Linkage error and linkage bias: A guide for IDI users

Amanda Kvalsvig
Sheree Gibb
Andrea Teng

July 2019

Department of Public Health | Te Tari Hauora Tūmatanui
University of Otago, Wellington | Te Whare Wānanga o Otāgo ki Te Whanga-nui-a-Tara

Box 7343 Wellington, New Zealand 6242
Street address: Level J, 23a Mein St, Newtown

www.otago.ac.nz/wellington

TABLE OF CONTENTS

Acknowledgements

- This report was funded by the *Healthier Lives* National Science Challenge, Ministry for Business, Innovation and Employment (MBIE).
- The authors would like to thank the advisors and reviewers for their helpful comments, particularly Michael Alspach, Anapapa Mulitalo and Nancy Wang from Stats NZ, and the Virtual Health Information Network (VHIN) Executive comprising university and public sector representatives.

Aims and scope of this report

This guide discusses linkage bias in the Integrated Data Infrastructure (IDI). This is an area in which there is almost no existing research. No previous studies have examined linkage error or linkage bias in the IDI. Of the 397 projects listed in the IDI project database only two mention linkage bias, and both of those are focussed solely on links within census data. The overwhelming majority of IDI analysis is completed without any attempt to understand, measure or correct for potential linkage bias. Because of this lack of research, we believe that most IDI researchers are poorly informed about linkage error, potential bias, and the extent to which this may impact their findings. This guide is a first step in attempting to increase literacy about linkage bias amongst IDI researchers.

The aim of this report is to provide IDI researchers with an introduction to the concepts of linkage error and linkage bias, describe how linkage happens in IDI, and identify some paths to progress research in this important area.

This guide is organised into three sections. In the first section we introduce some basic linkage concepts, including a description of common data linkage methods, linkage error; and linkage bias. The second section is an overview of data linkage in the IDI with a short discussion of how linkage quality is currently measured and reported in the IDI. The final section outlines some possible paths for research on linkage bias in the IDI, given what we know about the types of information available to researchers.

This report does not suggest solutions to the problem of linkage error and bias, or describe methods by which researchers can correct for linkage bias. Linkage bias is an area of emerging research internationally, and there is currently no research on linkage error or bias in the IDI. For this reason, much more research is needed before we are in a position to recommend methods that researchers can use within the IDI. Nonetheless, this paper represents a first step towards developing a programme of research on linkage error and bias in the IDI.

This guide focuses on linkage bias, but we recognise that there are many other potential sources of bias in IDI analyses. For example, selection bias may arise through unequal access to services determining whose information is captured in administrative datasets. These biases are important but are beyond the scope of this guide.

Summary and key points

Background

- The purpose of record linkage is to link an entity (e.g. a person) in one dataset to the same entity in other datasets. This process produces error which can lead to bias.
- No previous studies have examined linkage error or bias in the IDI. There is no current information or guidance for researchers about linkage bias in analyses using IDI data.
- The aim of this report is to explain the concepts of linkage, linkage error and linkage bias as they apply to the IDI, and provide some suggestions for advancing research in this area.

Record linkage

- Linkage procedures are used in the IDI to link tax, births and visa data to create the spine; to link datasets within collections, e.g. datasets in the Ministry of Health collection; and to link collections to the spine.
- There are several types of record linkage procedure: deterministic, e.g. using a unique common identifier such as National Health Index (NHI); probabilistic, where record pairs are assigned a weight which should correlate with the probability of being a link; and a group of newer approaches e.g. machine learning. The IDI uses deterministic and probabilistic approaches.

Linkage error and linkage bias

- Linkage error is inevitable in any record linkage project, including in the IDI. If linkage errors are not distributed evenly across groups, this can lead to bias. Bias in an analysis may lead to incorrect conclusions.
- The impact of linkage bias can be high even when the error is small; and even a large amount of error will not necessarily produce bias.
- The amount of bias will vary depending on the study population used, variables analysed and their relationships with one another. As a result, all IDI researchers need to have a working knowledge of linkage error and bias to guide interpretation of their findings.

Next steps

- To better understand linkage bias in IDI research we need to: measure linkage error better; make use of existing bias methods such as quantitative bias analysis; and increase researcher literacy around linkage error and bias.
- A longer term goal is to develop automated methods that measure and correct for linkage bias in IDI analyses.

Section 1

Introduction to linkage concepts

"The primary goal of record linkage is to link an entity (e.g. person and household) from one file to the same entity in other file(s)".[1]

This section describes some of the theory behind record linkage and introduces readers to some of the important concepts in record linkage. These will form a base for later sections which discuss record linkage and bias as they pertain to the IDI.

Record linkage has a number of applications in population research because of the important insights that can be gained by linking data from multiple sources. Linked data can help researchers and policymakers understand the complexities of people's lives and the structural factors that underlie these complexities. For example, record linkage of health data to non-health data can provide key contextual information for researchers investigating health issues that have causes or consequences beyond the health system itself.

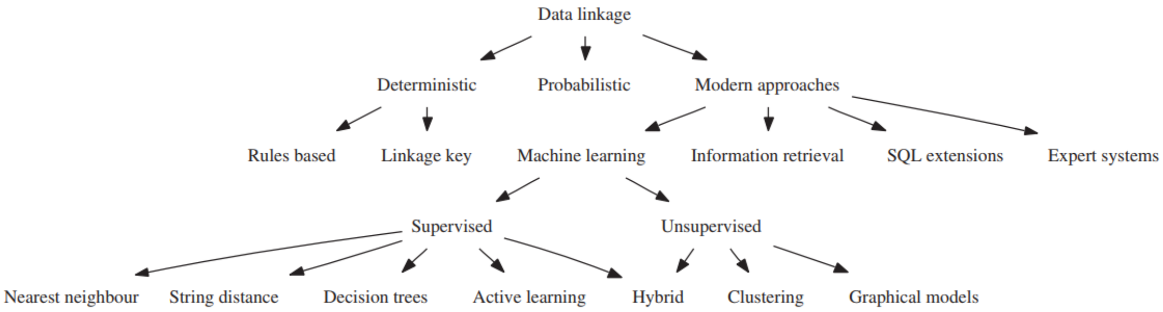
Data linkage methods

There are several main types of record linkage procedures:

- Deterministic, where records are linked according to a pre-specified rule about the level of agreement in the matching fields that is required. Record linkage using a unique common identifier (e.g. National Health Index (NHI)) is a common deterministic approach within health data.
- Probabilistic, where record pairs are assigned a weight which reflects the probability of being a correct link. Potential links are accepted or rejected using a set of rules and cut-offs (e.g. the Fellegi-Sunter method[2]; see below where this method is described in more detail as it applies to the IDI).
- A group of emerging techniques, which Christen and Goiser refer to collectively as “modern approaches”;[3] these include machine learning, information retrieval, SQL extensions, and expert systems.

