Stats NZ
Tatauranga Aotearoa

Data sources and imputation for cigarette smoking behaviour in the 2023 Census

Te Kāwanatanga o Aotearoa
New Zealand Government

Citation
Stats NZ (2024). *Data sources and imputation for cigarette smoking behaviour in the 2023 Census*. Retrieved from www.stats.govt.nz.

ISBN 978-1-991307-14-9 (online)

Published in October 2024 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Contact
Stats NZ Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600

www.stats.govt.nz

Table of contents

List of tables and figures

List of tables

List of figures

Purpose and summary

Purpose

Data sources and imputation for cigarette smoking *behaviour in the 2023 Census* describes the data sources and methods used to produce cigarette smoking behaviour information in the 2023 Census, focusing on the use of alternative data sources to mitigate missing census responses.

It also discusses changes since the 2018 Census, the reasoning behind these changes and the effect on the quality of the variable.

Summary of key points

It is important that quality data about cigarette smoking behaviour can be produced from the census for the health sector to monitor changes in smoking prevalence in New Zealand. However, non-response in the 2023 Census meant that some data was missing. Non-response impacted some populations and regions more than others. Census responses for cigarette smoking behaviour were combined with historical census data and statistical imputation to provide information where it would otherwise have been missing due to non-response.

The methodology for determining cigarette smoking behaviour in the 2018 Census used historical census data for current and past smoking behaviour and did not use historical census data to inform statistical imputation. This likely resulted in an overestimated smoking rate and underestimated ex-smoker rate.

We made changes to the methodology for determining cigarette smoking behaviour in the 2023 Census. Where a record had a missing or residual value, the 2023 Census methodology:

- used historical census data for past smoking behaviour for those that have smoked as this is something that will not change over time – that is, someone that has smoked in the past will never count as 'Never smoked' in the future
- did not copy historical values for current smoking behaviour – that is, if someone was a regular smoker in 2018 or 2013, this was not assumed to still be the case in 2023
- used statistical imputation for all categories and included historical smoking information as a predictor variable. This accounted for changes since the last census and used relevant information available for imputation.

The updated methodology for the use of alternative data and imputation for cigarette smoking behaviour resulted in an increased overall quality rating for the concept from moderate in 2018 to high in 2023.

Cigarette smoking behaviour – 2023 Census: Information by concept has more information on the smoking variables and their quality rating in the 2023 Census.

Editing, data sources, and imputation in the 2023 Census provides more information on the process used for other census variables.

Background to the cigarette smoking behaviour concept

This section describes what cigarette smoking behaviour information is used for, how the cigarette smoking behaviour questions were asked, and how this relates to the output variable.

How cigarette smoking behaviour data is used

Cigarette smoking behaviour data is used by the health sector to monitor changes in smoking prevalence in New Zealand. This information enables health professionals to understand more about smokers, to better target at-risk groups in the community with future health education programmes, and to assess the success of ongoing health education programmes.

Smoking data from the census is also used as a general health indicator. It is often used as an indicator when measuring negative wellbeing and deprivation.

Cigarette smoking behaviour questions and output

Cigarette smoking behaviour was derived using two input variables from two questions on the census individual form:

- The regular smoker indicator captures if someone is currently a regular smoker (question shown in Figure 1).
- The ever smoked indicator captures if someone (that is not currently a regular smoker) has ever been a regular smoker (question shown in Figure 2).

Figure 1: Regular smoker indicator question on the 2023 Census English individual paper form

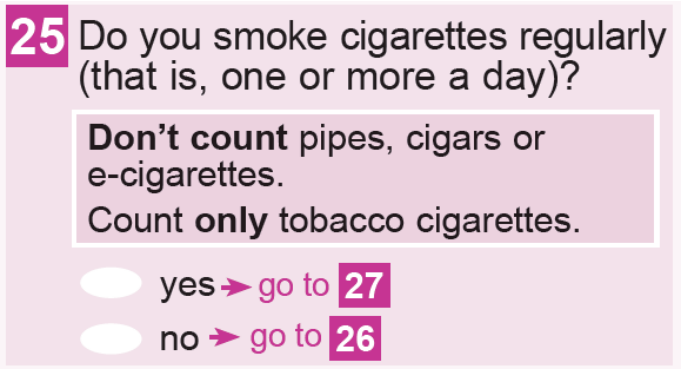
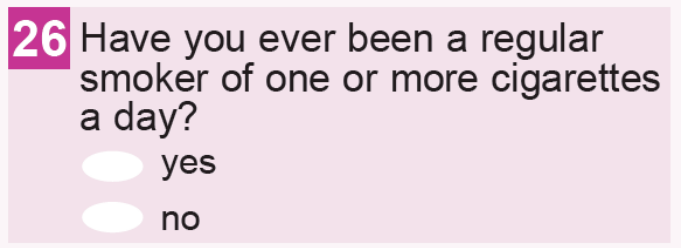


