Stats NZ
Tatauranga Aotearoa

Methodologies for filling gaps in gender and sex at birth concepts for the 2023 Census

2023 Census | Tatauranga 2023

January 2024

Te Kāwanatanga o Aotearoa
New Zealand Government

Citation
Stats NZ (2024). *Methodologies for filling gaps in gender and sex at birth concepts for the 2023 Census*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-104965-0 (online)

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

Updated in December 2024

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

Contents

List of tables and figures

List of tables

List of figures

December 2024: We updated table A1 in Appendix 1 to describe how DIA now allows a person to select 'non-binary' as a marker on their New Zealand birth certificate (via application).

Purpose

Methodologies for filling gaps in gender and sex at birth concepts for the 2023 Census provides information on the methods that will be used to fill gaps in the 'gender' and 'sex at birth' variables in the 2023 Census.

This paper also:

- provides background, context, and information on gender and sex at birth concepts
- summarises the engagement and consultation on these concepts that have informed decisions about the methodologies employed in the 2023 Census
- covers uncertainties, risks, and limitations, and has an outlook to future work required to ensure that Stats NZ produces outputs related to the concepts of gender and sex at birth that meet stakeholders' needs.

Gaps in the 'sexual identity' and 'variations of sex characteristics' variables will not be filled by alternative data sources. Therefore, these variables are outside the scope of this paper.

Summary

Stats NZ will use the following methodologies to fill gaps (to address missingness) in the gender and sex at birth variables for the 2023 Census.

Gender: If a census response does not include a usable value for this variable from the 2023 Census individual form, then the gender variable will, in order, be sourced from:

- the 2023 Census dwelling or household summary form
- from admin data sources, where the admin data and Stats NZ concepts align
- statistical imputation.

Sex at birth: If a census response does not include a usable value for this variable from the 2023 Census individual form, then the sex at birth variable will, in order, be sourced from:

- admin data sources where the admin data and Stats NZ concepts align
- historical census responses for sex

- wider admin data sources with different concepts of sex or gender
- statistical imputation.

While the statistical imputation of one variable will not be dependent on the other, gender and sex at birth are imputed simultaneously. If both need imputing, they will be imputed from the same record.

Decisions about the methods employed for gender and sex at birth variables have been informed by:

- internal Stats NZ technical review
- engagement with key stakeholders and experts in data about Rainbow communities
- a public consultation process[1]
- discussion with the Data Ethics Advisory Group.

We recognise that many people in Rainbow communities have experiences where sex, gender, or other key elements of their identity are denied or must be carefully guarded. For these reasons, input was sought from various groups and experts working with rainbow data. Additionally, there were opportunities for members of the wider community to provide their thoughts and feedback on the approaches proposed by Stats NZ.

This feedback will also support the appropriate presentation of the outputs from the 2023 Census and clear explanation of the variables for data users.

[1] Gender and sex concepts 2023 Census methodology consultation provides more information.

Background

Gender and sex at birth concepts

In addition to providing population and dwelling counts, the census of population and dwellings provides accurate aggregated information about the population on a variety of topics of interest, nationally and down to small geographical levels and sub-populations. This information is de-identified for analysis, and everything published from the data is confidentialised. It is about groups in the population, and no individuals or individual households can be identified in statistics or research that are published.[2]

Data on gender and sex at birth (see Glossary for explanation of terms) is used to understand the demographic characteristics of a population for statistical, policy, and research purposes. Given their importance, Stats NZ developed a data standard for both concepts, which was finalised in 2021. This involved public consultation, and input from data users, international experts, and an external advisory group of subject matter experts and community advocates. The data standard 'standardises' the definitions and measures of collection of gender and sex concepts:

- The gender concept used by Stats NZ has three response categories: 'male', 'female', and 'another gender'.
- The sex at birth concept used by Stats NZ has two response categories: 'male' and 'female'.

Data standard for gender, sex, and variations of sex characteristics provides more information on the definitions and measures, and for the collection and dissemination of sex and gender concepts and data.

For the 2023 Census:

- a question on gender has been explicitly included for the first time
- a question on sex at birth has been included, replacing the question that read 'Are you: Male/Female' in the 1996 to 2018 Censuses (this question was used for the sex concept but was not labelled as such on forms)
- in accordance with the data standard, gender is the default variable for demographic breakdowns (as opposed to sex at birth); this means in the

[2] Data confidentiality degrees of identification in data (PDF) provides more information about de-identified and confidentialised data.

2023 Census, it will be included in most data outputs that have demographic breakdowns

- in line with the data standard, sex at birth will not be used for default demographic breakdowns and will only be used in outputs where there is a specific need for this concept instead of or in addition to gender
- in line with the data standard, intersex population data is based on the variations of sex characteristics variable, not gender or sex at birth
- related concepts (called derived variables) include:
    - ‘cisgender and transgender status’ – the value for this variable is determined by responses to both the gender and sex at birth questions
    - ‘Rainbow LGBTIQ+ indicator’ – which is determined by responses to the gender, sexual identity, and variations of sex characteristics variables, and by the derived cisgender and transgender status.

One of the uses of census data is to create time series, which measure change in the population over time. Time series for the new gender and sex at birth variables will be available, using the historical sex variable for the 2018 and 2013 data. The gap-filling methodology presented in this paper is not expected to have a notable effect on the time series. There will be some discontinuity for these time series, as the data in 2023 is for different concepts to 2018 and 2013. Timeseries tables will have supporting information to reflect the conceptual changes across collection periods.

Completeness of variables

The best quality data in a census is achieved when people respond fully and accurately to questions asked. However, some people do not respond to the census at all, some do not respond to every question, and others do respond but provide unusable responses. All these situations lead to gaps or ‘missingness’ in the variables.

As census data is widely used and vital for public policy and investment decisions, it is important that the gender and sex at birth variables are the highest quality possible, meaning there should not be any gaps.

If a person had a missing response for either the gender or sex at birth variables, that person and all their information could not be taken into consideration during analyses that involve that variable, meaning the analyses themselves would be less accurate and less representative of the population distribution, particularly at low geographical levels or for smaller sub-populations. This ‘missingness’ can disadvantage communities

because local policies and decisions about services would be informed by less representative and less accurate information.

