

# DATA CLEANING GUIDANCE

# 2024 ADULT INPATIENT SURVEY

 Last updated: February 2025

Contacts

The Survey Coordination Centre for the NHS Patient Survey Programme

Picker Institute Europe

Suite 6,

Fountain House,

1200 Parkway Court,

John Smith Drive,

Oxford,

OX4 2JY

Tel: 01865 208127

E-mail: inpatient@surveycoordination.com

Website: [www.nhssurveys.org](http://www.nhssurveys.org)

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 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 [NHS Surveys website](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/).

Questions and comments

If you have any questions or concerns regarding this document, please contact the Survey Coordination Centre (SCC) using the details provided at the top of this page.



**For trusts and contractors taking part in the survey:**

Contractors submitting final data for the Adult Inpatient Survey **must not** clean their data before submitting it to the Survey Coordination Centre. Please refer to [Survey Handbook](https://nhssurveys.org/wp-content/surveys/02-adults-inpatients/03-instructions-guidance/2024/Survey%20handbook.docx) and [Entering and Submitting Final Data](http://www.nhssurveys.org/Filestore/Generic_instructions/Generic_Entering_submitting_data_V2.pdf) instructions for more details.

##

## Contents

[Data cleaning – an overview 5](#_Toc160704262)

[Introduction 5](#_Toc160704263)

[Definitions 5](#_Toc160704264)

[Raw / uncleaned data: 5](#_Toc160704265)

[Free text comments: 5](#_Toc160704266)

[Data cleaning:](#_Data_cleaning:) 6

[Ask-all questions: 6](#_Toc160704268)

[Routing questions: 6](#_Toc160704269)

[Filtered questions: 6](#_Toc160704270)

[Multiple response questions: 6](#_Toc160704271)

[Multiple questionnaire responses: 6](#_Toc160704272)

[Sample data:](#_Sample_data:) 7

[Response data: 7](#_Toc160704274)

[Out-of-range data: 7](#_Toc160704275)

[Outcome: 7](#_Toc160704276)

[Non-specific response: 7](#_Toc160704277)

[Missing responses:](#_Missing_responses:) 8

[Entering and coding data prior to submission](#_Entering_and_coding) 9

[Multiple Questionnaire Responses - De-duplication and Inclusion](#_Multiple_Questionnaire_Responses) 10

[Outcome code priorities](#_Outcome_code_priorities) 10

[Editing and cleaning data after submission](#_Editing_and_cleaning) 11

[Approach and rationale](#_Approach_and_rationale) 11

[Dealing with filtered questions](#_Dealing_with_filtered) 11

[Example 1](#_Example_1:) 12

[Example 2](#_Example_2:) 12

[Cleaning multi-code questions 1](#_Toc160704287)3

[Eligibility](#_Cleaning_special_cases) 13

[Age / Year of birth](#_Age_/_Year) 13

[Demographics 1](#_Toc160704290)4

[Out-of-range data 1](#_Toc160704291)5

[Usability 1](#_Toc160704292)5

[Example 3 15](#_Toc160704293)

[Conditions for usability](#_Conditions_for_usability)……………………………………………………………………………………..15

[Partial responses](#_Partial_responses)…………………………………………………………………………………………...16

[Missing question responses 1](#_Toc160704294)6

[Question suppressions](#_Question_suppressions)…………………………………………………………………………………….17

[Non-specific responses 1](#_Toc160704295)7

[Weighting](#_Weighting) 19

[Population weight](#_Population_weight_(pop_weight)) 19

[Trust data weights](#_Trust_weights_(qx_tr_weights)) 20

[Other data weights](#_Public_service_agreement) 20

[Appendix A: Example of cleaning](#_Appendix_A:_Example) 21

[Appendix B: In-range data](#_Toc160704301) 23

# Data cleaning – an overview

## Introduction

Once fieldwork for the 2024 Adult Inpatient Survey has been completed, data needs to be submitted to the SCC in a **raw, uncleaned** format (for details of this, see the guidance on the NHS surveys website on [Entering and Submitting Final Data](https://nhssurveys.org/survey-instructions/entering-and-submitting-final-data/). To ensure that the cleaning process is comparable across all NHS trusts, the SCC cleans the full dataset of all trusts.

This document provides a description of the processes that will be used by the SCC to clean and standardise data submitted for the 2024 Adult Inpatient Survey. By following the guidance contained in this document it should be possible for all data users to replicate this cleaning process on raw uncleaned data. 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. This document should be used alongside the IP24 [data mapping document](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/) which provides further information on specific and non-specific responses.



