

## Estimating rheumatic fever incidence in New Zealand using multiple data sources

J. OLIVER\*, N. PIERSE AND M. G. BAKER

*Department of Public Health, University of Otago, Wellington, New Zealand*

*Received 20 August 2013; Final revision 15 December 2013; Accepted 24 January 2014;  
first published online 6 March 2014*

### SUMMARY

Rheumatic fever (RF) is an important public health problem in New Zealand (NZ). There are three sources of RF surveillance data, all with major limitations that prevent NZ generating accurate epidemiological information. We aimed to estimate the likely RF incidence using multiple surveillance data sources. National RF hospitalization and notification data were obtained, covering the periods 1988–2011 and 1997–2011, respectively. Data were also obtained from four regional registers: Wellington, Waikato, Hawke's Bay and Rotorua. Coded patient identifiers were used to calculate the proportion of individuals who could be matched between datasets. Capture–recapture analyses were used to calculate the likely number of true RF cases for the period 1997–2011. A range of scenarios were used to correct for likely dataset incompleteness. The estimated sensitivity of each data source was calculated. Patients who were male, Māori or Pacific, aged 5–15 years and met the Jones criteria, were most likely to be matched between national datasets. All registers appeared incomplete. An average of 113 new initial cases occurred annually. Sensitivity was estimated at 80% for the hospitalization dataset and 60% for the notification dataset. There is a clear need to develop a high-quality RF surveillance system, such as a national register. Such a system could link important data sources to provide effective, comprehensive national surveillance to support both strategy-focused and control-focused activities, helping reduce the incidence and impact of this disease. It is important to remind clinicians that RF cases do occur outside the well-characterized high-risk group.

**Key words:** Rheumatic fever, New Zealand, population surveillance, statistical models.

### INTRODUCTION

Rheumatic fever (RF) and its sequela rheumatic heart disease (RHD) cause an important burden of disease in New Zealand (NZ), Australia and developing countries [1, 2]. In NZ, RF is now almost exclusively a disease of Māori and Pacific children [3–6]. RF is not usually fatal, but may produce

arthritis, arthralgia, chorea and RHD [7, 8]. Between 42% and 60% of RF cases develop RHD if not treated monthly with intramuscular antibiotics over a period of at least 10 years [8]. RHD is a leading cause of preventable mortality in NZ Māori and Pacific populations. The mean age of death for Māori with RHD is 57·4 years, and 55·4 years for Pacific peoples [3]. By contrast, the NZ population as a whole has a life expectancy of 80·9 years [9]. The economic impact of RHD and RF is substantial with direct hospital costs alone amounting to NZ\$12 million per annum over 2000–2009 [3].

\* Author for correspondence: Ms. J. Oliver, 23A Mein Street, Newtown, Wellington, New Zealand 6242. PO Box 7343, Wellington.  
(Email: olija865@student.otago.ac.nz)

NZ has three major sources of RF surveillance information: national hospitalization data, national notification data, and regional registers. All of these surveillance sources have major limitations that negatively affect their ability to generate national case totals and descriptive epidemiological data [10–13]. These limitations reduce the ability of RF surveillance to support strategy-focused activities (e.g. describing the incidence and distribution of disease to help identify and evaluate interventions) and also control-focused activities (e.g. case management, including delivery of penicillin prophylaxis) to reduce the incidence and impact of this disease [14].

The National Health Index (NHI) number is used to identify patients across all healthcare events. This number is linked to variables such as sex and ethnicity, allowing such variables to be checked, thus improving the internal completeness of datasets. The NHI also links healthcare events over time so re-admissions, recurrences and hospital transfers can be identified with a fairly high degree of certainty.

District health boards are required to electronically submit data on all publicly funded hospitalizations. The Ministry of Health collates these data to form the National Minimum Dataset [15]. Because hospitalization is recommended for all RF cases [16], the hospitalization dataset should be a comprehensive surveillance data source [10]. However, such data contains misdiagnosed and miscoded cases. An analysis of Waikato RF hospitalization data found that this source over-counted cases by 25% [12]. An analysis of Auckland hospitalization data found that RF cases in this region were over-counted by 30% [13].

RF was made a notifiable disease in 1986 [17]. The notification dataset includes the case's demographic and diagnostic details [18]. Severe under-notification of cases has been documented, with incompleteness ranging between 10% and 50% in different regions [10].

The NZ population of 4.4 million people includes 1.4 million concentrated in the Auckland region [19]. National healthcare delivery is managed by 20 district health boards. There are 11 regional registers [20], which aim to track patients and therefore support delivery of prophylactic antibiotics to prevent RF recurrences. Consequently cases are scrutinized carefully before placing them on these databases [21, 22]. The registers' use as a surveillance tool is a by-product of this patient management function [16, 18]. Of these, the Auckland register is the longest running and most complete, covering about 45% of recorded cases in NZ. It is also seen as having the most rigorous case

definition and therefore supplying the most accurate estimates of RF incidence, although data collection is restricted to the Auckland region [13, 20].