To best benefit census data users and the communities impacted by the policies and services informed by census data, the gender and sex at birth variables need to be 'complete', meaning all records have usable information for these variables.

Consistent with the design of the 2023 Census, the procedure for the 2023 Census is to fill gaps in variables where this can be done appropriately, applying what is called the 'combined census model'[3].

However, gender and sex at birth can be particularly personal topics, especially to many in Rainbow communities. Stats NZ has aimed to find an approach that minimises unintended consequences and maximises benefit to the communities of Aotearoa New Zealand by making sure the data that will be output from the census is the highest quality possible.

1. Gender is required to be a complete variable for the 2023 Census. This enables Stats NZ to provide the distribution and counts of gender at a national level as required. Therefore, where no response was provided in the census, one must be sourced from alternative data. This requirement enables key uses of census data, as missing values would exclude respective records from a variety of important analyses.
2. Sex at birth is required to be a complete variable for the 2023 Census, requiring sourcing from alternative data to fill gaps. Consultations and discussions revealed that this would support the quality of planned outputs.
3. The derived variables cisgender and transgender status and Rainbow LGBTIQ+ indicator will only be produced for records where individual census form responses provide the necessary information; they will not use alternatively sourced data or imputed data to achieve complete variables. For records where that is not possible, the value will show 'cisgender and transgender status unidentifiable' or 'LGBTIQ+ status unidentifiable', while the data source indicator will be coded to 'no

---

[3] Using a combined census model for the 2023 Census provides information on the planned use of alternative data sources to produce the best quality data possible in the 2023 Census.

Editing, data sourcing, and imputation: Planned approach for the 2023 Census provides more information, including historical use of imputation and admin data in past censuses.

information'. The data quality concerns that arise from using alternatively sourced or imputed data in these derived variables are explained further in the limitations section of the document.

For all variables output in the 2023 Census, Stats NZ will provide information on the breakdown of data sources used, and where customised data is requested, it will be possible to have data source flags and filter on the different data sources used for variables.

Circumstances of data collection

People could respond to the census either online or on paper. While people were asked to answer all questions, on paper respondents may either intentionally or unintentionally miss questions in their census forms. Online, people were required to provide a response to at least one of the gender or the sex at birth questions to progress further and submit the form.

Each person needed to complete an individual form (either on paper or online). Each household had to complete either a paper dwelling form, or if completing the census online, there were two forms – the household summary form and the dwelling form. The paper dwelling and online household summary forms also asked for each person's gender and age, but not for their sex at birth.

Using alternative data sources

The 2023 Census includes a combined census model by design. This involves using admin data, historical census data, and statistical imputation to fill gaps in variables where appropriate. Together these are called alternative data sources[4]. Definitions of terms used in this paper are provided in the Glossary.

Historical census data and admin data are usually the best proxy for a record's correct value in a variable, leading to higher quality data.

- Stats NZ is using data from the 2013 and 2018 Censuses in the 2023 Census.
- There are two key admin data sources for gender and sex at birth variables (these are discussed further in the next section – Methodologies):

[4] Editing, data sourcing, and imputation: Planned approach for the 2023 Census sets out Stats NZ's intended approach for filling gaps in census responses.

- Births data from the Department of Internal Affairs has been shown to be a very good proxy for respondents' sex at birth responses.
- Gender data collected by the Ministry of Social Development, while only covering a small part of the population, where available has been found to largely be consistent with census responses for gender.

Appendix 1 provides information on the historical census and admin data that will be used.

The last step used for filling gaps in the data is statistical imputation using CANCEIS (Canadian Census Editing and Imputation System). This system uses 'nearest neighbour donor imputation' to fill gaps in the data. The purpose of CANCEIS imputation in this context is not to capture precise information about the specific individual, but to achieve representative aggregated data by using information from similar individuals to provide realistic values for individuals' missing attributes. For each record with missing information, CANCEIS finds the most similar record in the 2023 Census dataset and copies its information for any variables that need to be imputed. CANCEIS searches for similar records by comparing their values in a defined set of variables called 'matching variables'. 'Donor' records copy over (or 'donate') their values to 'donee' records that have a missing value. To evaluate how well imputation performs, we consider the consistency or accuracy of imputed values compared to true values, at both an individual and aggregate level.[5]

Data source flags are created while processing the data to mark where the value for each variable on each record came from. If someone wants to understand the impact of admin data or statistical imputation or prefers to not consider alternatively sourced data in their analyses, they can request customised data and look at confidentialised, aggregated outputs that, for instance, only use what was on census forms.

[5] Editing, data sourcing, and imputation: Planned approach for the 2023 Census provides more information on statistical imputation using CANCEIS.

Methodologies

The methods for filling gaps in the gender and sex at birth variables in the 2023 Census are presented in Figure 1.

The principle for including each step is that the step must be ethically appropriate and the benefits of having the additional data must outweigh the risk of some of it potentially being inaccurate.