Figure 2: Ever smoked indicator question on the 2023 Census English individual paper form



A valid response for each question was a single 'yes' or 'no' response. For people that responded 'yes' to the regular smoker indicator question, no response to the ever smoked question was required. If a response was recorded it is disregarded, and they were considered a regular smoker.

The questions and routing for cigarette smoking behaviour in the 2023 Census were the same as in previous censuses.

On the online form, the questions and routing were the same as for the paper form shown in Figures 1 and 2. The ever smoked indicator question (Figure 2) was not presented to those that answered 'yes' to the regular smoker indicator question (Figure 1) on the online form.

Table 1

How responses to the input variables were used to derive the cigarette smoking behaviour output variable

Input variables		Cigarette smoking behaviour
Regular smoker indicator	Ever smoked indicator	
Yes	NA[1]	Regular smoker
No	Yes	Ex-smoker
No	No	Never smoked

1. This value was not considered when deriving cigarette smoking behaviour for people that had 'yes' for the regular smoker indicator. If a response was recorded, it was removed via an edit as this question should not have been answered.

The combined census model and missing information

This section describes the combined census model, alternative sources for cigarette smoking behaviour, and gives an overview of how missing information was handled for cigarette smoking behaviour.

The combined census model

The 2023 Census was a combined model by design. This means that alternative data sources were used to supplement census responses.

When an individual had not responded to the census and there was confidence in the quality of the admin record, then admin data was used to add records (referred to as admin enumerations). This achieved a higher, and more representative, coverage across subpopulations. This is important because non-response was not evenly distributed across the population – some subpopulations had higher rates of non-response than others.

Methodology for using admin data to count people in the 2023 Census provides more information about the admin enumeration process.

When there was missing information within a census record – for example, if an individual did not answer the date of birth question – there were processes to use historical census data, admin data, and statistical imputation to fill in the gaps. This was important to reduce biases in the data and under-representation of subgroups. Alternative data was not used if there was a valid census response.

Editing, data sources, and imputation in the 2023 Census provides more information.

Alternative sources for cigarette smoking behaviour

Historical census data was prioritised as the first potential alternative data source for cigarette smoking behaviour. In the 2018 Census, historical census data from 2013 was used. In the 2023 Census, historical census data from 2013 and 2018 was used.

Statistical imputation was only used when there was no census response or historical information available. Statistical imputation of the smoking variables in the 2018 and 2023 Censuses used a form of nearest neighbour imputation, which is where a missing value is imputed by copying information from a similar record. The imputation methodology is designed to maintain the population-level distribution while adjusting for the differences between the non-responding and responding populations. This is achieved by matching to donors using statistically relevant characteristics (such as age and location).

Editing, data sources, and imputation in the 2023 Census and Data sources, editing, and imputation in the 2018 Census have more details.

Missing information and derivation of cigarette smoking behaviour

Where there was a valid and complete census response to the cigarette smoking behaviour questions, this was used to derive cigarette smoking behaviour, and alternative data sources were not used.

However, cigarette smoking behaviour could not be derived if an individual had missing information for the regular smoker indicator, or if an individual that was not currently a smoker had missing information for the ever smoked indicator. In these cases, alternative data (historical census data and statistical imputation) was used to fill in missing information for each input variable separately.

Because the input variables were handled separately, this could mean that there is a combination of census response and alternative data used for the same individual. For example, if someone had responded ‘no’ to the regular smoker indicator but not answered the ever smoked indicator, only the ever smoked indicator would be alternatively sourced.

Cigarette smoking behaviour was derived after alternative data sourcing for the input variables.

Methods for using data sources for cigarette smoking behaviour

Overview of the 2018 method

In 2018, the following steps were used (in the order listed) to search for a valid combination of values for the input indicators for cigarette smoking behaviour for each record (including census respondents and admin enumerated records). Valid combinations of the input variables are outlined in Table 1. When a valid combination of values was obtained, the search was halted.