Christen and Goiser’s linkage taxonomy is reproduced below (Figure 1). Readers who are interested in reading in more detail about the range of potential linkage methods are referred to their paper ‘Quality and complexity measures for data linkage and deduplication’.[3] We do not further discuss the “modern approaches” in this guide because currently, all data linkage in the IDI is deterministic or probabilistic. However, it is important for analysts to be aware of the existence of newer linkage methods as they may be increasingly used in the future.

Figure 1. Taxonomy of data linkage techniques. Image reproduced from Christen and Goiser.[3]

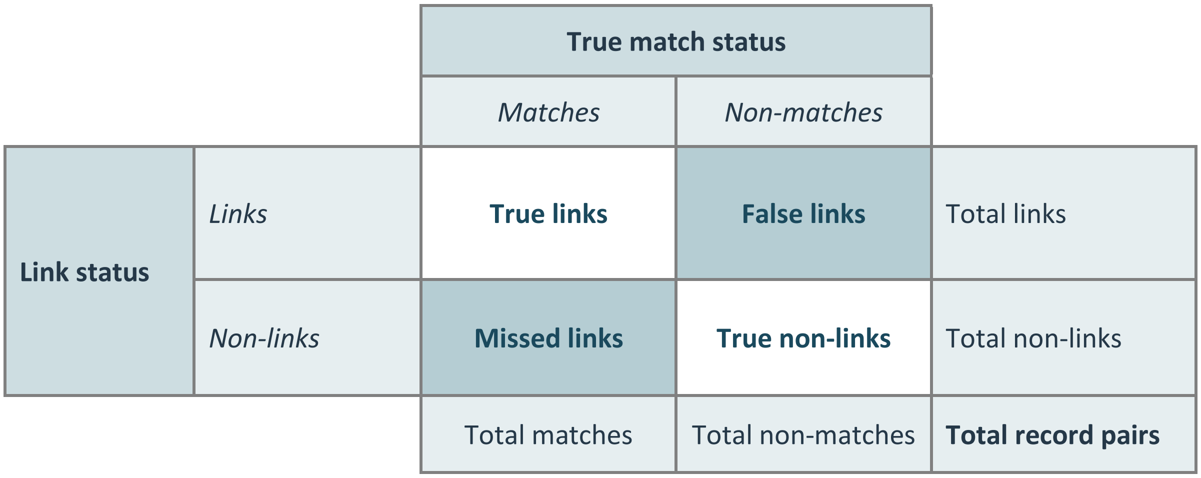


Linkage error

All types of linking processes will produce errors. As the first step in assessing the potential for linkage bias, analysts need to understand linkage error in their study datasets. This understanding requires consideration of how linkage error might arise in the data, and quantitative information about linkage quality that might be available from the linking process itself or from logic checks within the dataset.

As an introduction to linkage error terminology, Figure 2 shows the two-way relationship between link status (i.e. links or non-links assigned by the linkage procedure) and match status (records that are true matches or true non-matches); the two main types of error are then summarised in Table 1. In the next section the two-way relationship shown in Figure 2 is used as the basis for calculating common measures of linkage error and linkage quality.

Figure 2. Potential combinations of matches and links. Adapted from Bohensky.[4]



		True match status		
		Matches	*Non-matches*	
Link status	*Links*	**True links**	**False links**	Total links
	Non-links	**Missed links**	**True non-links**	Total non-links
		Total matches	Total non-matches	**Total record pairs**

As seen in Figure 2, two important types of linkage error are false links and missed links. Table 1 summarises these types of error, showing that they reflect different challenges in the linkage process, and can result in different types of linkage bias. Linkage procedures frequently involve managing trade-offs between false links and missed links because reducing false links will tend to increase the risk of missed links, and vice versa. Linkage error can result in merging (separate individuals are combined into one) or splitting (one individual is represented multiple times in the data).

Table 1. Types of linkage error and how they arise. Information adapted from Harron et al.[5]

Error type	False links	Missed links
Also known as	False positives	False negatives
What is the error?	Records are linked but they actually belong to different individuals	Records from the same individual are not linked
Common sources of error	Identifiers do not discriminate well between individuals: • Large file sizes • Many people share identifiers e.g. age and sex	Usually from errors in identifiers: • Typographical errors • Changes over time (e.g. married women changing their surnames) • Missing or invalid data
Type of bias that might result	Information bias (i.e. misclassification or measurement error)	Selection bias or information bias

Measuring linkage quality

In general, linkage quality is described in terms of the types of linkage error and the magnitude of these errors. An understanding of the relationships between matches and links can help analysts interpret data about linkage quality and decide what further information they would need to assess linkage quality.

Figure 3 shows how the joint distribution of matches and links shown in Figure 2 can be used to calculate standard measures of linkage error. The letters (a,b,c,d) assigned to cells follow the conventional notation for calculating sensitivity, specificity, and positive- and negative predictive value, and readers familiar with these measures will find that they have useful parallels in measurement of linkage quality. Note however that the link rate is not strictly a measure of quality – it is an indicator of how much the populations of two datasets overlap. It is mentioned here because it is reported together with false positive rates for IDI linkage projects.