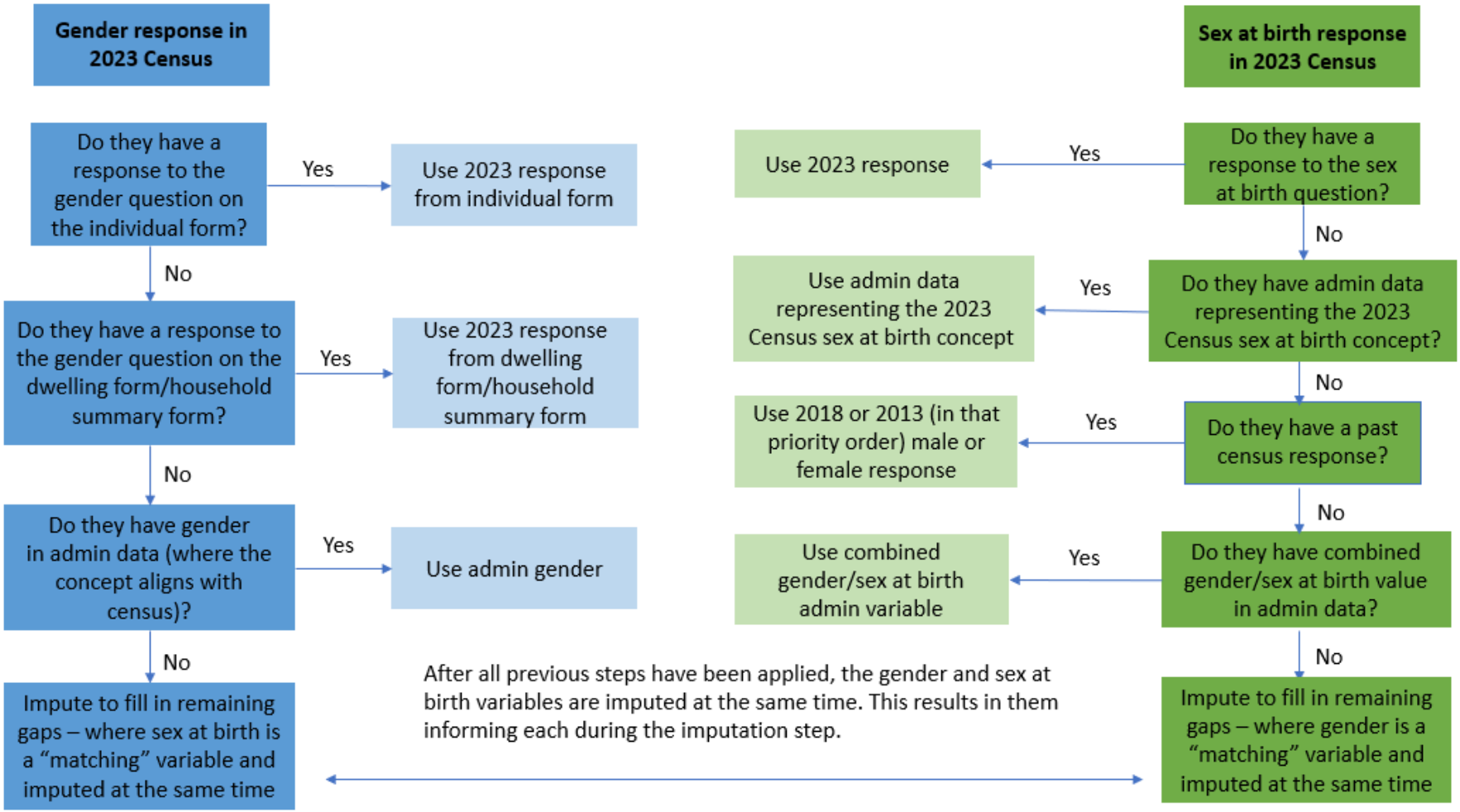
The chosen methodology has been informed by public consultation feedback and engagement with key stakeholders and external experts.

A variety of options were considered and evaluated. The main alternatives are summarised in Appendix 3. The results of testing the individual accuracy and distributional accuracy of alternative data sources are summarised in Appendix 2.

For gender, the methodology is presented in blue on the left-hand side in Figure 1. If a usable response is sourced in a step, that response is used, otherwise the next step is attempted:

1. Use the gender response from the individual form where it is available.
2. Use a gender response from the paper dwelling or online household summary form. This response, while recent, may not always be accurate because often one person fills in this form for the entire household and they may not know how each person identifies. A response on this form is likely to be more accurate than most admin data.
3. Use an admin source that contains gender information using a concept that aligns with the concept in the data standard or is similar to it. This data is high quality and provides accurate values for each record, and the values filled in this way are more likely to be correct than use of statistical imputation in the next step. Currently, the only admin dataset that meets that requirement is Ministry of Social Development data collected since December 2019, which currently has low coverage.
4. Use statistical imputation using CANCEIS. The key matching variable used is sex at birth, so CANCEIS will almost always choose a donor record with the same sex at birth as the donee record. From this donor the response for gender is then copied over. For most of the records in the donor pool, the sex at birth value will match the gender, which means the imputed gender will usually match the donee's record for sex at birth, but there will be instances where it will not. Other matching variables that influence the choice of a most similar record for gender include age, ethnicity, and usual residence location.

Figure 1: Overview of methods for filling gaps in the gender and sex at birth variables



This step for imputing gender will be carried out simultaneously with the step for imputing sex at birth. The imputation of sex at birth for records will not influence the imputation of gender for other records, or vice versa. If both values need imputing, they will be imputed from the same donor record.

Analyses showed this imputation approach to be highly accurate at an individual level for records whose value for gender is either ‘male’ or ‘female’, but to be inaccurate for records whose value is ‘another gender’. However, while the individual consistency is low for another gender, the imputed aggregated distribution of all three categories is consistent with expectations. Using imputation reduces the risk that ‘another gender’ is under-represented overall. None of the other imputation methods explored worked well for ‘another gender’ and often resulted in poorer aggregated distributions. Despite the low individual consistency for records with ‘another gender’, this method is better than not filling the gaps in the gender variable, because not filling them would lead to an underrepresentation of records.

Note: Historical census data about the sex concept is not used for gender because historical census data only had two response categories. Therefore, using it for gender would lead to a further under-representation of ‘another gender’.

For sex at birth, the methodology is presented in green, on the right-hand side, of Figure 1. If a usable response is sourced in a step, that response is used, otherwise the next step is attempted:

1. Use the sex at birth response from the individual form where it is available.
2. Use admin data that is closely aligned with the census concept for sex at birth. Currently, Department of Internal Affairs births data has been found to have a very high consistency with respondents’ answers to the census question. Using this data source will lead to highly accurate data at the individual level, therefore increasing chances that the data used in analyses and to inform policies has the correct distribution. This source is suggested to be used next because there is no other information on the census individual or dwelling form that could inform the sex at birth value. See Appendix 3 for further discussion of this data source.
3. Use historical census responses about the sex concept. The phrasing and the answering options of the previous census questions means people may have interpreted the question as either gender or sex at birth. Testing shows that this historical census data aligns quite closely with the sex at birth concept and is more accurate than the remaining two steps to inform the sex at birth value. If there is a response to the sex concept from the 2013 Census, that will be used; if not, a response from the 2018 Census will be used. It is thought that an earlier record is more likely to correspond with sex at birth.