Please note the only data cleaning to be undertaken on the data file before it is submitted to the SCC is the de-duplication of cases and prioritisation of outcome codes where multiple questionnaires have been returned for a respondent. No further data cleaning should be applied to the raw data before it has been submitted.

## Definitions

Definitions of terms commonly used in this document, as they apply to the 2024 Adult Inpatient Survey are as follows:

## Raw / uncleaned data:

‘Raw’ or ‘uncleaned’ data are data that have been entered verbatim from completed questionnaires without any editing taking place to remove contradictory or inappropriate responses; thus, all response boxes crossed on the questionnaire should be included in the data entry spreadsheet (see the [Entering and Submitting Final Data document](http://nhssurveys.org/survey-instructions/entering-and-submitting-final-data/)).

While data are required in a raw/uncleaned format, data should still be checked for errors resulting from problems with data entry and should have duplicates removed. Ensuring high data quality is paramount and errors resulting from data entry problems can and should be corrected, as detailed in the [Final Data Entry Checklist](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/).

## Free text comments:

These are verbatim comments provided by a patient in response to the three open questions at the end of the survey: “Was there anything particularly good about your hospital care?”; “Was there anything that could be improved?”; and “Any other comments?”. These responses should be included within the data entry spreadsheet. A patient may have only answered the free text comments and none of the quantitative questions. We would still want their free text comments to be provided to the SCC even though they have not answered the other questions in the survey.

## Data cleaning:

The SCC uses the term ‘data cleaning’ to refer to all editing processes undertaken upon survey data once the survey has been completed and the data has been entered and collated.

## Ask-all questions:

These are items in the questionnaire which are not subject to any filtering, and which should therefore be answered by all respondents. For the 2024 Adult Inpatient Survey, the ask-all questions are **Q1, Q5, Q8-Q9, Q11-Q30, Q31\_1–Q31\_5, Q32, Q35-Q39, Q41-Q50, and Q52-Q59.**

## Routing questions:

These are items in the questionnaire which instruct respondents either to continue to the next question or to skip irrelevant questions, depending on their response to the routing question. For the 2024 Adult Inpatient Survey, the routing questions in the questionnaire are **Q1, Q5-Q6, Q9, Q32, Q39, and Q50.**

## Filtered questions:

These are items in the questionnaire which are not intended to be answered by all respondents. Whether individual respondents are expected to answer filtered questions depends on their responses to preceding routing questions. For the 2024 Adult Inpatient Survey, the filtered questions in the questionnaire are **Q2-Q4, Q6-Q7, Q10, Q33-Q34, Q40, and Q51.**

## Multiple response questions:

These are items in the questionnaire where either multiple responses to a single item are permissible, or the question is treated this way for analysis purposes. For the 2024 Adult Inpatient Survey, the multiple response questions are **Q6, Q8, Q16, Q41, Q50, and Q52.**

## Multiple questionnaire responses:

For the 2024 Adult Inpatient Survey, respondents have the option to complete the survey either online or on paper. This may lead to a respondent completing both the online and paper questionnaire. Below is a description of how multiple questionnaire responses may occur:

All eligible respondents receive the first letter inviting them to complete the survey online by including a link and log in details, and a QR code. Those aged over 80 years old, also get a paper copy of the questionnaire. After 5 working days from the first mailing (day 7), those who haven’t completed the survey receive an SMS text reminder with a unique link to the online survey. Following this, a first postal reminder letter is sent on day 14 from the first mailing to those who haven’t responded. After 5 working days (day 21 from the first mailing), non-respondents receive a second SMS text message reminder. Finally, on day 35 from the first mailing, those who haven’t completed the survey receive a second postal reminder letter which also includes a paper questionnaire.

Though the reminders are only sent out to non-respondents, there is the potential for an individual to complete both the online and paper version. This situation may arise if there is an overlap between when the respondent completes the online survey and when the paper questionnaire arrives in the mail. A respondent having just completed the online survey may assume that their response wasn’t recorded if they receive the paper questionnaire. Then they may fill out and return the paper questionnaire as well.

To address such cases of multiple responses from the same individual, [procedures are in place](#_Multiple_Questionnaire_Responses) during the coding of data prior to submission to identify and remove duplicate entries.