1. Regular smoker indicator:
    a. Start with the individual’s response to the regular smoker indicator in the 2018 Census.

   b. If no valid response, then take smoke regularly from the 2013 regular smoker indicator historical data.
   c. If still no valid value, then use statistical imputation.
2. Ever smoked indicator (if smoke regularly question resulted in 'no'):
   a. Start with the individual's response to the ever smoked indicator in the 2018 Census.
   b. If no valid response, then take ever smoked from the 2013 ever smoked indicator historical data.
   c. If still no valid value, then use statistical imputation.

Limitations of the 2018 Census methodology

Unlike attributes such as date of birth or birthplace, which should be stable over time, attributes that represent *current* information will always produce some inaccuracy when historical census data is used.

Because of the high levels of non-response in the 2018 Census, and the known differences between responding and non-responding populations, a methodology to fill gaps in the census data was required. The methodology developed within the required timeframes included the use of historical census data. This methodology carried the assumption that if someone was a regular smoker in 2013, this was still the case in 2018, despite decreasing smoking rates. It also assumed that if someone *wasn't* a regular smoker in 2013, this was still true in 2018 – this was more likely to be true given the reduction in people picking up smoking.

Concerns about the use of historical data were more pronounced in 2023 as some of the data was 10 years old. Also, according to the New Zealand Health Survey, the quit rate[1] for daily smokers was higher between 2018 and 2023 than between 2013 and 2018 (see quit rate indicator in the New Zealand Health Survey Annual Data Explorer). This increased the likelihood that people who were smoking at the time of the previous census would no longer be smoking in 2023, therefore using historical data would provide an inaccurate smoker and ex-smoker rate.

Changes in method from 2018 to 2023 Census

In the 2023 Census, the methodology for using alternative data for cigarette smoking behaviour had several key changes from the 2018 Census design:

1. Historical census data use was restricted to only filling in missing information for the ever smoked indicator (which captures past behaviour). It was not used for filling in missing information for the regular smoker indicator (which captures current behaviour).
2. Historical census data for filling missing information in the ever smoked indicator was restricted to when the historical data indicated that the individual *has* previously smoked. It was not used if there was no evidence of smoking in the historical data for the individual.
3. Historical census data for filling missing information in the ever smoked indicator was expanded to include information from historical regular smoker indicator values in addition to historical ever smoked indicator values.

[1] New Zealand Health Survey calculates the quit rate by dividing the number of people who have quit smoking in the past 12 months by the number of daily smokers who are still smoking daily plus the number of people who have quit smoking in the past 12 months.

To be considered someone who has 'quit smoking', an individual has to have smoked more than 100 cigarettes in their whole life and stopped smoking more than one month ago.

4. Statistical imputation was adjusted to also account for historical smoking information, where available, which reduced the impact of the non-smoker bias in the responding population.

Because historical census data on smoking behaviour is not necessarily reflective of current behaviour, the 2023 Census did not use historical data for the regular smoker indicator or for the 'Never smoked' category of the ever smoked indicator.

However, historical information could inform past behaviour (that is, the 'Have ever smoked' category for the ever smoked indicator). As such, historical census data was only used in the 2023 Census when it indicated that someone had previously smoked.

Overview of the 2023 method

We used the following steps (in the order listed) to search for a valid combination of values for the input indicators for cigarette smoking behaviour for each record (including census respondents and admin enumerated records) in the 2023 Census. Valid combinations of the input variables are outlined in Table 1. When a valid combination of values was obtained, the search was halted.

1. Start with the individual's responses to the regular smoker and ever smoked indicator questions in the 2023 Census.
2. For the ever smoked indicator only, if no valid response, then look for past smoking behaviour in:
    a. 2018 and 2013 ever smoked indicator ('yes' only)
    b. 2018 and 2013 regular smoker indicator ('yes' only).
3. For both indicators, if still no valid combinations of values, then use statistical imputation for missing information (with historical information accounted for where available).

Comparison between methodologies

Table 2 summarises the data sources used for the input variables in 2018 and 2023.

Table 2

Summary of data sources used for the input variables in 2018 and 2023

Data source	Regular smoker indicator		Ever smoked indicator	
	2018	2023	2018	2023
Census response	✓	✓	✓	✓
Historical census data	✓[1]	✗	✓[2]	✓[3]
Statistical imputation	✓	✓[4]	✓	✓[4]

1. From the regular smoker indicator.

2. From the ever smoked indicator.

3. From the regular smoker and ever smoked indicator where there is evidence that they have smoked.

4. Accounting for historical information in donor matching where available.

Because of the substantial reduction in historical data use in the 2023 Census, there was an increase in the number of records that required statistical imputation.