Due to the limitations described here, none of the existing surveillance data sources can provide a complete or accurate base for describing the incidence and distribution of RF across NZ [10, 11, 16]. One of the top ten goals of the NZ Government sector is to achieve a two-thirds reduction in the incidence of initial bouts of RF by mid-2017, reaching the rate of 1.4/100 000 [23]. However the actual current incidence is unclear. In order to meet our aim of estimating the likely incidence of RF in NZ, we performed capture–recapture analyses using national hospitalization and notification data covering the 15-year period 1997–2011. The resulting incidence rate estimate may be useful as a baseline for measuring progress towards the Government's RF reduction target. These analyses formed part of a comprehensive surveillance sector review which led to recommendations being made to the Ministry of Health on what potential improvements could be made to the NZ surveillance sector with the aim of supporting optimal RF control and prevention activities [24].

## METHODS

### Data sources

We obtained national RF hospitalization data covering the period 1988–2011 from the Ministry of Health Analytical Services Section. We extracted cases with a principal diagnosis of RF (ICD-9 390–392, ICD-10 I00–I02). Each record included an encrypted NHI number, used to distinguish first admissions from later ones.

We received national RF notification data compiled by Environmental Science and Research Ltd (ESR) for the period 1997–2011 containing encrypted NHIs for most entries.

We also received data from four regional registers covering the following areas: Hawke's Bay, Rotorua, Waikato and Wellington. All registers contained encrypted NHIs and all except the Rotorua register contained additional case demographic information.

We received a selection of summary statistics on the Auckland Register from D. Lennon (Professor of Pediatrics, University of Auckland). Unfortunately the raw data were not made available to us. The odds ratio of matching these groups to notification and hospitalization datasets were calculated from the statistics provided.

### Creating data subsets

The program R v. 2.15.0 was used throughout the analysis [25]. Entries lacking encrypted NHIs were removed. Where we detected repeated entries, all entries for that individual were removed except the first.

Our initial hospitalization dataset contained only NZ residents with a principal diagnosis of RF (assigned between 1997 and 2011), who had never previously been assigned a principal or additional diagnosis of RF or a principal diagnosis of RHD. Hospital transfers were excluded, so only the first hospitalization was recorded for each case. In doing this, we attempted to make the dataset exclusive to new initial RF presentations. This method is in accordance with the approach recently adopted by the Ministry of Health to measure the incidence of new RF cases. The initial notification dataset contained individuals notified between 1997 and 2011, who had no previous episodes of RF recorded in their first notification, and could not be matched to a hospitalization event for RHD prior to RF being assigned as a principal diagnosis. Basic descriptive analyses were performed on the initial hospitalization and initial notification datasets.

### Matching individuals between datasets

An overlap occurred when an encrypted NHI in one dataset could be matched to the same encrypted NHI in the other. Most analyses involved matching data from the initial notification dataset to the initial hospitalization dataset, and vice versa.

In order to discover the number of individuals who could not be matched, we subtracted matched individuals in each dataset from the total number in the dataset. When investigating whether people with certain characteristics were more likely to be matched, we performed stratified analyses using logistic models. These models were used to calculate odds ratios (ORs) and 95% confidence intervals (CIs) of matching subgroups, compared to a reference subgroup. Reference subgroups were usually selected on the basis that they contained the highest number of individuals.

All analyses containing register data were completed using the reduced RF hospitalization dataset and overall notification dataset, rather than the initial datasets. It would be erroneous to use initial datasets for the register analyses as people with RHD, previous RF attacks and additional RF diagnoses may be receiving prophylaxis. The reduced hospitalization

dataset included all first RF hospitalizations between 1997 and 2011 that had been assigned principal or additional diagnoses of RF, regardless of whether individuals had been hospitalized with RHD previously. The overall notification dataset included all notifications between 1997 and 2011.

### Capture–recapture analysis

The total size of a population may be detected by imperfect surveillance systems [26–30]. Two such systems exist in the context of this study; the initial notification and initial hospitalization datasets. Using the equation derived by Chapman [26] the likely true number of initial RF cases may be estimated as:

$$N = \left( \frac{(M + 1) * (C + 1)}{R + 1} \right) - 1,$$

where  $N$  is the number of true initial RF cases (according to the Chapman estimate);  $M$  is the total cases in the initial hospitalization dataset;  $C$  is the total cases in the initial notification dataset; and  $R$  is the number of cases in both the initial hospitalization dataset and the initial notification dataset

Based on the central limit theorem, the standard error of the estimate of  $N$  can be calculated by:

$$\text{var}(N) = \frac{(M + 1)(C + 1)(M - R)(C - R)}{(R + 1)(R + 1)(R + 2)},$$

95% CIs can then be calculated using the equation:

$$95\% \text{ CI} = N \pm (1.96\sqrt{\text{var}(N)}).$$

Dividing the number of initial cases recorded in the dataset in question by  $N$  gives an estimate of its sensitivity [26].

The Chapman estimate assumes that the datasets are independent and that only true cases are recorded. Neither of these assumptions is likely to be satisfied. We have therefore presented five scenarios where we have adjusted the data to reflect the lack of independence and data inaccuracies.