An important limitation of these calculations is that some aspects of the distribution can be hard to observe. Barriers to measuring and reporting on linkage error include:

- In some datasets, including the IDI, the true match status can never be known, i.e. whether a given linked record belongs to group a+c (it is a true match) versus b+d (it is a true non-match). Manual audit of links can provide an estimate of the linkage error but this process is itself subject to error. In a manual audit process a sample of links (a+b) is inspected and the proportion of false links from that sample (b/(a+b)) is reported. Predictive approaches can also be useful.[6]
- Missed links are difficult to estimate. A manual audit for missed links is difficult because it would require auditors to inspect all possible links for a given unlinked record to distinguish between a missed link and a true non-link, and in a large dataset like the IDI there might be millions of records to examine for each potential link. See Section 3 for further discussion of strategies to quantify missed links.

Figure 3. Measures of linkage error and how they are calculated.

		True match status		
		Matches	*Non-matches*	
Link status (Links assigned by algorithm)	*Links*	**True links** **a**	**False links** **b**	Total links **a+b**
	Non-links	**Missed links** **c**	**True non-links** **d**	Total non-links **c+d**
		Total matches **a+c**	Total non-matches **b+d**	**Total record pairs** **a+b+c+d**

Measure	What it means	How it is calculated
Match rate	Proportion of matches that are correctly identified; can be thought of as the sensitivity of the linkage method	a/(a+c)
True negative rate	Proportion of non-matches that are correctly identified (i.e. the specificity of the linkage method)	d/(b+d)
Precision rate	The proportion of matches that are true links (i.e. the positive predictive value of linkage)	a/(a+b)
Negative predictive value	The proportion of non-matches that are not true links	d/(c+d)
False positive rate	The proportion of total matches that are false matches; also known as the ‘false match rate’	b/(a+b)
Link rate	Proportion of total records that are linked	a+b/(a+b+c+d)

How does linkage error produce bias?

The classic definition of bias is “the lack of internal validity or incorrect assessment of the association between an exposure and an effect in the target population”;[7] bias is also used to describe incorrect assessment of other types of parameter e.g. prevalence.

In epidemiological investigations, the potential for linkage bias arises when the probability of correct linkage varies according to factors that are relevant to the topic under investigation.[8] For example, in Aotearoa New Zealand, linkage error might vary with age, ethnicity, or socio-economic position. This type of differential linkage error might cause distortions in the data in such a way that analysis of the linked population can generate incorrect estimates of prevalence or effect.[9] Linkage bias is a concern because biased findings may lead researchers to draw unfounded conclusions about effective policy or practice.

Note that the term ‘linkage bias’ describes the origin of the bias (i.e., in the linkage procedure) but it is not a bias mechanism as such. The two major mechanisms of systematic error in linkage procedures are selection bias and information bias.

In the presence of missed links (false negatives), errors may lead to biased findings through a selection bias mechanism when they determine inclusion or exclusion in the analysis, or misclassification if errors occur in variables used to classify exposures or outcomes. In the presence of false positive links, incorrect information may be linked to individuals in the dataset, leading to biased findings through mismeasurement or misclassification of key variables.

An important implication of the above mechanisms of bias is that the impact of linkage bias may vary for a given linked dataset depending on the analysis of interest, because linkage *error* is a property of the linked data, but linkage *bias* is a property of the analysis. Findings about linkage bias in one analysis cannot necessarily be transferred to a different analysis with a different causal structure, even when it uses the same population and the same data. This specificity of the bias effect requires analysts to have a working knowledge of linkage bias and how to assess its impact.

Impact of linkage bias

The impact of linkage bias can be high even when the error is small;[10] conversely, a large amount of error will not necessarily produce bias. This is because the impact of linkage error depends more on how it alters the *structure* of the data than on the number of errors that have occurred. For example, if an event is rare it would require only a small decrease in specificity for many or the majority of assigned events to be false, with consequent implications for any conclusions drawn from the data.

In the hypothetical example below, we demonstrate these patterns. The analyst is using linked data to compare the risk of diabetes in individuals from Island A compared with the reference population (i.e. those who are not from Island A).

In this example, more missed links occur in the Island A population than in the reference population. An example of how this could happen is if the linking software has been trained on typical surnames found in the reference population and as a result, the algorithms frequently fail to identify different spellings of the same Island A name. For this example, the prevalence of diabetes in the reference population was set to be 1%, and the observed risk ratio (RR) of diabetes in the Island A population (compared with the reference population) was set to be RR=1.8.

, and Table 3 shows the numbers that the RRs are based on, demonstrating the effect of missed links on distributions and sample size. The method for calculating the bias is shown in the Appendix.

Table 2 shows the true and observed risk ratios obtained under two different linkage error scenarios, and Table 3 shows the numbers that the RRs are based on, demonstrating the effect of missed links on distributions and sample size. The method for calculating the bias is shown in the Appendix.

Table 2. Worked examples illustrating the complex relationship between linkage error and linkage bias.

	Scenario	Percentage of missed links (%)				Risk ratio of diabetes (Island A compared to reference population)	
		Diabetes		No diabetes			
		Island A	Ref	Island A	Ref	Observed	Corrected
1	Island A individuals are much less likely to be linked than the reference population, but the proportions linked are similar whether they have diabetes or not	40	2	40	2	1.80	1.80
2	Island A individuals with diabetes are more likely to be linked than those without diabetes (and overall are less likely to be linked than the reference population)	5	2	20	2	1.80	1.52

The tables illustrate two important concepts in understanding linkage bias:

- Even when there is a large difference in missed links between the Island A population and the reference population, this difference will not in itself result in a biased risk ratio (although it may reduce the precision of the estimate by decreasing the sample size);
- However, when there is *differential* linkage error (e.g. by both Island and outcome), the estimate can be strongly biased in the presence of a much smaller magnitude of error.[6 11]

An additional point is that as noted above, missed links are typically harder to quantify than false links. Without knowledge of the true percentage of missed links or some indication of the likely magnitude and joint distribution of these errors,[6] the analyst would not know which scenario they were dealing with and hence, whether the observed RR of 1.8 was valid or biased.

For more detail on the literature around reporting linkage error and its effects on different types of analysis, see the systematic review reported in Harron et al.[5]

Table 3. Distribution of cases and non-cases in the worked-example scenarios.