Non-smokers had a higher response rate and comprised the majority of the population, even for subpopulations with higher smoking rates. In most subpopulations, those that had never smoked comprised the majority of non-smokers. As outlined, the method of statistical imputation used in the census involves copying a census response from a donor form to fill in missing information for another individual (donee form). Because non-smokers had a higher response rate, this skewed the donor pool towards people that were not smokers and especially towards those that had never smoked.

The bias towards non-smokers in the responding population is accounted for by matching to donors with similar characteristics to the donee (non-responding individual), however the skew cannot be entirely removed. To improve the accuracy of matching to donors for imputation, historical smoking values were used in addition to the normal matching variables from the 2023 Census, where historical information was available.

Testing of different methodologies for statistical imputation demonstrated that when historical smoking behaviour was not accounted for, 'Never smoked' was over-imputed. By accounting for past smoking behaviour, donees who had previously smoked were more likely to be matched to donors that had also smoked in the past. This slightly increased the proportion that were imputed to regular smoker and ex-smoker, thus reducing the non-smoker bias in the donor pool.

Table 3 shows the distribution of the cigarette smoking behaviour categories using the 2018 methodology for 2018 data, the 2018 methodology for 2023 data, and the final distribution for 2023 using the new methodology for 2023 data.

Table 3

Distribution of cigarette smoking behaviour

2018 methodology and new 2023 methodology

2018 and 2023 Censuses data

Census year	Regular smoker	Ex-smoker	Never smoked regularly
2018	13.2%	22.0%	64.8%
2023 (2018 methodology)	8.9%	23.2%	68.0%
2023 (new 2023 methodology)[1]	7.7%	25.0%	67.4%

1. Note that this is the final version of the cigarette smoking behaviour output variable.

Note: Census data has had fixed random rounding applied to protect confidentiality. Individual figures may not sum to totals.

Source: 2023 and 2018 Censuses, Stats NZ

This table demonstrates that if the methodology from 2018 was used in 2023, the regular smoker rate would be higher and the ex-smoker rate would be lower. This was largely because of historical data use. When the 2018 methodology was used on 2023 Census data, one quarter of regular smokers were coded using historical data, in contrast only 9% of the overall population were coded using historical data. This suggested that historical data was overestimating the regular smoker population in 2023. This was supported by comparing the smoking status of those who responded to the 2023 Census with their smoking status in the historical data. This comparison highlighted that

over half of 2023 Census respondents that were regular smokers in 2018 or 2013 were not regular smokers in 2023.

It was expected that those requiring alternative data sourcing would have higher smoking rates due to the overlap between the non-responding population and the populations with higher smoking rates. However, such a marked difference, combined with the known changes in the smoking behaviours of the responding population, suggested that the historical data was not providing accurate information for current smoking behaviour. This supported the change in methodology for the 2023 Census.

Data sources for the output variable

Because the smoking input variables (regular smoker indicator and ever smoked indicator) are finalised independently before the derivation of the output variable (cigarette smoking behaviour), there could be a different data source for the two input variables (see Table 4 for distribution of this in the 2023 Census). For example, there may have been a census response of 'no' for the regular smoker indicator but no response for the ever smoked indicator, which was then sourced from historical data.

For regular smokers, the data source indicator value for cigarette smoking behaviour in the 2023 Census was taken from the regular smoker indicator. For not regular smokers, the data source was taken from the ever smoked indicator. The majority of records (96%) had the same data source for both input variables or were regular smokers and therefore the ever smoked indicator wasn't relevant. Most of the remaining records had statistical imputation for the regular smoker indicator and historical data for the ever smoked indicator. These records had the historical data source captured as the data source for cigarette smoking behaviour.

In 2018, data source indicators were only published for the two input variables. In 2023, a data source indicator for the output variable has been published and data source indicators for the input variables are also available.

Table 4

Distribution of data sources

Regular smoker data source by ever smoked data source

2023 Census

Regular smoker data source	Ever smoked data source				
	2023 Census Response	Historical census response	Statistical imputation	No information[1]	Total
2023 Census response	78.6%	0.1%	0.3%	6.1%	85.2%
Statistical imputation	0.3%	3.0%	10.0%	1.5%	14.8%
Total	78.9%	3.1%	10.3%	7.7%	100.0%

1. These people are regular smokers and as such do not have a value for the ever smoked indicator. Missing information is not an issue for these records.

Note: Census data has had fixed random rounding applied to protect confidentiality. Individual figures may not sum to totals.