The baseline scenario, scenario 1, assumes the datasets are independent and accurate. This is most likely incorrect [12, 13]. In scenario 2 we have adjusted for dataset inaccuracies by assuming that the notification dataset overstates cases by 2.7% and the hospitalization dataset overstates cases by 25.0% (as found by a Waikato analysis) [12]. The positive predictive value (PPV) for the initial notification dataset was therefore set at 97.3% and the PPV for the initial hospitalization dataset was set at 75.0%. In scenario 3 we have the same assumptions and PPVs as scenario 2,

Table 1. Descriptive characteristics of notification, hospitalization, and selected register datasets

	Initial notification dataset	Initial hospitalization dataset	Auckland RF register*	Hawke's Bay and Waikato RF registers†
No. cases	1179	1953	548	491
Age (years), median (range)	11 (0–55)	11 (1–86)	–	18.0 (4.6–62.0)
<5	1.2%	2.0%	1.8%	0.0%
5–15	81.0%	77.1%	98.2%	39.0%
16–30	15.6%	15.0%	–	46.3%
>30	2.2%	5.9%	–	14.7%
Unknown	0.0%	–	–	–
Sex				
Male	51.8%	56.5%	57.5%	–
Female	39.7%	43.5%	42.5%	–
Unknown	8.5%	0.0%	–	–
Ethnicity				–
Māori	58.0%	49.5%	37.4%	–
Pacific	32.0%	36.6%	57.1%	–
European	4.8%	9.5%	3.3%	–
Other	5.2%	4.4%	2.0%	–
Meet Jones criteria				–
Yes	72.0%	–	–	–
No	4.0%	–	–	–
Unknown	24.0%	–	–	–

RF, Rheumatic fever.

\* All cases entered into the Auckland register over the period 1998–2010

† All cases entered into the Hawke's Bay and Waikato registers over the period 1997–2011.

but have also adjusted for positive dependence between the datasets. Consequently we have reduced the overlap rate to account for hospitals directly notifying cases.

In scenario 4 we used the results of a case audit in the Auckland region, which indicated slightly different PPVs for the hospital and notification datasets (i.e. 67.0% and 78.7%, respectively). The PPV of the overlap section was 88.0% [12, 13]. However the Auckland register is considered the best regional register and thus these findings do not reflect the situation across the whole country. As 47.6% of all RF cases come from Auckland, we present scenario 5. In scenario 5 we have assigned the PPVs calculated using the Auckland analysis [13] a weight of 50.0%. PPVs used in scenarios 2 and 3 have also been weighted at 50.0%.

## RESULTS

### RF case data

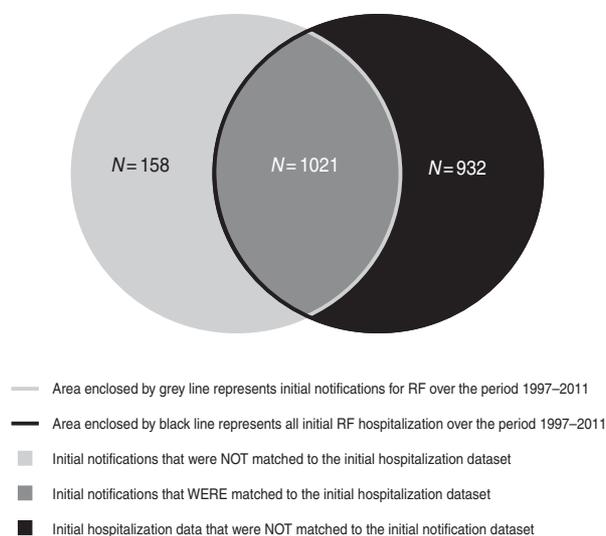
The distribution of initial RF cases arising during 1997–2011 according to major national data sources is shown in Table 1. There were considerably higher

proportions of Māori and people of other ethnicities in the notification dataset than in the initial hospitalization dataset ( $P=0.0003$ ).

Māori were greatly overrepresented in the initial hospitalization dataset, compared to national hospitalization data (i.e. all hospitalizations for 5- to 15-year-olds over 1997–2011, not only those diagnosed with RF or RHD). The proportion of Māori in the initial hospitalization dataset was 2.1 times greater. Pacific peoples were also over-represented; the initial hospitalization dataset contained a 3.6 times greater proportion. There was a slightly higher proportion of males in the initial hospitalization dataset. RF is much more common in socio-economically deprived groups, with 64% of hospitalized cases coming from the most deprived 20% of NZ census area units. By contrast, 30.1% of all hospitalizations were from the two most deprived deciles.

### Matched RF data

Altogether 2111 individuals were identified in the initial notification and initial hospitalization datasets.



**Fig. 1.** Overlap between initial case datasets, 1997–2011.

The majority of individuals recorded as hospitalized with RF (1021/1953 individuals, i.e. 86.6%) could be matched to the notification dataset. Of those on the initial notification dataset, 158 (13.4%) notifications were not matched to the initial hospitalization dataset (Fig. 1).

