*Epidemiol. Infect.* (2010), **138**, 1052–1058. © Cambridge University Press 2009
doi:10.1017/S0950268809991130

# Estimating unreported malaria cases in England: a capture–recapture study

S. J. CATHCART[1]*, J. LAWRENCE[2], A. GRANT[2], D. QUINN[2], C. J. M. WHITTY[3], J. JONES[2], P. L. CHIODINI[3] AND G. FRASER[4]

[1] *North East and North Central London Health Protection Unit, Health Protection Agency, London, UK*
[2] *Centre for Infections, Health Protection Agency, London, UK*
[3] *HPA Malaria Reference Laboratory, London School of Hygiene and Tropical Medicine, London, UK*
[4] *Regional Epidemiology Unit, Health Protection Agency, London, UK*

(*Accepted 14 October 2009; first published online 18 November 2009*)

## SUMMARY

A capture–recapture study was undertaken to estimate the incidence and likely total burden of malaria cases in England. Cases diagnosed by the national Malaria Reference Laboratory (MRL) between July 2003 and December 2004 were matched with cases reported to Hospital Episode Statistics using demographic, geographical, parasitological, and temporal information. A total of 3861 cases were recorded in one or both datasets; the 'unknown population' was estimated as 746 cases (95% CI 677–822) giving a total of 4607 cases (95% CI 4446–4767) over 18 months. Eighty-four percent (95% CI 83–85) of cases were recorded in one or both datasets. Fifty-six percent (95% CI 54–58) of cases were captured by the MRL surveillance system; ascertainment for *Plasmodium falciparum* and London cases was higher at 66% and 62%, respectively. Improving case ascertainment will facilitate effective measures to reduce the burden of this preventable disease in the UK.

**Key words**: Estimating, infectious disease epidemiology, malaria, *Plasmodium*, prevalence of disease, surveillance.

## INTRODUCTION

Around 40% of the world's population are at risk of acquiring malaria, a preventable disease that kills more than one million people every year [1]. In non-endemic countries, such as the UK, malaria occurs in people who have returned from endemic countries. Between 1987 and 2006, around 2000 cases of malaria were reported in the UK each year. Most (96%) of these were in England, where London accounts for 58% of all reported malaria cases [2]. Almost three quarters of cases in the UK result from the potentially fatal *Plasmodium falciparum* infection, the majority having been acquired in West Africa [3]. Travellers visiting friends and relatives in these regions are at particular risk of infection [4, 5].

Malaria surveillance in the UK is undertaken by the national Malaria Reference Laboratory (MRL), part of the Health Protection Agency. Hospital laboratory and clinical staff diagnosing malaria cases complete standard reports accompanying specimen referrals, with supplementary information gathered on cases including travel destination, reason for travel, and chemoprophylaxis. The specificity of the

* Author for correspondence: Dr S. J. Cathcart, North East and North Central London Health Protection Unit, 7th Floor, Holborn Gate, 330 High Holborn, London, WC1V 7PP.
(Email: Simon.Cathcart@hpa.org.uk)

system is high; the great majority of cases reported are laboratory confirmed by the MRL, with most other cases coming from laboratories that are part of the national quality assurance scheme.

The sensitivity of the surveillance system is unknown but is likely to be lower. It is known anecdotally that not all laboratories report cases; in one study in the early 1990s, of 135 enquiries about malaria cases made to the Hospital for Tropical Diseases by UK hospitals, only 79 (59%) of the cases were finally reported to the MRL [6]. Under-reporting of malaria has been documented in other European countries [7, 8] and the USA [9]. Further, although malaria is a notifiable disease, the number of notifications annually is only one third of the number of cases reported to the MRL system [6].

It is likely that the true burden, and cost, of malaria in the UK is underestimated. Significant under-reporting raises the possibility of bias in surveillance information for public health purposes. Capture–recapture (CRC) techniques have been used to estimate the degree of under-reporting of imported malaria in other countries [7, 8]. CRC was developed originally to estimate animal populations that do not easily lend themselves to complete and direct enumeration [8]. It has increasingly been applied to epidemiological studies, using statistical calculations to estimate the total number of incident cases of a disease in a given period. Two or more separate data sources ('lists') are used, which are believed to have incomplete coverage, even when taken together [9, 10].