4. Use the combined sex/gender variable from admin data where its value is ‘male’ or ‘female’. For many variables, Stats NZ estimates the best admin demographic information based on multiple collections of admin data using a set of specific rules. The combined sex/gender variable is not fully aligned with the 2023 Census sex at birth concept. While not ideal, at least there is confidence that this value was associated with the individual record and therefore leads to higher quality and more representative data than statistical imputation for filling gaps in the data, according to our testing.

   This variable is useful for filling gaps in the sex at birth variable, but not appropriate to be used for filling gaps in the gender variable. Analysis has shown that the only values in this mixed admin variable that aligns most closely with the 2023 Census gender concept are from Ministry of Social Development data, which already is one of the steps for filling gaps in the gender variable. As the remaining sources for this variable do not contain ‘another gender’, using it to fill gaps in the gender variable would lead to an undercount of ‘another gender’.
5. Statistical imputation using CANCEIS. We will use the same tool used for filling gaps in the gender variable, and the imputation for both variables occurs simultaneously yet independently as explained in the gender method. This imputation process uses several matching variables to determine the most similar record. For sex at birth, the most important matching variable is gender, meaning CANCEIS will almost always choose a donor who has the same gender response as the donee record. It was found that this method has very high consistency. While we do know that some imputed values will be incorrect at the individual level, this is to be expected of imputation. The key outcome of imputation remains the creation of accurate aggregated data and distributions, and we know the totals that include imputed numbers are more accurate than the totals without them. Given the importance of this variable, we assess that the benefit of having the more appropriate aggregate numbers for each breakdown of the variable outweighs the downside of some values being inaccurate at an individual level and that imputation is preferable to these records being excluded from analyses due to having gaps.

Engagement on gender and sex methodologies

Stats NZ has undertaken significant engagement and consulted publicly to inform decisions related to the collection and dissemination of data on gender, sex, and variation of sex characteristics, and on the methodologies for gender and sex concepts discussed in this paper.

Data standard for gender, sex, and variations of sex characteristics

The data standard was developed through review processes, extensive public consultation, and with an external advisory group.

Sex and gender identity statistical standards: Findings from public consultation July-August 2020 provides more information.

2023 Census gender and sex concepts methodology consultation

From December 2022 to January 2023, Stats NZ consulted publicly[6] on how to fill gaps in census responses to sex at birth, gender, and two new, related questions. This engagement sought to understand sentiment to the concepts and proposed methods to fill gaps in census responses, and to identify topics of concern.

A technical users' forum was delivered in December 2022 as part of this consultation process. The public were invited to attend this meeting.

Overall, the public consultation process yielded information on the principles, values, and reasoning about filling in blanks in these variables, and some information on the use of historical census responses, admin data, or statistical imputation. The feedback received was evaluated and informed further investigations and decisions about the chosen methodologies for filling gaps in sex at birth and gender.

Gender and sex concepts 2023 Census methodology consultation – feedback provides a summary of the feedback received from the public consultation.

[6] Gender and sex concepts 2023 Census methodology consultation provides information on the consultation process.

Data Ethics Advisory Group

Given the sensitivity around sex and gender data and methods, the proposed methodologies were presented for discussion to the Data Ethics Advisory Group[7]. This advisory group provides independent expert advice across different areas to government agencies, including on the responsible management of potential risks and harms.

Stats NZ received feedback to clearly articulate the value of filling gaps in the data, and the benefits and limitations of the methodologies chosen. Continued engagement with other experts was recommended. The advisory group also discussed several considerations around imputed census data when it becomes part of the Integrated Data Infrastructure (IDI), for example, flagging which responses were imputed, which is standard information provided for data users.

Consultations on final methodologies

In accordance with the recommendation provided by the Data Ethics Advisory Group, Stats NZ has conducted external discussions on the proposed methodologies with key stakeholders. These have included representatives from organisations that work with Rainbow communities and/or data, as well as researchers at New Zealand universities who work closely with Rainbow communities' data. These engagements supported increased understanding and refinement of the methodologies. It is key to note that the attendees do not represent all stakeholders and were not engaged to "approve" the methodologies. Participants emphasised that Stats NZ must carefully consider how to frame the outputs of these variables and plan for the future governance of the data; both issues are recognised in the Future consideration for these concepts and methodologies section of this paper.

[7] Data Ethics Advisory Group provides information on this group.

Considerations and limitations

Distinction between people and census individual records

There is a crucial distinction between the people who provided the original data via their census form, admin data, or historical census data and the resulting records for use in statistics and research.

The methodologies in this paper assign a value to fill a gap in a variable on a particular record. When we assign that value, we are not saying the person associated with that record has the characteristic implied by that value. Instead, we are filling the gap with as realistic a value as possible to ensure that the record can be included in related analysis. This helps the overall statistics better represent communities across Aotearoa New Zealand, even though the records will not be accurate for every individual.

It is also important to note that careful processes for data de-identification and confidentialising tables, analyses, and other published data outputs protect individuals' identities. This means that if a person chose not to fill in 'gender' or 'sex at birth' because they are uncomfortable with the concepts, but a value is provided to their record by admin data or imputation, this is not at risk of being visible to others and being associated with them. The data files for the census are processed in secure environments.

Personal identifiers on records are kept separate from linked attributes, so the datasets that can be used for research and statistics do not have identifying information (for example, date of birth, name, or address).

Apply to use microdata for research explains what can be accessed and the process for applying to do research using individual-level de-identified records.

Data is aggregated and used only to inform analyses, not to identify people. All outputs based on the data, produced by Stats NZ and other researchers, must be aggregated and confidentialised so individuals cannot be recognised.[8]

Respondents usually have reasons for answering questions or leaving them blank, and decisions about the gender and sex at birth variables can be particularly personal. If people or organisations prefer to specifically analyse only the data for people whose

[8] Applying confidentiality rules to 2018 Census data explains rules very similar to those that will be applied in the 2023 Census.

gender or sex at birth values came from census responses, they can do so by requesting customised data that includes the data source flags for gender or sex at birth. Similarly, qualified researchers who are interested in exploring the nuances of blanks or responses to these variables can apply to do such a research project with individual level but de-identified records.