The likelihood of matching cases between datasets was influenced by a patient's age, sex, ethnicity and whether their disease presentation met the Jones criteria. If older than the reference group (5–15 years), a patient's odds of matching between both datasets declined. The OR for hospitalized patients being matched to the initial notification dataset was 0.46 (95% CI 0.36–0.60) for the 16–30 years age group, and 0.08 (95% CI 0.05–0.15) for those aged >30 years, compared to the reference group. Similarly, the OR of notified patients being matched to the initial hospitalization dataset was 0.30 (95% CI 0.20–0.44) for the 16–30 years age group, and 0.15 (95% CI 0.07–0.35) for those aged >30 years. Notified women were less likely to be matched to the initial hospitalization dataset (OR 0.65, 95% CI 0.46–0.93), and also less likely to be matched from the initial hospitalization dataset to the initial notification dataset (OR 0.83, 95% CI 0.70–0.997). Māori had the highest odds of matching to the initial notification dataset. If the Jones criteria were not met, notified patients were less likely to be matched to the initial hospitalization dataset (OR 0.27, 95% CI 0.25–0.64) (Table 2).

Aucklanders in the initial hospitalization dataset were significantly less likely to be matched to the initial notification dataset, compared to those from Wellington, Waikato, Hawke's Bay, Northland and

Tairāwhiti. Wellington and Tairāwhiti notifications were less likely than all other areas to be matched to the initial hospitalization dataset.

We also matched from a selection of regional registers to the hospitalization and notification datasets. Of those on the regional registers, 51% could be matched to the overall notification dataset and 59% to the reduced hospitalization dataset. Of those on the overall notification dataset, only about 80% were recorded as receiving register-based prophylaxis in the region they were notified as being in. Both the overall notification dataset and reduced hospitalization datasets contained cases that could not be matched to their local register.

### Estimated incidence of RF

Table 3 shows the best estimates of the likely true case numbers according to the scenarios outlined in the Methods section.

Scenario 1 is the baseline with no adjustment; scenario 2 is adjusted for the over-reporting in the datasets (based on a Waikato audit), scenario 3 is adjusted for the over-reporting and positive dependency. Scenario 4 adjusts for both the over-reporting and positive dependency (based on the Auckland data).

Scenario 5 uses the average values from scenarios 3 and 4 to give the most plausible estimate of the true number of cases.

Overall, notification data are consistently less sensitive than hospitalization data. Estimates of its sensitivity vary between 50% and 70%, with the best estimate (based on scenario 5 as +62%). Hospitalization data, according to our estimates, is between 70% and 97% sensitive, with the best estimate being 82% sensitive.

It is likely that there has been a 165% increase in the annual number of initial true RF cases over the period 1997–2011, with between 146 and 158 new cases occurring during 2011.

The annual number of RF notifications is consistently lower than overall hospital diagnoses (i.e. principal and additional diagnoses of RF including those with previous RF and RHD admissions); however, the notification curve roughly follows the curves and troughs seen in this hospitalization curve. This pattern may be indicative of ICD code misclassification. The overall hospital diagnoses curve is consistently higher than the initial RF curves and the discrepancy has widened slightly in recent years. This pattern suggests

Table 2. Descriptive characteristics that influence the chances of matching initial cases, 1997–2011

	Matching from initial hospitalization dataset to initial notification dataset OR (95% CI)	Matching from initial notification dataset to initial hospitalization dataset OR (95% CI)	Matching from Auckland register† to overall notification dataset OR	Matching from Auckland register† to overall hospitalization dataset OR	Matching from Upper North Island registers‡ to overall notification dataset OR (95% CI)	Matching from Upper North Island registers‡ to overall hospitalization dataset OR (95% CI)
Age (years)						
<5	<b>0.36 (0.18–0.70)</b>	<b>&gt;100 (0–∞)</b>	1.7	∞	1.1 (0.3–3.7)	<b>2.94 (1.1–7.5)</b>
5–15*	–	–	1	1	1	1
16–30	<b>0.46 (0.36–0.60)</b>	<b>0.30 (0.20–0.44)</b>	–	–	<b>0.3 (0.3–0.5)</b>	<b>0.6 (0.4–0.9)</b>
≥31	<b>0.08 (0.05–0.15)</b>	<b>0.15 (0.07–0.35)</b>	–	–	<b>0.2 (0.1–0.3)</b>	<b>0.3 (0.1–0.9)</b>
Sex						
Male*	–	–	1	1	1	1
Female	<b>0.83 (0.70–1.0)</b>	<b>0.65 (0.46–0.93)</b>	1.3	1.3	<b>0.7 (0.6–1)</b>	0.8 (0.6–1.1)
Unknown	–	<b>0.55 (0.31–0.98)</b>	–	–	<b>0.2 (0.1–0.4)</b>	–
Ethnicity						
Māori*	–	–	1	1	1	1
Pacific	<b>0.72 (0.59–0.87)</b>	0.93 (0.63–1.35)	1	1	<b>0.2 (0.1–0.2)</b>	<b>0.2 (0.1–0.2)</b>
European	<b>0.29 (0.20–0.41)</b>	0.56 (0.28–1.13)	0.5	∞	1 (0.6–1.7)	<b>0.5 (0.3–0.8)</b>
Other	<b>0.34 (0.21–0.55)</b>	<b>0.32 (0.18–0.58)</b>	1.6	0.4	<b>0.3 (0.1–0.7)</b>	<b>0.3 (0.1–0.6)</b>
Meet Jones criteria						
Yes*	–	–	1	1	1	1
No	–	<b>0.27 (0.25–0.64)</b>	–	–	–	1.48 (0.9–2.5)
Unknown	–	–	–	–	–	<b>0.4 (0.3–0.6)</b>

OR, Odds ratio; CI, confidence interval.