## Sample data:

Patient data that is provided by the trust as part of the sampling process. This includes: gender, year of birth, ethnicity, mobile phone indicator, date of admission and discharge, length of stay, treatment function code, admission method code, ICD-10 chapter code, NHS site code for both admission and discharge, and a virtual ward indicator.

## Response data:

Data from the completed questionnaire which is provided from the patient. This includes answers to Q1 through to Q59.

## Out-of-range data:

This refers to instances where data within a variable has a value that is not permissible. For categorical data – most of the 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). Out-of-range responses entered into the dataset should not be automatically (e.g., algorithmically) removed prior to submitting the data to the SCC. A full list of valid responses for the 2024 Adult Inpatient Survey can be found in the [data mapping document](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/). In range **sample** data is listed in [Appendix B: In-range data](#AppB).

## Outcome:

An outcome code is given to each patient to indicate the end result of their participation in the survey. This is used when calculating the adjusted response rate for the survey and is therefore vital to ensure all patients are coded appropriately. The coding for outcome is as follows:

Outcome 1: Returned completed questionnaire (including accessible versions)

Outcome 2: Undelivered / moved house

Outcome 3: Deceased after fieldwork started

Outcome 4: Too ill / opt out

Outcome 5: Ineligible

Outcome 6: Unknown

Outcome 7: Deceased prior to fieldwork

## Non-specific response:

This refers to response options that essentially indicate the question is not directly applicable or relevant to the respondent. Most commonly, these are responses such as “Don’t know / can’t remember”. Likewise, responses that indicate the question is not applicable to the respondent are considered ‘non-specific’ – for example, responses such as “I do not need support for this”. Please note that non-specific responses are set to user missing in the final respondent-level dataset[[1]](#footnote-2). This does not delete the data in any way but alters how that data is used in analysis.

## 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, several different missing response codes are used to identify how data for a particular respondent has been handled.

These codes are as follows:

* 999 ‘missing’: this code is used for out-of-range question responses and when someone should have answered a question but did not. For example, ask-all questions or filtered questions where the respondent meets the filter criteria.
* 998 ‘inapplicable’: this code is used when someone answered a question but should not have. For example, filtered questions where the respondent does not meet the filter criteria.
* 997 ‘incompatible’: this code is used when someone provided 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.
* 98 ‘inconclusive’: This code is used for respondents who have circled more than one answer for the scale of 0 to 10 overall experience question (Q48), as it cannot be determined which answer they intended to select.

**Please note, contractors should submit raw, uncleaned data to the SCC, as per the ‘Entering and coding data prior to submission’ section below.**

# Entering and coding data prior to submission

For the 2024 survey, contractors are required to submit raw (‘uncleaned’) data to the SCC. For clarification, raw data is created as follows:

* All responses should be entered into the dataset, regardless of whether or not the respondent was meant to respond to the question (e.g., where patients answer questions that they have been directed to skip past, these responses should still be entered).
* Where a respondent has selected more than one response category on a question, this question should be left blank for that respondent in the data. The exceptions to this are for the ‘multiple response’ questions (e.g., Q6), where respondents may select more than one response option. These should be coded ‘0’ for response not selected and ‘1’ for response selected[[2]](#footnote-3).
* Where a respondent has crossed out a response, this should not be entered in the data (the response should be left blank). Where a respondent has crossed out a response and instead selected a second response option, the second choice should be entered into the data.
* Where a respondent has given their response inconsistently with the formatting of the questionnaire but where their intended response is nonetheless unambiguous upon inspection of the completed questionnaire, then the respondent’s intended response should be entered. For example, where a respondent has written their date of birth underneath the boxes at Q53 (“What was your year of birth?”), then their year of birth should be entered.
* For the year of birth question, unrealistic responses should still be entered except following the rule above. For example, if a respondent enters ‘2024’ 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.
* 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 patient’s intended response has not been captured. This includes ‘out-of-range’ responses, which must not be removed from the dataset. Responses in the dataset should only be changed before submission to the SCC where they are found to have been entered inconsistently with the respondent’s intended response.
* The data file should be de-duped. In practice, this means removing multiple questionnaire submissions, so the file only contains one record per patient. This may include applying the outcome code priority order detailed in the next section.
* Free text comments given in the final three questions in the survey (“Was there anything particularly good about your hospital care?”; “Was there anything that could be improved?”; and “Any other comments?”) should be submitted within the [data entry spreadsheet](https://nhssurveys.org/wp-content/surveys/02-adults-inpatients/03-instructions-guidance/2024/Data%20Entry%20Spreadsheet.xlsx). Free text comments should be cleaned of any symbols and/or non-readable characters that may result from the survey software before submission.