		Island A	**Reference**
Observed numbers and distribution in linked data			
	N	100,000	100,000
	Diabetes	1800	1000
	Diabetes %	0.9%	0.5%
	RR	1.80	
True numbers and distribution			
Scenario 1: Non-differential missed links	N	166667	102041
	Diabetes	3000	1020
	Diabetes %	1.8%	1.0%
	RR	1.80	
Scenario 2: Differential missed links	N	124655	102041
	Diabetes	1895	1020
	Diabetes %	1.5%	1.0%
	RR	1.52	

Section 2
Linkage error and linkage bias in the IDI

Part 2 of this report explains record linkage in the IDI, including how records are linked, and how linkage quality is measured. This information was current as at July 2019 but may change in the future.

What is the IDI?

The Integrated Data Infrastructure (IDI) a database consisting of linked data. Data from government agencies, Stats NZ surveys, and other sources are linked to form a national-level longitudinal dataset that can be used for research, policy development, and reporting of national statistics.

A detailed description of IDI linkage can be found in the report “Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project”;[12] what follows is an overview of elements of IDI linkage that are relevant to error and bias.

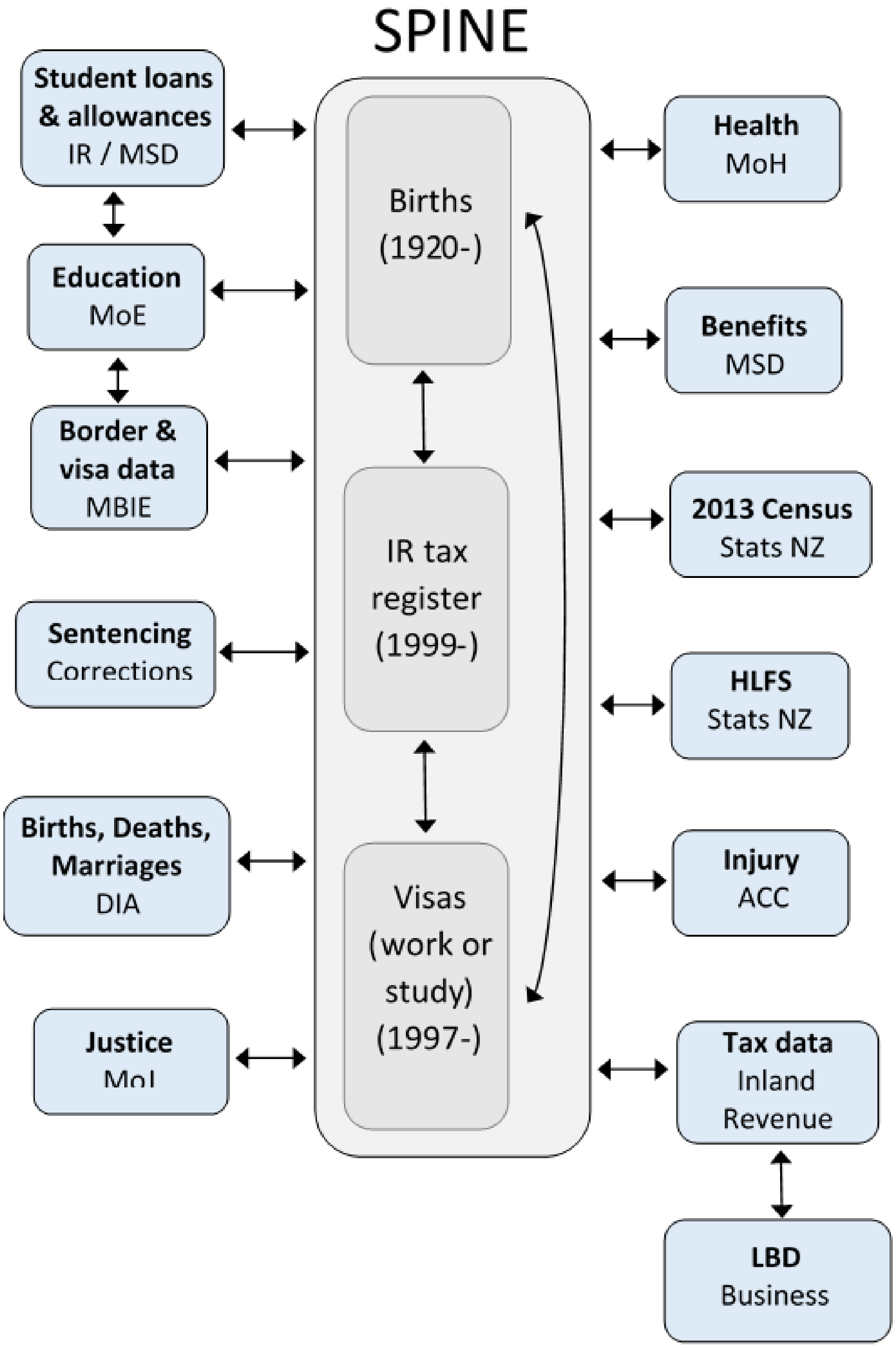
The IDI consists of a central spine and many nodes (collections of datasets linked to the spine).

The **IDI spine** is intended to capture the ‘ever-resident’ NZ population, and is itself the result of record linkage between three key datasets: tax from 1999 onwards, NZ births from 1920 onwards, and long-term visa approvals from 1997 onwards. For a detailed description of the IDI spine and how it is constructed, see “The IDI prototype spine’s creation and coverage” by Andrew Black.[13]

Nodes are collections of datasets that share a common identifier, and usually collected by the same agency. For example, the health node includes datasets such as pharmaceutical dispensing, lab tests, and hospital discharges, all linked by National Health Index (NHI). **Nodes** are linked to the spine; in a few cases they are also are linked to one another.

Figure 4 shows the relationships between some of the datasets in the IDI.

Figure 4. Linkage between datasets in the Integrated Data Infrastructure (IDI). (Diagram by Sheree Gibb)



How are records linked in IDI?