Bold values are statistically significant.

\* Compared to reference groups: age 5–15 years; male; Māori ethnicity; meets the Jones criteria.

† Auckland register data covers cases added to the register between 1998 and 2010.

‡ Upper North Island registers are Hawke's Bay and Waikato registers.

Table 3. Range of likely true initial rheumatic fever case numbers, 1997–2011

Scenario	True cases on initial notification dataset	True cases on initial hospitalization dataset	Overlap	Estimate of total initial true cases ( <i>N</i> )	95% CI
Scenario 1	1179	1953	1021	2252	2217–2287
Scenario 2	1147	1465	1021	1646	1628–1664
Scenario 3	1147	1465	796	2110	2055–2165
Scenario 4	928	1309	898	1351	1342–1360
Scenario 5	1038	1387	847	1698	1668–1728

CI, Confidence interval.

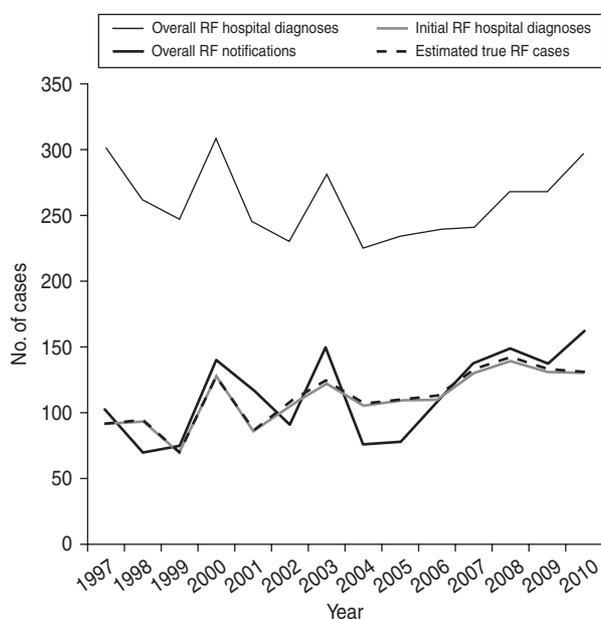


Fig. 2. New Zealand rheumatic fever (RF) hospitalizations and notifications, 1997–2010.

RF cases are increasingly being hospitalized for complications and/or comorbidities. From 2005, overall hospitalizations and notifications broadly increased, both peaking in 2010. Our estimated number of cases very closely follows the number of initial RF hospitalizations (Fig. 2).

## DISCUSSION

Here we present the first national estimate of new RF cases (excluding those with previous RF/RHD) in NZ residents, based on combined data from national notification and hospitalization data and knowledge gained from regional RF registers. The capture–recapture analysis indicates there were, on average, about 113 cases arising per year over the period 1997–2011.

RF hospitalization rates of around 300 events per year, and 100–160 new initial cases annually (Fig. 2), are comparable to that seen in Indigenous Australian populations, which are known to have among the highest RF rates in the world [31]. These rates may also be comparable to true South African incidence rates; however, under-diagnosis is likely to be a major issue in that region [8, 32]. Importantly, this analysis showed that none of the current national RF surveillance systems are complete or accurate. The finding that just over 50% of hospitalized cases could be matched to the notification dataset is consistent with under-notification documented in previous literature [10]. The relatively poor sensitivity of the notification dataset also indicates that under-notification is a serious problem, particularly in the populous Auckland region. As the overall notification and reduced hospitalization datasets contained cases that could not be matched to their local register, this indicates that the registers may be incomplete.

Despite the higher sensitivity, a high rate of misdiagnosis and/or case miscoding affects the hospitalization dataset. Our findings support previous research, by showing that 47.7% of all hospitalized cases assigned RF codes could not be matched to the notification dataset [10]. It is possible that the higher PPV observed with the notification dataset might result from clinicians being more likely to notify severely unwell patients.

Hospitalized cases have diagnostic codes assigned at discharge. This is an imperfect process and does not include mechanisms to revise subsequent diagnoses. Furthermore, there are no ICD codes denoting ‘possible’, ‘probable’ or ‘definite’ RF status. Cases may be notified as ‘probable’, ‘confirmed’ or ‘under investigation’ when notifying. Thus clinical information concerning case status [16] does not translate well to national data sources.