## Multiple Questionnaire Responses - De-duplication and Inclusion

This section outlines how to approach situations when a patient returns multiple questionnaires. The below table details how to approach different scenarios where this may occur.

Table 1. Selecting a questionnaire if multiple questionnaires are returned by a patient in the 2024 Adult Inpatient Survey

|  |  |
| --- | --- |
| **Scenario** | **Priority** |
| The total number of completed questions should be calculated, and the questionnaire with the highest number of completed questions should be selected.  | **First** |
| In the event that the total number of completed questions is equal on all questionnaires, the data used are selected according to a priority order, and the earliest questionnaire received (either online or paper) should be selected.  | **Second** |
| In the rare event that the total number of completed questions is equal on all questionnaires, and the questionnaires were received at the same time, priority will be given to the response completed online.  | **Third** |

## Outcome code priorities

There may be scenarios when a patient falls into multiple outcome codes, for example a patient returns a completed questionnaire, but they are later flagged as deceased during fieldwork. In such cases, the following priority list should be used to determine which outcome codes should be used:

Table 2. Selecting an outcome code if multiple questionnaires or outcome codes in the 2024 Adult Inpatient Survey

|  |  |
| --- | --- |
| **Outcome Code** | **Priority** |
| Outcome 1: Returned completed questionnaire | **1st** |
| Outcome 7: Deceased prior to fieldwork  | **2nd** |
| Outcome 3: Deceased during fieldwork  | **3rd** |
| Outcome 5: Ineligible | **4th** |
| Outcome 4: Opted out | **5th** |
| Outcome 2: Undelivered / moved house | **6th** |
| Outcome 6: Unknown | **7th** |

# Editing and cleaning data after submission

## Approach and rationale

The aim of the SCC in cleaning the collated final data 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 to do so in a relatively permissive way to enable as many responses as possible to contribute to the overall survey results.

## Dealing with filtered questions

Some of the questions included in the survey are only relevant to a subset of respondents, and in these cases filter instructions are included in the questionnaire to route respondents past questions that are not applicable to them.

It is necessary to clean the data to recode responses where filter instructions have been incorrectly followed. In such cases, participants’ responses to questions that were not relevant to them are recoded in the dataset. Responses are only recoded where respondents have answered filtered questions despite selecting an earlier response on a routing question instructing them to skip these questions.

In such cases, participants’ responses to questions that were not relevant to them are recoded to ‘998’ to indicate a non-applicable response.[[3]](#footnote-4) See table 3 for a list of all routing questions included in the 2024 Adult Inpatient Survey, the response values that require cleaning, and the appropriate filtered questions to recode as ‘998’.

Table 3. Appropriate cleaning for routing questions in the 2024 Adult Inpatient Survey

|  |  |
| --- | --- |
| **Condition for routing questions** | **Recoding for filtered questions (if answered)** |
| **Q1 options 2 or 3** | **Q2 - Q4 = 998** |
| **Q5 option 1** | **Q6-Q7 = 998** |
| **Q6 options 6 or 7** | **Q7 = 998** |
| **Q9 options 3 or 4** | **Q10 = 998** |
| **Q32 options 2 or 3** | **Q33-Q34 = 998** |
| **Q39 options 2 or 3** | **Q40 = 998** |
| **Q50 options 17 or 18** | **Q51 = 998** |

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

A worked example of the cleaning process for recoding non-applicable responses to filtered questions is included in [Appendix A: example of cleaning](#_Appendix_A:_Example).

Responses are only recoded as ‘998’ where respondents have answered filtered questions despite ticking an earlier response on a routing question instructing them to skip these questions.

## Example 1:



In the example above (example 1), the response to Q2 would be recoded to ‘998’ because according to their answer from Q1 (the routing question), they were supposed to skip Q2.

Responses to filtered questions are not removed where the response to the routing question is missing.

## Example 2:



In the example above (example 2), the response to Q7 would remain as code 4 and Q6 would be coded as missing (999).

## Cleaning multi-code questions – Incompatible answer codes

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 Q8, participants cannot select both "Noise from staff" and "I was not prevented from sleeping" as reasons for being prevented from sleeping at night.