Record linkage happens at several different places in IDI:

- To link tax, births and visa data **to create the spine**. As the three datasets do not have any agency identifiers in common, probabilistic matching is used, based mainly on name, date of birth, and sex, and in some cases passport data. The creation of the spine involves three probabilistic projects, linking tax to births, tax to visa, and births to visa.[13]
- To **link datasets within nodes**. For example, within the Ministry of Health collection there are a range of health datasets that must be linked together. This is done using deterministic linkage with agency identifiers (in the case of health, NHI number).
- To **link nodes to the spine**. As above, most datasets do not share agency identifiers so matching is done using a combination of deterministic and probabilistic matching on name, date of birth, and sex. Address information is used to link all household surveys in the IDI.

It is important for researchers to understand the linkage procedures that were used to generate the data they are using, particularly in the case of probabilistic linking which involves a number of steps and judgement decisions. As mentioned above, the Stats NZ report on linking methodology[12] is an excellent source of additional detail. For readers interested in understanding deterministic and probabilistic linkage in more detail, “Overview of data linkage methods for policy design and evaluation” by Natalie Shlomo[14] is a good introduction. Another good resource is Stats NZ’s data integration manual which has IDI-specific discussion of linkage methods.[15]

Probabilistic linkage in the IDI

Probabilistic record linkage, as the name implies, involves calculations that estimate the probability that two records are a match. As described by Shlomo,[14] the key technical challenges for probabilistic linkage are:

- “The availability of good-quality **identifiers** to discriminate between the entity to whom the record refers and all other entities
- Deciding whether or not **discrepancies in identifiers** are due to mistakes in reporting for a single entity
- Processing **a large volume of data** within a reasonable amount of computer processing time”.

The data linkage approach used in the IDI is designed to address these challenges and in the next paragraphs, we examine them one by one. For a detailed discussion of the design of linking methodologies in the IDI, we recommend that readers consult Stats NZ’s data integration manual (http://archive.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx).[15]

Commonly-used identifiers

In the IDI, some identifiers are available across all projects:

- first names
- last names
- sex
- year of birth, month of birth, day of birth.

These identifiers are compared one by one in a series of **passes.** Sometimes there is a unique identifier available in both datasets that is not of high enough quality or coverage to use for deterministic linking, but it is useful information that can be incorporated into a probabilistic link. For example, the first pass could be an exact match on IRD number, followed by probabilistic linkage in subsequent passes.

Discrepancies in identifiers

Probabilistic linkage is undertaken in situations where record identifiers are non-unique (e.g. a first name may be shared by many individuals) and may also vary for a given individual in different datasets (e.g. street name changing as people move). Some records can be uniquely identified and linked using a set of identifiers as a whole, but pairs of records often have both similarities and differences in their identifiers (as in the examples above) and linkage procedures need to be able take this uncertainty into account when assigning link status.

The mathematical model developed by Fellegi and Sunter[2] is designed to optimise these linkage decisions, and this model is incorporated into linkage software used in the IDI (see below). For those who are interested, a detailed description of the method and its challenges can be found in the original paper[2] and in Schlomo's overview of data linkage methods.[14] There are two key aspects that IDI users need to be aware of from the point of view of understanding linkage bias:

- Although some procedures are automated, several parameters and decisions are determined by human intervention; and
- The method is still vulnerable to certain types of error in the data.

In essence, a probabilistic data linkage procedure using the Fellegi-Sunter model generates a weight for each record pair, such that a higher weight reflects a higher probability of a match[12] (if certain assumptions hold). Note however that technically the weight is a score that correlates with the likelihood of a match, and is not itself a measure of probability.[16]

Based on the weights that are assigned, a record pair can be assessed as 'near-exact' (strong or complete agreement) or 'non-exact' (all other links). The distributions of near-exact and non-exact links can be visualised using a histogram. Visual inspection of the weights distribution and other criteria[17] can be used to determine the cut-off value, which in turn determines whether a record pair is assigned as a link (i.e. the weight is equal to or above the cut-off) or non-link (i.e the weight is below the cut-off). Alternatively, upper and lower cut-off limits can be set, and records between the limits are assigned to clerical review.[17]

Processing a large amount of data: software and algorithms

Even when linking two small datasets there will be a large number of potential record pairs because the number of potential matches is the product of the number of records in each dataset. IDI datasets are typically very large and it would not be feasible to compare records manually in this way. Therefore, probabilistic record linkage in large datasets needs to be automated. In the IDI record linkage is implemented using QualityStage, an IBM programme that is based on the Fellegi-Sunter method. QualityStage performs many different types of transformations and comparisons, for example Soundex, an algorithm that uses phonetic coding to compare records with similar-sounding names.

Even with an automated procedure the process can be slow and for that reason, blocking variables are used to limit the number of comparisons that need to be made. For example, if date of birth is the blocking variable, only record pairs with matching date of birth are compared.

If blocking variables have errors, then record pairs that are potential matches may not be compared, resulting in missed links.[14] To mitigate this risk, multiple passes are used with different combinations of blocking variables.

How is linkage error currently measured in the IDI?

Linkage information that is routinely provided for IDI users (for example, in an IDI refresh report) focuses on two measures: link rates and false positive rates.

The **link rate** is calculated as:

$$\frac{\text{No. of units linked to the IDI spine}}{\text{No. of units in the node dataset}} \text{ x100}$$

The **false positive rate** is:

$$\frac{\text{No. of false positive links}}{\text{No. of units linked to the IDI spine}} \text{ x100}$$

Linkage weights can also be made available to researchers on request and Section 3 describes how they could be used to assess linkage bias.

Section 3
Recommendations for a path forward

Part 3 of this report describes some possible approaches to better understanding linkage bias in the IDI.

This section is not intended as a ‘how to’ guide for researchers. The purpose of this section is to give researchers a sense of some of the potential methods available to investigate linkage bias in IDI. There are specific challenges to understanding linkage bias in IDI, including the large number of links and the lack of researcher access to linking variables or the linkage process (which is necessary for security but makes it difficult for analysts to investigate and adjust for the effects of linkage bias in their analyses, as they cannot compare linked and unlinked data and do not have automatic access to information about linkage quality [18]). Given these challenges, the focus of this section is on methods that will be suitable for use in IDI.

Further work is required to develop linkage bias analysis methods that are tailored to IDI data, and to develop technical support for the IDI community to enable researchers to undertake their own analyses. This work is urgently needed given the increasingly influential role of the IDI in determining policy decisions in Aotearoa New Zealand.