The proportion of people aged >30 years assigned RF codes is almost three times higher than the proportion of notified cases in this age group. As an initial episode of RF is rare in people aged >30 years [6, 26, 33], it seems likely that diagnoses in this group were frequently not confirmed, resulting in few notifications and inaccurate hospitalization data.

It is concerning that the odds of matching female and older cases were lower than the odds of matching younger males between both national datasets. Our findings here may, in part, be attributable to a higher rate of miscoding in these groups. However, there is pre-existing evidence that women and older patients with cardiovascular pathologies are less likely to have their symptoms investigated or receive treatment. Such findings could be at least partially attributed to clinician preconceptions and underuse of active therapies [34]. Our results suggest these findings may also hold relevance to RF in NZ. This suggestion may be supported by the fact that while there are more male RF patients in NZ [27], a significantly greater proportion of women die from RHD [3]. Additionally, previous research has shown that Māori and Pacific women with mechanical or bioprosthetic cardiac valves have a 7- to 8-fold increased risk of dying, compared to European and Asian women [35]. Together, this research indicates that management and treatment outcomes for Māori and Pacific women with RF and RHD are of particular concern.

Calculating an overall total for RF is complicated by the fact that there is clinically detected RF, and RF that is not detected. Patients may be hospitalized with RHD without having been diagnosed first with RF. Five scenarios for the total number of detected RF cases are presented, of which scenario 5 is probably the most robust. This indicates the likely true number of RF cases arising in NZ over the 15-year period 1997–2011 ranges from 1668 to 1728, with an average of 113 cases annually.

Our major limitation is that the capture–recapture analysis is designed to work with datasets which are largely independent, but the notification and hospitalization datasets are most likely non-independent. This assumption is almost always violated in capture–recapture analyses, therefore attempts to reduce the effect of bias are recommended [36]. We have no way of knowing the extent to which the datasets are non-independent or how greatly we violated the assumption of independence in our analyses. The Chapman estimate formula, however, has been designed to reduce the effect of bias when working with non-independent

datasets [26, 37, 38]. We have also adjusted the estimates to take account of the estimated (and large) inaccuracies in these datasets. Capture–recapture analysis has been widely used to enhance infectious disease surveillance [32], notably for calculating likely true case numbers using multiple incomplete data sources and for assessing the magnitude of under-reporting, much in the manner performed here [39–48].

If an individual had more than one NHI number, they may have been counted multiple times in our analyses (although the Ministry of Health indicates that it is now unusual for younger patients in NZ to have multiple NHIs). A number of cases in the notification dataset had not been assigned NHI numbers and could not be linked to any recorded NHI number and consequently were excluded from the analysis. If the datasets are independent then the conclusions will not be affected by excluding such cases. If, however, the datasets are not independent as is plausible, then given the positive dependence between the datasets we are slightly overestimating the total number of cases and underestimating the cases in the older datasets. Our analyses would have also been affected by inaccuracies in ICD coding; however, we attempted to correct for this by using different error scenarios (Table 3).

We assume that well maintained regional registers are more accurate than notification or hospitalization data. Both the Waikato [12] and Auckland [13] RF case audits indicate that these registers are more sensitive and specific than the national datasets. The PPVs used in our scenarios have been calculated using dataset sensitivity estimates based on these findings.

The major implication of this research is the need for NZ to develop a single national RF register. Such a register could take over the functions of the current regional registers and, ideally, the RF notification database. It could build-in additional quality features (such as case definitions used by the Auckland Regional register) and an expanded dataset covering important RF risk factors. This proposed register has potential to save time and effort by requiring clinicians to report cases to a single database instead of two. It would also provide a much more comprehensive base for national RF surveillance through enabling effective case management, to support disease control and by producing strategy-focused surveillance information [11, 24].

Our findings also suggest the need to investigate clinicians' awareness and index of suspicion of RF in individuals who are not male, Māori and aged

5–15 years. The possibility of under-diagnosis and reduced hospitalization of female cases should also be investigated. If verified by subsequent research, intervention is necessary.

### What is already known

RF is an important cause of morbidity and mortality in NZ, especially in Māori and Pacific children. Accurate national RF epidemiological information is not able to be generated as major limitations affect all three sources of national RF surveillance data (i.e. the notification database, hospitalization records, and regional registers). Capture–recapture analyses are useful when calculating the likely true size of a population in situations where that population is detectable only using flawed surveillance systems.

### What this study adds

Over the period 1997–2011, it is likely that 1668–1728 new RF cases occurred in NZ, an average of 113 per year. The hospitalization dataset is our most comprehensive national RF surveillance system and it detects about 82% of these cases, while the notification dataset detects about 62%. Male, Māori patients aged 5–15 years are most likely to be matched between hospitalization and notification datasets, and vice versa. This implies they are more likely than people of other demographics to be both hospitalized and notified with RF. There is a clear need to improve information linkage and develop a more comprehensive national surveillance system, such as a national RF register.