The purpose of creating complete variables through alternative data sourcing is not to focus on any individual – indeed, protocols mean that those authorised to do research on individual records cannot see identifying information like name or address.

Filling the blanks through these methodologies does, however, create a dataset with the best quality possible to inform the public and decision-making.

Census data is produced to benefit society. It is considered that having complete variables informed by historical census or admin data for gender and sex at birth, produced with privacy, confidentiality, security, and transparency, outweighs some of the feedback we received to leave gaps in the data unfilled.

Limitations of approach

Low individual consistency for ‘another gender’ category

We tested how well different imputation approaches would perform by taking 2023 Census responses with a completed gender variable and checking what our approach would have filled in if gender had been missing (see Appendix 2 for more detail). A limitation of the approach (and all others tested) is that there is low individual consistency around ‘another gender’. Low individual consistency means that when we compare each person’s actual response with the value this approach would fill in, the values usually do not match.

There currently are few usable admin data sources for gender; therefore, it is hard for admin data to sufficiently fill blanks. After that, the imputation approach leads to high accuracy of the population distribution, or broadly correct numbers of people identifying as ‘another gender’. For example, when we test the percentage of ‘another gender’ responses in a geographic area, the actual percentage from 2023 Census responses is close to the percentage that would result from filled in values. Unfortunately, at the individual record level, the value of ‘another gender’ has very low consistency rates. This somewhat lowers the quality of the ‘another gender’ category in microdata analyses, as this introduces noise into that variable. However, missing values would pose a different problem in those microdata analyses.

While the methodology produces consistent gender values of ‘male’ and ‘female’, it must be acknowledged that it does not provide that individual-level consistency for values of ‘another gender’. However, this methodology is effective for aggregate statistics, and it produces better representation than if gaps were not filled.

Undercount of transgender people

A second limitation of the methodology is an undercount of transgender people through the derived variable ‘cisgender and transgender status’.

Data standard for gender, sex, and variations of sex characteristics explains how and why this status is derived in two steps, separately asking about gender and sex at birth.[9] However, if one of the two variables is missing from the census individual form, then there is not enough information to derive the cisgender and transgender status variable with high quality. For many other derived variables in the 2023 Census, alternatively sourced variables are used in their derivation, with data quality gains from alternatively sourced components flowing through to the derived variables. However, in this case, if alternatively sourced variables were used, the relatively small transgender population and the two-step process to derive the status would compound any quality problems in the sources; filling the gaps well would require an accuracy not present in the alternative data for these two variables.

Filling the blanks directly for the derived variable through alternative data sourcing would not produce high quality data at this time either. Available admin data does not have this information, and it was not part of prior census data. Imputation quality is hard to check without established estimates of the variable’s distribution to check against. More importantly, there are no reliable matching variables to inform the imputation, and there currently is no good way to determine whether there is a strong bias in these variables, which would impact the performance and evaluation of the imputation.

Because of these data quality issues, many records will have the value ‘cisgender and transgender status unidentifiable’, and the number of cisgender and transgender people will be undercounted.

Clearly identifying and communicating the limitations of the data will be important for its interpretation and use.

[9] Appendix B of the data standard provides more information.

Engagement limitations

Stats NZ has engaged and consulted on the 2023 Census methodologies for the gender and sex at birth concepts (for more information see Engagement on gender and sex methodologies). However, not all potential methodological approaches were covered at each engagement, and the people who engaged with us will not represent all people of Aotearoa New Zealand.

Future considerations for these concepts and methodologies

Governance of 2023 Census data

Stats NZ has received feedback from respondents who participated in consultation processes for the methodologies that there is a concern that identities that were not disclosed on census forms may show up in statistical records, first in the 2023 Census data and later in the IDI census data, as a result of using admin data. This is a particular concern for people whose gender is different from their sex at birth; the records would have both values appear together. This is a data governance matter, related to rules about how data is made available and used.[10]

Feedback also outlined that projects looking at both gender and sex at birth must acknowledge and mitigate risks to Rainbow communities as part of getting approval to undertake research using IDI data. This is in line with current rules.[11]

In particular, those consulted requested restrictions on data users deriving indicators for cisgender and transgender status or Rainbow LGBTIQ+ status. Unless an approved project is focused on understanding the gender, sex at birth, sexual identity, or variations of sex characteristics concepts better, there should be appropriate safeguards in place, so researchers do not inappropriately or inconsistently derive their own indicators.

We are reviewing the safeguards for accessing this data via research projects to ensure they are implemented appropriately for these concepts.

---

[10] The Five Safes framework and Ngā Tikanga Paihere provide more information on the principles that guide and inform data practice.

[11] How to apply Ngā Tikanga Paihere to microdata research projects helps researchers identify potential impacts of their research and be responsive to Treaty of Waitangi and human rights obligations. For instance, in applying for a research project, researchers must show they are accountable to the communities impacted by their research findings.

Supporting publication, communications, and engagement

Outputs from the 2023 Census will be the first to include gender as a standard output in crosstabs (for example, tables that show counts or percentages of people by age by gender by region) and is the first census to collect the specific concepts of gender and sex at birth. Stats NZ will proactively offer guidance on the meaning and interpretation of these variables and the methods used to fill gaps to support use of the data by census customers, community organisations, and in the media. A focus will need to be on clear plain English explanations and guidance about what the data can be used for (and limitations on its use), particularly for small population groups such as the transgender population.

Because gender is a new concept in the 2023 Census and there is a conceptual change from sex to sex at birth, any methodological approach for filling gaps is limited by unknowns. We anticipate that in the future, as more alternative sources become available for these variables and after we receive feedback from users of 2023 Census data, we will update this methodology appropriately.

Glossary

Definition of terms and acronyms, as they relate to census.