Where participants have selected incompatible answer codes at a multi-code question, or have selected more than one answer code at a single code questions, these responses should be recoded as 997, as detailed in table 4.

Table 4. List of multi-code questions and answer codes that can only be single-coded

|  |  |
| --- | --- |
| **Condition for multi-code questions** | **Recode** |
| Q6 any option 1 to 5 = 1 AND Q6 any option 6 = 1 | Q6 = 997 |
| Q6 any option 1 to 5 = 1 AND Q6 any option 7 = 1 | Q6 = 997 |
| Q8 any option 1 to 7 = 1 AND Q8 option 8 = 1 | Q8 = 997 |
| Q16 option 1 = 1 AND Q16 any of option 2 3 4 5 = 1 | Q16 = 997 |
| Q16 option 5 = 1 AND Q16 any of option 2 3 4 = 1 | Q16 = 997 |
| Q41 any option 1 to 4 = 1 AND Q41 option 5 = 1 | Q41 = 997 |
| Q41 any option 1 to 4 = 1 AND Q41 option 6 = 1 | Q41 = 997 |
| Q41 option 5 = 1 AND Q41 option 6 = 1 | Q41 = 997 |
| Q50 any option 1 to 16 = 1 AND Q50 option 17 = 1 | Q50 = 997 |
| Q50 any option 1 to 16 = 1 AND Q50 option 18 = 1 | Q50 = 997 |
| Q50 option 17 = 1 AND Q50 option 18 = 1 | Q50 = 997 |
| Q52 any option 1 to 3 = 1 AND Q52 option 4 = 1 | Q52 = 997 |

## Eligibility

## Age / Year of birth

There may be instances where the sample and response data are mismatched, and the response data indicates that the respondent is under the age of 16. When this occurs, respondents will be considered eligible for the survey if their sample data is not missing and therefore remain as outcome 1. This approach aims to avoid removing legitimate responses because of an overly conservative approach to assessing eligibility; in other words, where the respondent’s 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, so respondents are not excluded based on potentially inaccurate data. We cannot be certain whether the mismatch occurs due to an error in the sample file or an error in the patient’s completion of the questionnaire. It is also possible that there has been an error in data entry.

In instances where the sample data is missing, the response data is the only proof of age available. If the response data indicates the respondent is under the age of 16, the respondent will be considered ineligible (outcome 5). See table 5.

Table 5. 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 ≤ 2008 | Q53 > 2008 | Eligible | 1 |
| 1 | YoB ≤ 2008 | Q53 ≤ 2008 | Eligible  | 1 |
| 1 | YoB ≤ 2008 | Q53 = missing | Eligible | 1 |
| 1 | YoB = missing | Q53 ≤ 2008 | Eligible | 1 |
| 1 | YoB = missing | Q53 > 2008 | Ineligible | 5 |
| 1 | YoB = missing | Q53 = missing | Ineligible | 5 |

##

## Demographics

In a small number of cases, sample data and response data does not correspond for age and gender. For example, the sample may identify a patient as male only for them to report being female, or the sample data may identify an individual as being born in 1980 only for the patient to report being born in 1985.

Where patient responses to demographic questions in the questionnaire are present, it is assumed these are more likely to be accurate than the sample data (since it is assumed that respondents are best placed to know their own gender, age and other characteristics). However, because questions about demographics tend to produce relatively high item non-response rates, it is not appropriate to rely on response data alone. In cases of mismatch where it is obvious that the respondent data contains a clear and unambiguous error (e.g. when the respondent enters the current year instead of their year of birth), then sample data is used.

For demographic analysis on groups of cases, it is therefore necessary to use some combination of the information supplied in the sample data and response data. To do this, we first copy all valid responses to survey demographic questions into a new variable. Where response data is missing, we then copy in the relevant sample data[[4]](#footnote-5) (note that for a very small number of respondents demographic information may be missing in both the sample and response data; in such cases data must necessarily be left missing in the new variable)[[5]](#footnote-6).

## Out-of-range data

Out-of-range data must be set for invalid responses to all other questions in the survey. The out-of-range responses will depend on the number of response options given for each question. For instance, all questions with three response options (e.g., Q1, Q18 or Q23) with response data of **≤ 0 or ≥ 4** would be set to missing and coded as ‘999’.