### ACKNOWLEDGEMENTS

We acknowledge our co-advisors: Tomasz Kiedrzyński and Nicholas Jones. We thank Chris Lewis, from the Ministry of Health Analytical Services Section, for encrypting the notification data; and Jane Zhang, who provided the hospitalization and mortality data. We acknowledge John Tagg, Graham Mackereth, John Malcolm, Rob Siebers, and Nicholas Jones, who all reviewed this research. We also thank the staff who provided register data from the following organizations: Rotorua General Practice Group, Capital and Coast DHB, Hawke's Bay DHB and Waikato DHB; and Diana Lennon, who provided Auckland register statistics. This project was supported by a scholarship awarded to J.O. by the Ministry of Health.

### DECLARATION OF INTEREST

None.

### REFERENCES

1. **Ibrahim-Khalil S, et al.** An epidemiological survey of rheumatic fever and rheumatic heart disease in Sahafa Town, Sudan. *Journal of Epidemiology and Community Health* 1992; **46**: 477–479.
2. **Ogunbi O, et al.** An epidemiological study of rheumatic fever and rheumatic heart disease in Lagos. *Journal of Epidemiology and Community Health* 1978; **32**: 68–71.
3. **Milne RJ, et al.** Mortality and hospitalisation costs of rheumatic fever and rheumatic heart disease in New Zealand. *Journal of Paediatrics and Child Health* 2012; **48**: 692–697.
4. **Milne RJ, et al.** Incidence of acute rheumatic fever in New Zealand children and youth. *Journal of Paediatrics and Child Health* 2012; **48**: 685–691.
5. **Vlajinac H, et al.** Influence of socio-economic and other factors on rheumatic fever occurrence. *European Journal of Epidemiology* 1991; **7**: 702–704.
6. **Jaine R, Baker M, Venugopal K.** Epidemiology of acute rheumatic fever in New Zealand 1996–2005. *Journal of Paediatrics and Child Health* 2008; **44**: 564–571.
7. **Jones TD.** The diagnosis of rheumatic fever. *Journal of the American Medical Association* 1944; **126**: 481–484.
8. **Ralph AP, Carapetis JR.** Group A streptococcal diseases and their global burden. *Current Topics in Microbiology and Immunology* 2013; **368**: 1–27.
9. **New Zealand: World Bank.** Life expectancy, 2011 ([http://www.google.co.nz/publicdata/explore?ds=d5bncppjof8f9\\_&met\\_y=sp\\_dyn\\_le00\\_in&hl=en&dl=en&idim=country:NZL:AUS:CHE](http://www.google.co.nz/publicdata/explore?ds=d5bncppjof8f9_&met_y=sp_dyn_le00_in&hl=en&dl=en&idim=country:NZL:AUS:CHE)). Accessed 2 December 2013.
10. **Loring B.** Rheumatic fever in the Bay of Plenty and Lakes District Health Boards review of the evidence and recommendations for action. Toi te Ora District Health, 2008.
11. **Jackson C, Lennon D.** Rheumatic fever register scoping the development of a national web-based rheumatic fever register. Auckland: Ministry of Health, 2009.
12. **Atatoa-Carr P, Bell A, Lennon DR.** Acute rheumatic fever in the Waikato District Health Board region of New Zealand: 1998–2004. *New Zealand Medical Journal* 2008; **121**: 96–105.
13. **Moxon T, et al.** Is a rheumatic fever register the best way to evaluate the government programme to control rheumatic fever? Paper presented at The Paediatric Society of New Zealand 64th Annual Scientific Meeting, Auckland, New Zealand, 2012.
14. **Baker MG, Easther S, Wilson N.** A surveillance sector review applied to infectious diseases at a country level. *BMC Public Health* 2010; **10**: 332.
15. **Craig E, Jackson C, Han DY.** *Monitoring the Health of New Zealand Children and Young People: Indicator Handbook*. Auckland: Paediatric Society of New Zealand, New Zealand Child and Youth Epidemiology Service, 2007.