Term	Description
Admin enumeration	The use of admin data to add people to the usually resident census population when a census response has not been received.
Administrative (admin) data	Data collected by government or other organisations for non-statistical reasons, such as births, tax, health, and education records. These are typically records describing events or interactions with government agencies and have been obtained in the course of some statutory obligation or service provided by a government agency.
Alternative data sourcing	For the 2023 Census, refers to using admin data, historical census data, or statistical imputation when individuals or variables are missing from census responses. Alternative data sourcing can support admin enumeration and fill gaps from missingness.
Another gender	Encompasses any genders that are not male or female. This term is used in the Stats NZ gender question format and classification.
CANCEIS	Canadian Census Edit and Imputation System. A method for ‘imputing’ (filling-in) data for missing responses/respondents. Used by a number of national statistical institutes for census imputation, including Stats NZ.
Cisgender	Refers to a person whose gender is the same as the sex recorded at their birth.
Combined sex/gender variable from admin data	For many variables, Stats NZ estimates the best admin demographic information based on multiple collections of admin data using a set of specific rules. The combined sex/gender variable is not fully aligned with either the 2023 Census sex at birth concept or the gender concept.
Data source flags	Indicator flags created to identify which data source was used for each variable (see alternative data sourcing).

Term	Description
	For customised data requests it will be possible to have flags and filter the data sources used for variables.
Derived variable	A variable created by calculating or categorising other variables, rather than being collected directly from a respondent. For instance, the combination of responses to gender and sex at birth will determine the value of the derived variable ‘cisgender and transgender status’.
Gender	A person’s social and personal identity as male, female, or another gender or genders that may be non-binary. Gender may include gender identity and/or gender expression. A person’s current gender may differ from the sex recorded at their birth and may differ from what is indicated on their current legal documents. A person’s gender may change over time. Some people may not identify with any gender.
Historical census data	Data collected in past censuses. For the 2023 Census, 2013 and 2018 Census values are part of alternative data sourcing. For the sex concept, values are the 2013 and 2018 Census responses to the question ‘Are you Male/Female?’
Imputation	See statistical imputation below.
Integrated Data Infrastructure (IDI)	A large research database maintained by Stats NZ. It contains de-identified data about people and households sourced from government agencies, non-government organisations, and Stats NZ surveys (including the 2013 and 2018 Censuses). Data from different sources are linked together, typically at the individual (person) level.
Intersex	See ‘variations of sex characteristics’.
Matching variables	In CANCEIS, the set of variables used to indicate how similar two records are to each other to find a donor record for the missing value on a concept. Note that different sets of matching variables may be used for different concepts. This is because different sets of matching variables may produce the most accurate imputation for different concepts.

Term	Description
Missingness	The gaps in census responses. Missingness also refers to the amount or degree of missing data.
Sex at birth	The sex recorded at a person's birth (that is, what was recorded on their birth certificate). Sex at birth may also be referred to as sex assigned at birth.
Sexual identity	How a person thinks of their own sexuality and which term(s) they identify with. Sexual identity terms include lesbian, gay, straight, asexual, takatāpui, bisexual, or pansexual, among others.
Takatāpui	An umbrella term for all Māori with diverse gender identities, sexualities, and sex characteristics.
Statistical imputation	The replacement of missing information with values from a statistical process.
Transgender	Refers to a person whose gender is different from the sex recorded at their birth.
Variations of sex characteristics	Innate genetic, hormonal, or physical sex characteristics that do not conform to medical norms for female or male bodies. People may be born with these characteristics, or they may develop in puberty. It refers to a wide spectrum of variations to hormones, chromosomes, genitals, and/or reproductive organs. These variations of sex characteristics may be called 'intersex variations', and some people with these variations may identify as intersex.

Appendix 1: Alternative data sources being used in 2023 Census to fill gaps in gender and sex at birth

See Figure 1 which provides a flowchart setting out when these data sources will be used.

Table A1. Alternative data sources being used in 2023 Census to fill gaps in gender and sex at birth

Data source	Variable	Values	Collection methodology and notes
2013 Census	Sex	Male Female	Respondents were asked on the census form “Are you male or female?”. We assume that due to the absence of a gender variable and no explicit mention on the form of the concept being collected, people may have responded to this question as if asked about their gender, their current sex, or their sex at birth. This was supported by feedback received from the public consultation process. Missing responses were imputed.
2018 Census	Sex	Male Female	Respondents were asked on the census form “Are you male or female?”. Respondents who wished to identify as intersex were instructed to request paper forms and tick both male and female boxes. We assume that due to the absence of a gender variable and no explicit mention on the form of the concept being collected, people may have responded to this question as if asked about their gender, their current sex, or their sex at birth. This was supported by feedback received from the public consultation process. Missing responses were imputed.
Admin: IDI personal details table	Combined sex/gender from admin data	Male Female	This is the combined sex/gender variable from admin data, which merges sex and gender data from all datasets into one table.

Data source	Variable	Values	Collection methodology and notes
		Code = 3	The derivation of this variable is currently using a preassigned priority that researchers are not able to modify. Individuals will be assigned a value for this variable from the highest priority dataset which does not have a missing value. The current sex/gender variable prioritises Department of Internal Affairs (DIA) birth marker information above all others. This means information from lower priority datasets that may be more reflective of gender will not be used if a DIA value is available.
Admin: DIA births	DIA birth sex	Male Female Null	Birth certificates are initially issued when a birth is registered and record the sex assigned at birth. DIA allows a person to change the marker on their New Zealand birth certificate via application[12]. Prior to mid-2023, the available options were male, female, and indeterminate. It is now also possible to change the marker to non-binary. 'Indeterminate' can be used as the marker on a birth certificate when a medical professional cannot determine a child's sex to be male or female when they are born. The marker may be changed to 'indeterminate' if someone's sex was incorrectly registered as male or female when they were born. Currently, there is no timestamp available to Stats NZ to show when a record has been updated. Records other than male or female (for example, indeterminate, not recorded) were

[12] Change the registered sex on your birth certificate provides information about this application process.