A common error when completing year of birth questions is for respondents to accidentally write in the current year. In this case, the response to **Q53** would be considered as an out-of-range response. For the 2024 Adult Inpatient Survey, out-of-range responses for **Q53** are recoded as ‘997’. The out-of-range responses for **Q53** are defined as **Q53 ≤ 1900 or Q53 ≥ 2009.** This must only be done after eligibility has been set as described in the earlier section titled ‘[Eligibility](#_Cleaning_special_cases)’.

A list of in-range responses for the 2024 Adult Inpatient Survey are listed in [Appendix B: In-range data.](#_Appendix_B:_In-range)

## Usability

Sometimes questionnaires are returned with only a very small number of questions completed. As in previous years and across the NHS Patient Survey Programme, questionnaires containing fewer than five responses are considered ‘unusable’ – we will set all responses pertaining to such cases as system missing ‘left blank’ and recode the outcome to 6. This should only affect a very limited number of cases and so should not have a significant impact on response rates. The number of responses per questionnaire (including responses to the demographic questions) will be counted after all cleaning has been conducted.

Additional clarification regarding what constitutes as five responses in determining if a questionnaire is usable (or not):

* Verbatim comments in the other comments free-text response box (the final three questions in the survey “Was there anything particularly good about your care?”; “Was there anything that could be improved?”; and “Any other comments?”) are not counted towards the five responses.
* Multiple choice questions are counted as one response, even where multiple response options are selected. For instance, Q50 would be counted as one response in the below scenario.

## Example 3



## Conditions for usability

It is possible that a questionnaire could be considered usable because there are five or more responses, despite having an outcome code of 2, 3, 4 or 6. In this case the outcome would be recoded to 1 to indicate a complete usable questionnaire.

## Partial responses

For the online component, responses are to be accepted and recorded if the respondent has reached the ‘other comments’ free text question, although they do not need to provide a response in the text box to this question. As long as the survey software indicates that the respondent has viewed this question, the survey will be deemed complete, meaning that clicking the submit button is not required for the survey to be considered complete.

This approach ensures that all responses where the respondent had progressed to this specified point will be captured and included in the dataset.

## Missing question responses

It is useful to be able to see the number of missing responses for each question. Responses are considered to be missing when a respondent is expected to answer a question, but no response is present.

For ask-all 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 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.

Missing responses in the data are coded as ‘999’[[6]](#footnote-7). For results to be consistent with those produced by the SCC, missing responses should be presented but should not be included in the base number of respondents for percentages.

## Question suppressions

The SCC will suppress results at both national and trust level for questions that have fewer than 30 respondents[[7]](#footnote-8). The suppression of 30 is achieved by weights at national level or base size at trust level. Note: non-specific responses are excluded from this count.

## Non-specific responses

As well as excluding missing responses from results, non-specific responses are removed from the 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.

A list of non-specific responses is detailed in the [data mapping document](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/).

As shown in table 6, using hypothetical data, non-specific response option 4 has been excluded from the base number when calculating percentages for question Q34. This is because those selecting answer option 4 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 2024 Adult Inpatient Survey data would use the percentages in the column on the far right of table 6, excluding the non-specific response options.

*Table 6. Example of how percentages are calculated excluding non-specific response options with hypothetical data*

|  |
| --- |
| **Q34: Were you given enough information about the care and treatment you would receive while on a virtual ward?** |
| **Response options** | **Original** **base** **numbers** | **Percentage including non-specific response options** | **Base numbers for percentages** | **Percentage excluding non-specific response options** |
| 1. Yes, completely
 | 6,000 | 59.5% | 6,000 | 60.0% |
| 1. Yes, to some extent
 | 2,000 | 19.5% | 2,000 | 20.0% |
| 1. No
 | 2,000 | 19.5% | 2,000 | 20.0% |
| 1. Don’t know / can’t remember
 | 250 | 2.4% | - | - |
| Total base | 10,250 | - | 10,000 | - |

# Weighting

## Weights used in National Patient Surveys

For some analysis purposes, case weights are applied to the data. The weighted result comprises the case-level data multiplied by the case-level weight, aggregated to the reporting level and then divided by the sum of case weights at the reporting level.

### Population weight (pop\_weight)

The purpose of this weight is to adjust for differential non-response within trusts among demographic groups (by route of admission, gender and age) to ensure that the data is representative in terms of the demographic make-up. A single population weight is computed for each respondent.