How could we improve our understanding of linkage error and bias in IDI?

The linkage error measures that are currently available to IDI researchers (link rate and false positive rate) do not give researchers many options for better understanding linkage bias because:

- They only cover false links, there are no measures of missed links
- Link rate is strongly influenced by the population overlap between two datasets and less by the quality of the linking
- They cover whole datasets and may not accurately represent the level of error in specific sub-populations
- They are not provided at an individual level meaning that researchers cannot calculate linkage error rates for their own study population

The first step in improving our understanding of linkage bias in the IDI is to produce better measurements of linkage error. Once these are available, researchers can employ existing techniques such as quantitative bias analysis to understand the impact that linkage bias may be having on their results.

This section focuses on three major approaches to better understanding linkage bias in IDI: expanding the measures of linkage error available to researchers (a necessary but not sufficient step for understanding linkage bias); methods to quantify the bias introduced by these errors; and increasing researcher literacy around linkage error and bias.

Expanding estimates of linkage error in IDI

Quantifying false links
Methods to quantify false links include:

- Stats NZ have previously conducted clerical reviews and report the percentage of linked records and false positives. These clerical review datasets are available at individual level and could allow researchers to estimate false link rates for customised populations relevant to their research. There are, however, several limitations to this approach. First, clerical reviews are manual and judgement-based, and based on only a sample of links so it may be difficult to estimate false link rates for small populations. Second, Stats NZ no longer does regular clerical reviews and has moved to a predictive modelling approach using logistic models to estimate the number of false positives based on historical clerical review data. New clerical review datasets may be produced from time to time to check the model, but will no longer be regularly produced.
- Logic checks to identify groups of people with (highly probable) false links indicated by implausible values in the linked datasets. For example, hospital records might indicate that an individual was hospitalised after their date of death. Code for these logic checks, once developed, could be shared between IDI users.

Quantifying missed links
Quantifying missed links is more challenging. Again, logic checks can be used to identify:

- Records that are very likely to be missed links because they had missing or invalid data on key linking variables (for example, name may be missing).
- Records are likely to be missed links if a true link should exist (e.g. someone who accessed a health service as a resident must have an identity in the spine) but no link was made

The suggestions listed here are not exhaustive and researchers may need to develop innovative strategies based on their knowledge of the research topic and of how the data were generated.

Approaches to understanding linkage bias

If researchers have access to individual-level estimates of linkage error, they can begin to understand the extent to which these errors produce bias in their analysis. At a basic level this involves comparing error rates between different groups of individuals and considering how much and in which direction the results might have been influenced by error. Once the distribution of error is measured or estimated, analysis results can be adjusted to produce estimates that are closer to the true value. Together these techniques are referred to as 'quantitative bias analysis' and have been successfully applied to other sources of bias, especially in the epidemiology domain.[19] [20]

Depending on the planned analysis and the potential mechanisms of linkage error, comparisons of error could be made by:[8]

- Ethnicity
- Age
- Sex
- Outcome
- Rurality
- Region
- Deprivation

… and if feasible, by combined strata e.g. ethnicity by age to understand differences in the joint distributions of these variables. In particular, researchers could consider how the linkage procedure might have led to changes in the joint distribution of your exposure and outcome, or adjustment variables in your model. (See the paper by Hochang Choi describing an approach to correcting the distribution of key population indicators to account for linkage error in the IDI spine.[21])

Indirect approaches to assessing linkage bias
When information about error is not directly available, researchers can use indirect approaches to get an idea of where problematic error may be occurring. For example, researchers could use external, unlinked data sources such as Census or survey data and compare the distribution of key characteristics between the external data source and the linked data. Differences in these distributions may indicate error deriving from the linkage process. Thus, a systematic exploration of the differences between linked and unlinked records can give a useful indication of potential sources of bias.

Correcting for linkage bias

The goal of adjusting for linkage bias is to produce an analysis output (e.g. estimate of effect) that is closer to the true value than the unadjusted (biased) result.[22] To do this we must first estimate the amount and pattern of bias (using methods such as those described above) and then adjust the original outputs for this bias.

There are a range of methods for achieving this. A detailed description of these methods is beyond the scope of this paper, but they are mentioned here to give researchers a picture of the end goal of linkage bias research. Methods include:

- Weighting analyses to take linkage error into account. If individual-level information about match status (correct or incorrect) is available then the analysis can be weighted to take error into account, an approach that was successfully used in the NZ Census-Mortality Study.[8] As an example of a simple weighting approach, the hypothetical example presented in Section 1 was calculated using spreadsheets developed by Matthew Fox, Aliza Fink, and Timothy Lash (2007) based on the method for adjusting for selection bias described by Lash et al.;[22 23] the spreadsheet was slightly adapted for the linkage bias example by the authors of this report. These spreadsheets allow researchers to model and explore the potential impact of different biases.
- Using parameters derived from the linkage process. Harron et al. note that linkage error can be handled without requiring any identifiable data if there is access to records with match weights or probabilities. Several proposed methods take this approach, e.g. prior-informed imputation[18] and other imputation-based methods. For an overview of linkage adjustment methods, see Harron et al.[5] However, this area is currently an emerging field that has yet to develop accessible software for non-specialist researchers. One proposed approach is the “adjusted estimating function” described by Kim and Chambers[24] for regression analysis which can include weighting for a specific mechanism of error, if known (as for missing data). This approach uses an audit to develop parameters for the estimate. Another approach using linkage probabilities is reported by Chipperfield and Chambers,[25] who describe a bootstrap estimator method for use with probabilistically-linked data. The method requires access to detailed outputs from the linkage procedure (that are not currently available to IDI researchers).
- If direct adjustment is not possible but record-level linkage weights are available, researchers can gain some indication of the likelihood of differential linkage error using a sensitivity analysis approach, i.e. by repeating an analysis using different cut-offs to understand how sensitive the analysis results are to differing cut-offs; this approach can generate insights about linkage error by examining how the results change as the balance shifts between false positives and false negatives.

Increasing literacy amongst IDI researchers around linkage error and bias

At present there is very little awareness of linkage error and bias issues amongst IDI researchers. This limits their ability to understand the potential impact on their work, and to work together to find solutions.