Data source	Variable	Values	Collection methodology and notes
			mapped to ‘null’ on data ingest prior to the October 2023 IDI refresh.
Admin: Ministry of Social Development (MSD) benefits	MSD sex/gender	Male Female Gender diverse	We assume this is gender for benefit applications since December 2019, but we do not know whether this variable is sex or gender before then. The third category, ‘gender diverse’, is not fully aligned with but similar to the 2023 Census concept of ‘another gender’.

Appendix 2: Comparing gender and sex at birth data sources

We tested the accuracy of alternative data sources.

Testing which data sources agree with census responses

For the testing, we used real data including both gender and sex at birth, then removed some values at random. Because we knew the true values behind the missing data, we were able to test the accuracy of filling the gaps in different ways. We compared interim 2023 Census gender and sex at birth responses received between March and early April 2023 with 2013 and 2018 Census responses and relevant admin data records. We looked for consistency or agreement in the responses provided across different sources. The results are indicative only, performed on about 80 percent of census forms and before data processing and linking of census responses to the IDI.

Table A2. Comparison of 2023 Census gender with other data sources

2023 Census gender	Agreement with 2023 Census gender, percent			
	2018 Census sex	2013 Census sex	DIA births	Combined IDI sex/gender
Male	99.1	99.1	99.4	99.3
Female	99.2	99.1	99.3	99.2
Another gender	n/a	n/a	n/a	n/a

Results of testing the gender responses from 2023 Census forms against other data sources showed a high level of consistency for male and female categories. This is mathematically expected because sex at birth is the main matching variable and records where gender is the same as sex at birth are in the vast majority of those donating values and receiving values in this test.

We found the ‘another gender’ category is not reliably present in any other data sources. Separate reviews of Ministry of Social Development data showed its sex/gender variable to be high quality but currently very sparse.

We also compared the 2023 Census gender response from the individual form with the response from the dwelling form or household summary form and found that consistency is lower. However, this could be due to the circumstances of testing – using census responses before data processing and before all census responses were linked

to the IDI spine – or because the person filling out the dwelling form or household summary form filled out an incorrect gender.

Consistency between 2023 Census gender and other sources is lower than consistency between 2023 Census sex at birth and other sources.

Table A3. Comparison of 2023 Census sex at birth with other data sources

2023 Census sex at birth	Agreement with 2023 Census sex at birth, percent			
	2018 Census sex	2013 Census sex	DIA births	Combined IDI sex/gender
Male	99.6	99.6	99.9	99.8
Female	99.7	99.7	99.9	99.8

Results of testing the sex at birth responses from 2023 Census forms against other data sources showed a very high level of consistency for the male and female categories, especially between 2023 Census and Department of Internal Affairs birth registration records (99.9 percent).

Imputing for gender and sex at birth variables

Preparing a test dataset

We tested the potential performance of imputation for both gender and sex at birth variables by randomly selecting one million records from the usable records in the 2023 Census dataset that were able to be matched with other datasets.

To replicate conditions under which we expect the imputation to be run (based on what we saw in the census data) we removed gender and sex at birth data from some records:

- 27.5 percent of rows in the test dataset were selected to have some missing values
- missingness (number/amount of missing responses) for all variables was set to 14 percent.

The above adjustment resulted in census responses for gender and sex at birth both missing in 6.8 percent of records in the test dataset, and for about 0.2 percent of those records, admin gender or admin sex at birth were also missing.

Using CANCEIS imputation on this dataset we tested three scenarios:

1. Imputing gender and sex at birth without drawing on admin data.

2. Using admin gender as an additional variable to inform the imputation.
3. Using admin gender or admin sex at birth to first fill gaps, and then imputing remaining missingness.

For all analyses, imputed and real values were compared and the proportion of imputed values that recovered the ‘true’ respondent-supplied value was our consistency measure. We also checked accuracy of population distributions by comparing the overall imputed proportions of male, female, and another gender against the proportions observed in the response data.

Results of imputation testing

Table A4. Comparison of different imputing scenarios

Test scenario	Results
Scenario 1. Imputing gender and sex at birth without drawing on admin data	For the first test, we used available gender response values to inform imputation of sex at birth, and sex at birth response values to inform imputation of gender. Where one of the two variables was missing, the imputed values were consistent with the respondent-supplied values for around 76 percent of the records, but this dropped to 50 percent where both variables were missing. These results served as the lower bound for the other scenarios.
Scenario 2. Using admin data as an additional variable to inform the imputation	For this scenario, we added the combined IDI sex/gender data as an additional matching variable. The consistency rates were much higher than scenario 1 at 97.8 percent for gender and 99.0 percent for sex at birth. To show how statistical imputation performs for the records which originally had specific genders (for example, for male, female, or another gender), we compared the imputed values with the respondent-supplied data. We found that the consistency of imputation for male and female records was equally high with over 98 percent of records being imputed with the respondent-supplied gender. But the consistency for ‘another gender’ records was very low, with the vast majority being imputed as female or male rather than another gender.
Scenario 3. Using admin gender or	In scenario 3, for missing sex at birth records, we filled in the value with the individual’s admin sex at birth or historical

Test scenario	Results
admin sex at birth to first fill gaps, and then impute remaining missingness	census data, where available, and then used CANCEIS to impute the remaining missing sex at birth records. We used a similar method for gender but did not use the historical census data (as no historical census gender data is available). The consistency of records where we filled in sex at birth with admin sex at birth was 99.8 percent, the same as the match rate between 2023 Census and combined IDI sex/gender shown in table A2. After using CANCEIS to impute for the remaining records, the overall consistency rate for sex at birth was 99.2 percent.

For all scenarios, the distributions (male, female, another gender) of imputed gender data were very close to the respondent-supplied data, although imputed values on another gender records were assigned almost randomly due to lack of informative matching variables.