The pop\_weight is calculated for each case as follows:

1. Use sample data from age (i.e. 16-35; 36-50; 51-65; 66+), gender and route of admission to create sixteen groups;

|  |  |  |
| --- | --- | --- |
| Route of admission | Gender | Age group |
| Emergency | Male | 16-35 |
|  |  | 36-50 |
|  |  | 51-65 |
|  |  | 66+ |
|  | Female | 16-35 |
|  |  | 36-50 |
|  |  | 51-65 |
|  |  | 66+ |
| Planned | Male | 16-35 |
|  |  | 36-50 |
|  |  | 51-65 |
|  |  | 66+ |
|  | Female | 16-35 |
|  |  | 36-50 |
|  |  | 51-65 |
|  |  | 66+ |

2) Calculate the proportion of total cases within each trust who fit into each of the sixteen groups (population proportion)

3) Calculate the proportion of *respondent* cases (outcome 1) within each trust who fit into each of the sixteen groups (respondent population)

4) Create a pop\_weight for each case by dividing the population proportion (step 2) by the respondent proportion (step 3)

5) To contain the impact of extreme outliers we truncate weights to a maximum of 5, meaning any weights larger than 5 will be replaced by the value of 5, thus reducing the impact of any very large weights. Note that the outliers are checked and truncated *before* the weight is added to ensure the truncating focusses only on those outliers resulting from differential non-response.

Missing weights are set to one to avoid data loss.

The pop\_weight is applied to categorical data in national tables analysis (with the exception of demographic questions), when it is multiplied by the trust weight for the question.

### Trust weights (qx\_tr\_weights)

This weight has been used since the beginning of the survey programme. Its purpose is to adjust for the differing numbers of respondents between trusts for any one question so that each trust has equal influence on the national table figures.

One qx\_tr\_weight is calculated per trust per question (with the exception of demographic questions). For a given question and trust, the qx\_tr\_weight is the average number of specific responses across all trusts divided by the number of specific responses in that trust.

The weight is applied to categorical data in national tables analysis when it is multiplied by the pop\_weight.

### Public service agreement weight (psa\_weight)

This originates with the Department of Health public service agreement for which the national statistics indicators were developed. It has also been applied in trust benchmarking from the time this was undertaken by CQC and subsequently transferred to Picker. Its purpose is to standardise the scores between trusts to create a more level ‘playing field’ when comparing trusts.

One psa\_weight is calculated per respondent equal to the proportion in that person’s weighting group in the national achieved sample divided by the proportion of that person’s weighting group in the trust’s achieved sample.

Weights are based on route of admission, age group and gender for this survey and are capped to a maximum of 5.

The weight is applied to scored data in trust benchmarking. Trust weights are not applied for benchmarking but are implicit because trust scores are averaged to produce the national mean scores. psa\_weight is also used to weight the NHS National statistic output.

A psa\_weight will be produced for each level of benchmark data provided:

* psa\_weight (trust level)
* site\_psa\_weight (site level)
* trust\_med\_psa\_weight (trust level medical services)
* trust\_surg\_psa\_weight (trust level surgical services)
* site\_med\_psa\_weight (site level medical services)
* site\_surg\_psa\_weight (site level surgical services)

# Appendix A: Example of cleaning

Figure 1 shows hypothetical raw / uncleaned data for eight patients, five of whom have responded to the survey. It can be seen from this data that some of the respondents have followed instructions from routing questions incorrectly:

Respondents ‘003’ and ‘005’ have answered questions about their virtual ward experience (Q33 and Q34) even though they said they did not or did not know / could not remember being admitted to a virtual ward in Q32, meaning that they should have skipped Q33 and Q34, and gone straight to Q35. Respondent ‘008’ has followed the routing correctly.

Figure 1. Example of raw / uncleaned data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Record** | **Outcome** | **Q32** | **Q33** | **Q34** | **Q35** |
| Patient Record Number | Outcome of sending questionnaire (N)  | When leaving hospital, were you admitted onto a virtual ward, also known as hospital at home? | Before being admitted onto a virtual ward, did hospital staff give you information about the risks and benefits of continuing your treatment on a virtual ward? | Were you given enough information about the care and treatment you would receive while on a virtual ward? | To what extent did staff involve you in decisions about you leaving hospital? |
| 001 | 6 |  |  |  |  |
| 002 | 1 | 1 | 2 | 1 | 1 |
| 003 | 1 | 2 | 2 | 4 | 1 |
| 004 | 4 |  |  |  |  |
| 005 | 1 | 3 | 4 | 3 | 4 |
| 006 | 6 |  |  |  |  |
| 007 | 1 | 1 | 3 | 3 | 4 |
| 008 | 1 | 2 |  |  | 1 |

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

*Table 7: cleaning example*

|  |  |
| --- | --- |
| **Condition for routing question** | **Recoding for filtered questions (if answered)** |
| **Q32 options 2 or 3** | **Q33-Q34 = 998** |

In accordance with this, all responses for **Q33** and **Q34** must be recoded as ‘998’. In cases where the respondent has answered **Q32 = 2 or 3** (i.e., had not been admitted to a virtual ward).

