q13, Q22, Q25, Q26, Q29, Q36, Q40, Q42, Q44, Q48, Q53,	≤0 ≥5
Q3, Q7, Q8, Q10, Q15, Q18, Q24, Q27, Q28, Q34, Q43, Q56	≤0 ≥6

Variable	**Out-of-range data**
Q4A, Q5, Q23, Q31, Q32, Q33, Q35	≤0 ≥7
Q12	≤0 ≥8
Q55	≤0 ≥10
Q46	≤0 ≥12
Q57	≤0 ≥20
Q52	>2005

6 Appendix C: Non-specific responses

The following table lists every question included in the 2020 Adult Inpatient Survey which have a non-specific response. This includes scored and unscored questions and this is used across survey outputs covering national and trust level reporting.

Question number	Question wording	Non-specific response answer codes	Non-specific response answer wording
Q1	Was your most recent overnight hospital stay planned in advance or an emergency?	3	Don't know / can't remember
Q2	How did you feel about the length of time you were on the waiting list before your admission to hospital?	4	Don't know / can't remember
Q3	How long do you feel you had to wait to get to a bed on a ward after you arrived at the hospital?	5	Don't know / can't remember
Q4	Did you ever stay in a hospital room or ward for those with coronavirus (COVID-19) or suspected coronavirus?	3	Don't know
Q4A	There were restrictions on visitors in hospital during the coronavirus (COVID-19) pandemic. Were you able to keep in touch with your family and friends during your stay?	4	I did not need to
		5	There were no restrictions on visitors
		6	Don't know / can't remember
Q5	Were you ever prevented from sleeping at night by any of the following?	3	Noise from medical equipment
		5	Something else
Q6	Did you ever change wards during the night?	4	Don't know / can't remember

Question number	Question wording	Non-specific response answer codes	Non-specific response answer wording
Q7	Did the hospital staff explain the reasons for changing wards during the night in a way you could understand?	4	No, but I did not need an explanation
		5	Can't remember
Q8	How clean was the hospital room or ward that you were in?	5	Don't know / can't remember
Q9	Did you get enough help from staff to wash or keep yourself clean?	4	I did not need help
Q10	If you brought medication with you to hospital, were you able to take it when you needed to?	4	I had to stop taking my medication as part of my treatment
		5	I did not bring medication with me to hospital
Q11	Were you offered food that met any dietary requirements you had?	4	I did not have any dietary requirements
Q12	How would you rate the hospital food?	6	I was fed through tube feeding
		7	I did not have any hospital food
Q13	Did you get enough help from staff to eat your meals?	4	I did not need help to eat meals
Q14	During your time in hospital, did you get enough to drink?	4	No, for another reason
		5	I had a hydration drip
Q15	When you asked doctors questions, did you get answers you could understand?	4	I did not have any questions
		5	I did not feel able to ask questions
Q18	When you asked nurses questions, did you get answers you could understand?	4	I did not have any questions
		5	I did not feel able to ask questions
Q22	Thinking about your care and treatment, were you told something by a member of staff that was different to what you had been told by another member of staff?	4	Don't know / can't remember

Question number	Question wording	Non-specific response answer codes	Non-specific response answer wording
Q23	To what extent did staff looking after you involve you in decisions about your care and treatment?	5	I was not able to be involved
		6	I didn't want to be involved
Q24	How much information about your condition or treatment was given to you?	5	Don't know / can't remember
Q25	Did you feel able to talk to members of hospital staff about your worries and fears?	4	I had no worries or fears
Q26	Were you able to discuss your condition or treatment with hospital staff without being overheard?	4	I did not want this
Q27	Were you given enough privacy when being examined or treated?	4	I did not want this
		5	Don't know / can't remember
Q28	Do you think the hospital staff did everything they could to help control your pain?	4	I was not in any pain
		5	Don't know / can't remember
Q29	Were you able to get a member of staff to help you when you needed attention?	4	I did not need attention
Q31	Beforehand, how well did hospital staff answer your questions about the operations or procedures?	5	I did not have any questions
		6	Don't know / can't remember
Q32	Beforehand, how well did hospital staff explain how you might feel after you had the operations or procedures?	6	Don't know / can't remember
Q33	After the operations or procedures, how well did hospital staff explain how the operation or procedure had gone?	6	Don't know / can't remember

Question number	Question wording	Non-specific response answer codes	Non-specific response answer wording
Q34	To what extent did staff involve you in decisions about you leaving hospital?	5	I did not want to be involved in decisions
Q35	To what extent did hospital staff take your family or home situation into account when planning for you to leave hospital?	5	It was not necessary
		6	Don't know / can't remember
Q36	Did hospital staff discuss with you whether you would need any additional equipment in your home, or any changes to your home, after leaving the hospital?	3	No, it was not necessary to discuss it
		4	Don't know / can't remember
Q38	Before you left hospital, were you given any written information about what you should or should not do after leaving hospital?	3	Don't know / can't remember
Q39	Thinking about any medicine you were to take at home, were you given any of the following?	6	I had no medicine
Q40	Before you left hospital, did you know what would happen next with your care?	4	I did not need further care
Q41	Did hospital staff tell you who to contact if you were worried about your condition or treatment after you left hospital?	3	Don't know / can't remember
Q42	Did hospital staff discuss with you whether you may need any further health or social care services after leaving hospital?	3	No, it was not necessary to discuss it
		4	Don't know / can't remember
Q44	After leaving hospital, did you get enough support from health or social care services to help you recover or manage your condition?	4	I did not need any support
Q47	During your hospital stay, were you ever asked to give your views on the quality of your care?	3	Don't know / can't remember

Ipsos MORI