Source: 2023 and 2018 Censuses, Stats NZ

Quality assessment

The changes to the methodology resulted in a more appropriate use of alternative data. It accounted for both changes in smoking behaviour over time and the differences in smoking behaviour between the responding and non-responding populations. This was reflected in an increase in the overall quality rating from moderate in 2018 to high in 2023. The overall quality rating is taken from the lowest rating across the three metrics assessed.

Table 5

Cigarette smoking behaviour quality ratings
2018 and 2023 Censuses

Census year	Overall rating	Metric one rating	Metric two rating	Metric three rating
2018 Census	Moderate	High	Moderate	Moderate
2023 Census	High	High	High	Very high

The metric one quality rating (data sources and coverage) remained the same (high). The metric one rating for the:

- regular smoker indicator decreased from high in the 2018 Census to moderate in the 2023 Census
- ever smoked indicator remained at high.

The metric one rating for the regular smoker indicator was affected more strongly by the methodology change as historical data sourcing was removed from use for this variable entirely. The final metric one score in 2023 was assessed using the combined output data source indicator.

Improvements in the census response rate from 2018 to 2023 and improvements to the quality of imputation were countered by the reduced use of historical census data and an increase in statistical imputation. Although statistical imputation is more appropriate to use than historical data for the regular smoker indicator and the 'Never smoked' category for the ever smoked indicator, it has a lower rating than historical data. As such, the increased use of statistical imputation limited the effects of other improvements on the metric one score.

The metric two quality rating (consistency and coherence) increased from moderate to high. The data reflected the expected trends (declining smoking rate), and this was consistent across subpopulations. The changes to alternative data use increased confidence in the alternatively sourced values, resulting in the increase to high.

The metric three quality rating (accuracy of responses) increased from moderate to very high due to improvements in scanning and repair for responses on paper forms, which reduced the number of responses needing to be sourced from alternative sources.

Data quality assurance in the 2023 Census has more information on data quality ratings in the 2023 Census.

Cigarette smoking behaviour – 2023 Census: Information by concept provides more information on the cigarette smoking behaviour concept.

Time series comparison

Data from the 2023 Census can still be compared with 2018, provided users are aware of the changes in methodology, as the overall declining trend in smoking is still present.

It is unclear how much of the change in regular smoking rates between 2018 and 2023 is due to the methodology change, compared with real-world changes in smoking behaviour.

Overall assessment

Data on cigarette smoking behaviour from the census is important for decision making in policy and measuring health outcomes. However, use of historical census data to fill in missing information on current smoking behaviour was providing inaccurate smoking rates and did not reflect people quitting appropriately.

The change to only directly use historical census responses for past behaviour (the ever smoked indicator) and also include historical census responses from the regular smoker indicator for the current ever smoked indicator is a more appropriate use of historical census data. This has enabled a more accurate reflection of the declining smoking rate and reduced the rate of regular smokers coded from alternative data sources.

Statistical imputation comprised a much larger proportion of alternatively sourced values in the 2023 Census, compared with 2018, because of the reduced use of historical census data. While historical census responses were not directly used for the regular smoker indicator, or the 'Never smoked' category for the ever smoked indicator, they were used in statistical imputation as matching information. This improved imputation of smoking behaviour because information was imputed from records with the same smoking history (where historical information was available). Without accounting for historical smoking behaviour, someone who had previously smoked could have current smoking information imputed from someone who had never smoked. Although the donor record would be similar in important predictive characteristics, they were not comparable in smoking behaviour and therefore less accurate. This change reduced the effects of the non-smoker bias in the responding population and improved the quality of imputed values.

The smoking data from the 2018 Census can still be used in time series comparisons as it reflects the overall declining smoking rates in New Zealand. The change in methodology for alternative data use for cigarette smoking behaviour is a more appropriate and accurate use of historical data and statistical imputation. This has improved the overall data quality for this concept.

Further reading

Ministry of Health (2023). New Zealand Health Survey Annual Data Explorer. Retrieved from minhealthnz.shinyapps.io.

Stats NZ (2024a). Cigarette smoking behaviour – 2023 Census: Information by concept. Retrieved from datainfoplus.govt.nz.

Stats NZ (2024b). Editing, data sources, and imputation in the 2023 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2024c). Methodology for using admin data to count people in the 2023 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2024d). Data sources, editing, and imputation in the 2018 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2024e). Data quality assurance in the 2023 Census. Retrieved from www.stats.govt.nz.