Figure 2 below shows how the data would look after cleaning is done by the SCC to remove responses to filtered questions that should have been skipped.

Figure 2. Example of cleaned data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Record** | **Outcome** | **Q32** | **Q33** | **Q34** | **Q35** |
| Patient Record Number | Outcome of sending questionnaire (N)  | When leaving hospital, were you admitted onto a virtual ward, also known as hospital at home? | Before being admitted onto a virtual ward, did hospital staff give you information about the risks and benefits of continuing your treatment on a virtual ward? | Were you given enough information about the care and treatment you would receive while on a virtual ward? | To what extent did staff involve you in decisions about you leaving hospital? |
| 001 | 6 |  |  |  |  |
| 002 | 1 | 1 | 2 | 1 | 1 |
| 003 | 1 | 2 | 998 | 998 | 1 |
| 004 | 4 |  |  |  |  |
| 005 | 1 | 3 | 998 | 998 | 4 |
| 006 | 6 |  |  |  |  |
| 007 | 1 | 1 | 3 | 3 | 4 |
| 008 | 1 | 2 |  |  | 1 |

# Appendix B: In-range data

The [data mapping document](https://nhssurveys.org/surveys/survey/02-adults-inpatients/year/2024/) published for 2024 Adult Inpatient Survey indicates the in-range values for each survey question. Table 8 covers the in-range data for sample information, or any information completed during fieldwork.

Table 8. In-range sample data

|  |  |
| --- | --- |
| **Sample Variable** | **In-range data** |
| Mobile number indicator | 0, 1 |
| Year of Birth | ≥ 1900 or ≤ 2008 |
| Gender | 0, 1, 2, 9  |
| Ethnic category | Anything except I, O, Q, T-Y |
| Day of AdmissionDay of Discharge | ≥ 1≤ 31  |
| Month of AdmissionMonth of Discharge | ≥ 1≤ 12  |
| Year of Admission | 2023, 2024 |
| Year of Discharge | 2024 |
| Length of Stay | >0≤ 700  |
| Treatment Function Code (on discharge) | See list of Treatment Function codes by [clicking here](https://www.datadictionary.nhs.uk/attributes/treatment_function_code.html). |
| ICD-10 Chapter Code | I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XVII, XVIII, XIX, XX, XXI, XXII, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 |
| Admission method | See list of Admission method codes by [clicking here](https://www.datadictionary.nhs.uk/attributes/admission_method.html). |
| NHS Site code-Admitted | See list of NHS Site Codes by [clicking here](https://files.digital.nhs.uk/assets/ods/current/ets.zip). |
| NHS Site code-Discharged | See list of NHS Site Codes by [clicking here](https://files.digital.nhs.uk/assets/ods/current/ets.zip). |
| Virtual Ward Indicator | 0, 1 |
| Day Questionnaire Received | ≥1≤ 31 |
| Month Questionnaire Received  | ≥1≤5 |
| Year Questionnaire Received | 2025 |
| Outcome Code | ≥1≤ 7 |

1. Please note that this rule doesn’t apply to multiple response questions, where all responses are included in the final respondent level dataset. [↑](#footnote-ref-2)
2. In this case, the response should be left blank during data entry and would subsequently be coded 999. [↑](#footnote-ref-3)
3. Code ‘998’ is an arbitrary value chosen because it is out-of-range for all other questions on the survey. [↑](#footnote-ref-4)
4. The sample data have already been checked for eligibility upon submission. [↑](#footnote-ref-5)
5. The exception to this is when response rates are calculated. Because response rates vary between demographic groups, 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 data should be used to calculate response rates by demographic groups. [↑](#footnote-ref-6)
6. This is an arbitrary value chosen because it is ‘out-of-range’ for all other questions on the survey. [↑](#footnote-ref-7)
7. This does not include the demographic items included in the ‘About You’ section of the questionnaire. [↑](#footnote-ref-8)