Table A5: Evaluation criteria summary for investigated options

Scenario	Accuracy of population distribution	Consistency (individual record level)
1. All imputation, no admin data	Very high	Low when one variable missing Very low when both variables missing
2. Admin data to guide imputation	Very high	Very high for male and female categories Very low for another gender category
3. Admin data to fill gaps followed by imputation	Very high	Very high for male and female categories Very low for another gender category

Based on individual level consistency and agreement between distributions, scenario 1 – which does not draw on admin data at all – is not a preferred option. The consistency values were the lowest for this scenario, especially where both gender and sex at birth

responses are missing. As expected for approximately 50-50 male/female population split, imputation is correct only 50 percent of the time.

Scenarios 2 and 3 both performed better, giving very similar results to each other. This is not unexpected, considering that CANCEIS attempts to find the closest donor based on the matching variables and considering that census sex at birth and admin sex at birth match in 99.9 percent of cases in this test. Overall, both scenarios achieved very high individual consistencies and preserved the distributions for the three gender options well. However, the individual level accuracy was very low for ‘another gender’ records.

Appendix 3: Overview and discussion of alternative methods considered

This appendix discusses the strengths and weaknesses of different methodologies that were considered for filling gaps for gender and sex at birth variables in the 2023 Census but are not being used.

Using gender to inform sex at birth and vice versa

In cases where only one of the two variables is missing, an alternative methodology could have been to use the response provided for one variable to fill in the value of the missing response, that is, if gender was missing, the sex at birth variable on the individual form would be used to fill the gender variable. The benefit of this approach is that it uses only information provided by the respondent on the form. However, it neglects the difference between the two concepts and assumes that both values are the same. This approach would be inaccurate, because we know that this is not the case for everyone, and it would lead to an under-representation of the ‘another gender’ category. Therefore, it is more appropriate to use alternative ways of filling in the missing value that allows both variables to be different.

Using admin data as a matching variable in statistical imputation

Another alternative considered was whether, instead of using admin data directly to fill in the missing variables, it should be used as a matching variable in statistical imputation. The idea behind this alternative is that admin data is not necessarily 100 percent correct and there are some records where admin data and census responses do not match. Another key part of this proposal is that it may be viewed as being less ethically problematic, and that it may be seen as a step removed and less invasive. Using it as a matching variable in statistical imputation would mean that the values most of the time are the same, but have a chance to be different, reflecting the minor uncertainty in admin data.

Investigations showed that the overall results were very similar if admin data was used directly or as a matching variable. The reason for this is in how the process for statistical imputation works. The admin variable and the census response variables are very similar for sex at birth. For CANCEIS, when imputing gender, it makes no difference whether the sex at birth variable is missing and the admin variable is used for matching, or the admin variable was first used to fill the sex at birth variable which is then used for matching.

There are some caveats around using admin data directly. The first is that while the concepts used in the census and the Department of Internal Affairs births data are closely aligned, they should not be assumed to be the same. People are able to change their marker on their birth record (the process for which has recently been streamlined through the Births, Deaths, Marriages and Relationships Registration Act 2021). Secondly, as outlined in the findings of the consultations and engagement, sex at birth may be a particularly sensitive concept for respondents, and it is important to acknowledge this and to treat this variable appropriately. In this context, it is important to emphasise that using the alternative data is *not* about assigning a sex at birth to a person but ensuring the distribution at an aggregate level is representative. We have done thorough investigation into alternative methods. We have recognised that using admin data – which in most cases is as true as we can get to the concept for the individual records – is not always the most appropriate. The decision to use admin data, in the end, was made on balance, considering a range of factors and making trade-offs. Some of this discussion is detailed in the following section.

Not using some of the steps for filling gaps in specific scenarios

There were a few alternatives considered that would skip some of the steps of the methodology in specific scenarios. The scenario discussed most was if respondents had provided a gender on the individual form but not a sex at birth. A hypothetical case was discussed where the reason for missingness of a sex at birth response was due to discomfort with the concept. The rationale for not using admin data in these cases is that it would avoid accessing a person's admin data where they had chosen not to have this associated with them on their 2023 Census form. In these cases, it might be more considerate to use statistical imputation instead. As previously mentioned, for all variables in the 2023 Census, Stats NZ will provide information on the breakdown of data sources used, and for customised data requests it will be possible to have data source flags and filter on the different data sources used for variables.

On the other hand, we do not actually know why some people fill in one variable and leave the other blank. If using admin data there is at least some certainty that the value was associated with the person at some time, while statistical imputation is far less certain at the individual level.

A key argument, however, is that this same situation applies to a lot of other variables as well, and not using admin data in this case would create an exception that is difficult to justify. This is not desired because, overall, using admin data has been shown to substantially increase data quality and contribute to more representative data and better outcomes for communities, so the assessment is that the benefit outweighs the risk. And there is an important distinction to be made about records versus people: the

process of using admin data to fill gaps in the census data is not meant to assign the respective characteristic to the person, but to assign a value to the record that is the most accurate at an individual level for statistical purposes.

Not using statistical imputation for either or both variables

The last alternative considered was to not use statistical imputation for either variable, especially gender. As with the use of admin data above, rationales for avoiding imputation included not wanting any value of the variable associated with their data or not wanting ‘inaccurate’ values in the data. The main methodological argument against imputing gender is the low individual consistency for another gender. The key reasons for ultimately using imputation are the arguments mentioned earlier for why gender needs to be a complete variable. There is not a major methodological argument against using imputation on sex at birth because that imputation performs well on both individual consistency and accurate distributions.

Admin enumeration: Planned approach for the 2023 Census describes how we will count people who submitted no 2023 Census response.

There are limited usable admin data sources for gender currently available, so anyone who did not have a census response for gender, including admin enumerated people, will likely contribute to substantial missingness. Without imputation, the missing values would lead to a substantial undercount of another gender. Additionally, the proportion of admin enumerated people is not uniformly distributed across Aotearoa New Zealand, but higher in some areas and for some subpopulations. If the gender variable were not filled with imputation, the missingness would affect some populations far more than others, potentially disadvantaging people in these areas or populations who were under-represented.

There is considerable value to this imputation. Even though the consistency of the imputed values is low at the individual record level for another gender, the distributions appear highly accurate even at subnational levels, and crosstabs with other variables of interest appear accurate as well. Using imputation allows for accurate distributions for another gender, and we cannot achieve this without using imputation in our methodology.