We suggest that researchers can do the following things to increase their literacy around linkage bias:

- Understand the linkage procedures used to generate their data. This includes consideration of how the data were generated, which variables were used for matching, the likely errors or missingness in key matching variables, which key study variables originated in which dataset/s (and hence which characteristics might be mismeasured if there are false links), and how all of the above factors might differ between populations of importance to the analysis.
- Quantify and describe linkage error in their study population, including (where possible) the proportion of false links and missed links, overall and by variables of interest. These parameters would allow them to weight your estimates to take linkage error into account; you would then be able to generate estimates that are less biased and closer to the true effect.
- Promote linkage error and bias as an important issue for further development and research. This could be done through feedback to Stats NZ, through discussion in their own organisations, and through IDI user conversations such as the IDI Technical Forum and User Forum. Regular acknowledgement of the potential impact of linkage bias in papers and reports using IDI data would also help to increase the visibility of linkage error and bias.

To help increase literacy around linkage error and bias, and to enable research into linkage bias, it is critical for IDI researchers to have access to information about the linkage process. This includes up-to-date descriptions of linkage methods (such as those already available on the Stats NZ website) and also data from the linkage process. At present the latter is not readily available to researchers. Examples of data that could be useful to researchers include:

- Information on the quality (eg plausible values) and completeness of key linking variables e.g. name, sex, and date of birth. Ideally this information should be made available at the individual level so that researchers can see how it varies across their study population.
- Information on false links derived from historical clerical review datasets, and any new clerical review that is undertaken.
- Individual-level information about linkage probabilities and other outcomes of the linking process

References

1. Han Y, Lahiri P. Statistical Analysis with Linked Data. *International Statistical Review*;0(0) doi: doi:10.1111/insr.12295
2. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association* 1969;64(328):1183-210.
3. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. Quality measures in data mining: Springer 2007:127-51.
4. Bohensky M. Bias in data linkage studies. In: Harron K, Goldstein H, Dibben C, eds. Methodological Developments in Data Linkage2016:63-82.
5. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. Chichester, UK: John Wiley & Sons 2016.
6. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 2002;31(6):1246-52. doi: 10.1093/ije/31.6.1246
7. Delgado-Rodríguez M, Llorca J. Bias. *Journal of Epidemiology and Community Health* 2004;58(8):635-41. doi: 10.1136/jech.2003.008466
8. Fawcett J, Blakely T, Atkinson J. Weighting the 81, 86, 91 & 96 census-mortality cohorts to adjust for linkage bias: Department of Public Health, Wellington School of Medicine and Health … 2002.
9. Baldi I, Ponti A, Zanetti R, et al. The impact of record-linkage bias in the Cox model. *Journal of Evaluation in Clinical Practice* 2010;16(1):92-96. doi: doi:10.1111/j.1365-2753.2009.01119.x
10. Neter J, Maynes ES, Ramanathan R. The Effect of Mismatching on the Measurement of Response Error. *Journal of the American Statistical Association* 1965;60(312):1005-27. doi: 10.2307/2283401
11. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology* 1977;105(5):488-95.
12. Statistics New Zealand. Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project. Available from www.stats.govt.nz, 2014.
13. Black A. The IDI prototype spine's creation and coverage. (Statistics New Zealand Working Paper No. 16-03). http://archive.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine.aspx, 2016.
14. Shlomo N. Overview of Data Linkage Methods for Policy Design and Evaluation. In: Crato N, Paruolo P, eds. Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design. Cham: Springer International Publishing 2019:47-65.
15. Statistics New Zealand. Data integration manual: 2nd edition. www.stats.govt.nz, 2013.
16. Doidge JC, Harron K. Demystifying probabilistic linkage: Common myths and misconceptions. *Int J Popul Data Sci* 2018;3(1):410-10. doi: 10.23889/ijpds.v3i1.410
17. O'Sullivan L. Linking, selecting cut-offs, and examining quality in the Integrated Data Infrastructure (IDI). *Statistical Journal of the IAOS* 2015;31(1):41-49.
18. Harron K, Wade A, Gilbert R, et al. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology* 2014;14(1):36. doi: 10.1186/1471-2288-14-36
19. Blakely T, Barendregt JJ, Foster RH, et al. The association of active smoking with multiple cancers: national census-cancer registry cohorts with quantitative bias analysis. *Cancer Causes & Control* 2013;24(6):1243-55. doi: 10.1007/s10552-013-0204-2
20. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *International Journal of Epidemiology* 2014 doi: 10.1093/ije/dyu149

21. Choi H. Adjusting for linkage errors to analyse coverage of the administrative population. *Statistical Journal of the IAOS* 2018(Preprint):1-7.
22. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *International Journal of Epidemiology* 2014;43(6):1969-85. doi: 10.1093/ije/dyu149
23. Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data. New York: Springer 2009.
24. Kim G, Chambers R. Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis* 2012;56(9):2756-70. doi: https://doi.org/10.1016/j.csda.2012.02.026
25. Chipperfield JO, Chambers RL. Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data. 2015;31(3):397. doi: https://doi.org/10.1515/jos-2015-0024

Appendix: Linkage bias worked example

The worked examples in , and Table 3 shows the numbers that the RRs are based on, demonstrating the effect of missed links on distributions and sample size. The method for calculating the bias is shown in the Appendix.

Table 2 were calculated using the method of Lash et al.,[23] and the spreadsheet developed by Fox, Fink and Lash for selection bias was used to model the effect of missed links on a risk ratio (RR) analysis.

The bias spreadsheets developed by these authors are easy to use and multiple scenarios can be calculated based on different bias parameters, as for the two contrasting examples in , and Table 3 shows the numbers that the RRs are based on, demonstrating the effect of missed links on distributions and sample size. The method for calculating the bias is shown in the Appendix.

Table 2. The spreadsheets can be downloaded from https://sites.google.com/site/biasanalysis/. The original selection bias spreadsheet was adapted for the , and Table 3 shows the numbers that the RRs are based on, demonstrating the effect of missed links on distributions and sample size. The method for calculating the bias is shown in the Appendix.