16. **Atatoa-Carr P, Lennon D, Wilson N.** Rheumatic fever diagnosis, management, and secondary prevention: a New Zealand guideline. New Zealand Rheumatic Fever Guidelines Writing Group. *New Zealand Medical Journal* 2008; **121**: 59–69.
17. **Millen PG.** The infectious disease order 1986. *The New Zealand Gazette*, 21 August 1986.
18. **ESR.** Communicable disease research activities. Porirua: ESR, New Zealand Government, 2011 (<http://www.esr.cri.nz/competencies/Health/Pages/CDresearch.aspx>). Accessed 18 September 2012.
19. **Statistics New Zealand.** Estimated resident population of New Zealand. Wellington: Statistics New Zealand, 2013 ([http://www.stats.govt.nz/tools\\_and\\_services/population\\_clock.aspx](http://www.stats.govt.nz/tools_and_services/population_clock.aspx)). Accessed 1 December 2013.
20. **Yap M.** Review: Rheumatic fever registers as a component of secondary prevention programmes. Wellington: New Zealand Ministry of Health, June 2012.
21. **WHO.** Rheumatic fever and rheumatic heart disease. *WHO Technical Report Series*. Geneva: World Health Organisation, 2004.
22. **WHO.** WHO programme for the prevention of rheumatic fever/rheumatic heart disease in 16 developing countries: report from Phase I (1986–90). WHO Cardiovascular Diseases Unit and principal investigators. *Bulletin of the World Health Organization* 1992; **70**: 213–218.
23. **State Services Commission.** Better public services: supporting vulnerable children. Wellington: State Services Commission, 2012 (<http://www.ssc.govt.nz/bps-supporting-vulnerable-children>). Accessed 15 September 2012.
24. **Oliver J, Pierse N, Baker MG.** Improving rheumatic fever surveillance in New Zealand. A report prepared for the Ministry of Health. Wellington: University of Otago, 2013.
25. **R Foundation.** R: A language and environment for statistical computing [program]. version 2.10.1. Vienna, Austria, 2010.
26. **Chapman D.** Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Press* 1951; **1**: 131–160.
27. **Harvey JN, Craney L, Kelly D.** Estimation of the prevalence of diagnosed diabetes from primary care and secondary care source data: comparison of record linkage with capture-recapture analysis. *Journal of Epidemiology and Community Health* 2002; **56**: 18–23.
28. **Brenner H, Stegmaier C, Ziegler H.** Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany. *Journal of Epidemiology and Community Health* 1995; **49**: 426–430.
29. **Aaron DJ, et al.** Estimating the lesbian population: a capture-recapture approach. *Journal of Epidemiology and Community Health* 2003; **57**: 207–209.
30. **Gemmell I, Millar T, Hay G.** Capture-recapture estimates of problem drug use and the use of simulation based confidence intervals in a stratified analysis. *Journal of Epidemiology and Community Health* 2004; **58**: 758–765.
31. **Parnaby MG, Carapetis JR.** Rheumatic fever in indigenous Australian children. *Journal of Paediatrics and Child Health* 2010; **46**: 527–533.
32. **Nkgudi B, et al.** Notification of rheumatic fever in South Africa – evidence for underreporting by health care professionals and administrators. *South African Medical Journal* 2006; **96**: 206–208.
33. **Taranta A, Markowitz M.** *Rheumatic Fever A Guide to its Recognition, Prevention and Cure*. Lancaster: MTP Press Limited, 1981.
34. **Anderson RD, Pepine CJ.** Gender differences in the treatment for acute myocardial infarction: bias or biology? *Circulation* 2007; **115**: 823–826.
35. **North RA, et al.** Long-term survival and valve-related complications in young women with cardiac valve replacements. *Circulation* 1999; **99**: 2669–2676.
36. **VALID International Ltd.** Notes on using capture-recapture techniques to assess the sensitivity of rapid case-finding methods, version 0.71 ([www.brixtonhealth.com/CRCCaseFinding.pdf](http://www.brixtonhealth.com/CRCCaseFinding.pdf)). VALID International Ltd, July 2006.
37. **Brugal MT, et al.** A small area analysis estimating the prevalence of addiction to opioids in Barcelona, 1993. *Journal of Epidemiology and Community Health* 1999; **53**: 488–494.
38. **Corrao G, et al.** Capture-recapture methods to size alcohol related problems in a population. *Journal of Epidemiology and Community Health* 2000; **54**: 603–610.
39. **Giorgi Rossi P, et al.** Incidence of bacterial meningitis (2001–2005) in Lazio, Italy: the results of an integrated surveillance system. *BMC Infectious Diseases* 2009; **9**: 13.
40. **Lambo JA, et al.** Completeness of reporting and case ascertainment for neonatal tetanus in rural Pakistan. *International Journal of Infectious Diseases* 2011; **15**: e564–568.
41. **Milde-Busch A, et al.** Surveillance for rare infectious diseases: is one passive data source enough for *Haemophilus influenzae*? *European Journal of Public Health* 2008; **18**: 371–375.
42. **Van Hest NAH, et al.** Incidence and completeness of notification of Legionnaires' disease in The Netherlands: covariate capture-recapture analysis acknowledging regional differences. *Epidemiology and Infection* 2008; **136**: 540–550.
43. **Bitar D, et al.** Estimating the burden of mucormycosis infections in France (2005–2007) through a capture-recapture method on laboratory and administrative data. *Revue d'Épidémiologie et de Santé Publique* 2012; **60**: 383–387.
44. **Klein S, Bosman A.** Completeness of malaria notification in the Netherlands 1995–2003 assessed by capture-recapture method. *Eurosurveillance* 2005; **10**: 244–246.
45. **Trijbels-Smeulders M, et al.** Epidemiology of neonatal group B streptococcal disease in the Netherlands before and after introduction of guidelines for prevention. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2007; **92**: F271–276.

46. **Trijbels-Smeulders M, et al.** Epidemiology of neonatal group B streptococcal disease in The Netherlands 1997–98. *Paediatric and Perinatal Epidemiology* 2002; **16**: 334–341.
47. **van Hest NA, Smit F, Verhave JP.** Considerable underreporting of malaria in the Netherlands; a capture-recapture analysis [in Dutch]. *Nederlands Tijdschrift voor Geneeskunde* 2001; **145**: 175–179.
48. **van Hest NAH, et al.** Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiology and Infection* 2008; **136**: 14–22.