Table 2 worked example and a screenshot is shown in Figure 5. Screenshot of selection bias spreadsheet by Fox, Fink and Lash,[23] adapted to calculate bias from missed links as shown in Table 2.

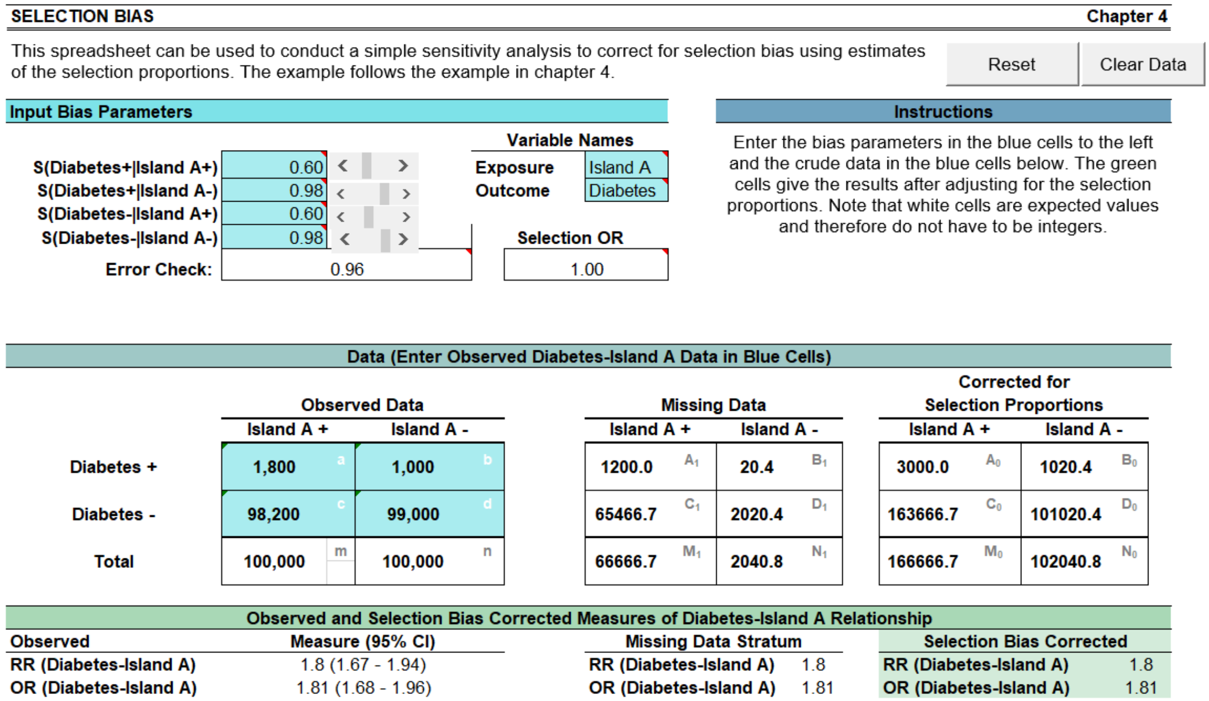
To calculate an adjusted RR using the selection bias spreadsheet, the user first enters bias parameters into the blue cells at the top left. In this example, the bias parameters (i.e. selection proportions) are equal to (1-missing link proportion) for each cell in the joint distribution of exposure and outcome. The user also completes the values in the 2x2 table of observed data.

Values for the corrected 2x2 table are generated by dividing each cell in the observed data 2x2 table by the selection proportion in the corresponding bias parameter cell. For example, if the selection proportion for the Diabetes+/Island A+ individuals is 0.60 (i.e. 40% missing links) and there are 1800 individuals in that cell in the observed data table, the corrected value will be 1800/0.60 = 3000. The corrected values can be seen in the rightmost 2x2 table.

The adjusted RR can then be calculated from the corrected table in the usual way, i.e. risk in exposed/ risk in unexposed, and the result is shown in the 'Selection Bias Corrected' column. A multidimensional analysis can be conducted by systematically altering the bias parameters and recalculating the RR.

The information bias spreadsheet by the same authors can be used in a similar way to adjust for misclassification errors.

Figure 5. Screenshot of selection bias spreadsheet by Fox, Fink and Lash,[23] adapted to calculate bias from missed links as shown in Table 2.



SELECTION BIAS **Chapter 4**

This spreadsheet can be used to conduct a simple sensitivity analysis to correct for selection bias using estimates of the selection proportions. The example follows the example in chapter 4.

Input Bias Parameters

			Variable Names	
S(Diabetes+|Island A+)	0.60		Exposure	Island A
S(Diabetes+|Island A-)	0.98		Outcome	Diabetes
S(Diabetes-|Island A+)	0.60			
S(Diabetes-|Island A-)	0.98		Selection OR	
Error Check:	0.96		1.00	

Instructions

Enter the bias parameters in the blue cells to the left and the crude data in the blue cells below. The green cells give the results after adjusting for the selection proportions. Note that white cells are expected values and therefore do not have to be integers.

Data (Enter Observed Diabetes-Island A Data in Blue Cells)

	Observed Data		Missing Data		Corrected for Selection Proportions	
	Island A +	Island A -	Island A +	Island A -	Island A +	Island A -
Diabetes +	1,800 a	1,000 b	1200.0 A_1	20.4 B_1	3000.0 A_0	1020.4 B_0
Diabetes -	98,200 c	99,000 d	65466.7 C_1	2020.4 D_1	163666.7 C_0	101020.4 D_0
Total	100,000 m	100,000 n	66666.7 M_1	2040.8 N_1	166666.7 M_0	102040.8 N_0

Observed and Selection Bias Corrected Measures of Diabetes-Island A Relationship

Observed	Measure (95% CI)	Missing Data Stratum		Selection Bias Corrected	
RR (Diabetes-Island A)	1.8 (1.67 - 1.94)	RR (Diabetes-Island A)	1.8	RR (Diabetes-Island A)	1.8
OR (Diabetes-Island A)	1.81 (1.68 - 1.96)	OR (Diabetes-Island A)	1.81	OR (Diabetes-Island A)	1.81