The purpose of this study was to estimate the true incidence, total burden, and degree of under-reporting, of malaria in England using CRC methodology.

## METHODS

### Data sources

Two data sources were used for the study:

- Malaria cases reported to the MRL surveillance system, with diagnosis dates between 1 April 2003 and 31 March 2005 inclusive.
- Hospital admissions with a diagnosis of malaria, using ICD-10 codes [11], admitted between 1 April 2003 and 31 March 2005 inclusive, were identified from the Hospital Episode Statistics (HES).

HES are the national statistical data warehouse for England that records care provided by National Health Service (NHS) hospitals and NHS patients treated elsewhere. The HES dataset does not include supplementary information about travel history and, unlike MRL data, cases are on the basis of clinician coding on discharge, whether or not a laboratory parasitological confirmation has been obtained.

Datasets were compared for cases resident in England only (HES were not available for the rest of the UK). Duplicate records in the same dataset were identified and removed using date of birth, postcode, and either patient name (MRL data) or a unique HES identification number. Further data cleaning within Microsoft Excel was undertaken to ensure that variable names and coding in each dataset were consistent for matching.

### Matching

Identifying records relating to the same episode of infection in the two datasets was undertaken using a Microsoft Access database designed to match both automatically in the first instance and then by manual methods. This database was adapted from a similar study on tuberculosis [12]. Core identifying variables were assigned a number of points depending on the perceived relative importance of that identifier (e.g. postcode and date of birth had high values compared with sex or parasitological diagnosis). Records were automatically matched if date of birth, sex, diagnosis, and full postcode in each dataset were exactly the same, and date of diagnosis (MRL) and admission date (HES) were within 7 days of each other. A 7-day time period was deemed to be an appropriate length of time according to normal diagnostic procedures for malaria. Deviation from a perfect match (e.g. where information from one record was missing or slightly different) resulted in points being deducted. A score for each potential match was then generated, reflecting the likelihood that two records were the same patient episode of infection.

Pairs of records not matched automatically were reviewed independently by three of the authors. Supplementary information, such as hospital or primary-care trust of treatment, or ethnicity was used to assist the decision on whether a pair of records matched or not; where opinions differed, a majority decision was taken. If a majority decision was not reached, the pairs were assigned as having insufficient data for matching. For the purposes of the initial analysis, these pairs were included as 'unmatched'.

Datasets were compared for a core period of 18 months: 1 July 2003 to 31 December 2004 inclusive;

https://doi.org/10.1017/S0950268809991130 Published online by Cambridge University Press

an additional 3 months at the beginning and end of this time was included in the matching process to account for possible paired records with one date occurring outside the core period. Ideally, to define the core dataset for analysis a date independent of those used for matching is required, e.g. onset date. Due to the nature of the data sources used, an appropriate independent date was not available for this study; therefore matched pairs, where the earlier of the two available dates (MRL date of diagnosis or HES admission date) fell between 1 July 2003 and 31 December 2004 inclusive, were included for analysis. For the unmatched cases, unmatched MRL cases with a diagnosis date, and unmatched HES cases with an admission date between 1 July 2003 and 31 December 2004 were included for analysis.

### Estimate of total cases

The Petersen maximum-likelihood estimator with a two-list CRC model was used to estimate the total number of malaria cases diagnosed in England over an 18-month period. The proportion of cases reported by MRL relative to HES as well as an overall ascertainment in both lists was also estimated. The formula for calculating the unknown population using the Petersen maximum-likelihood estimator is $n_{00} = n_{10}n_{01}/n_{11}$, where $n_{00}$ are the unknown cases, $n_{10}$ are the unmatched cases in the MRL dataset, $n_{01}$ are the unmatched cases in the HES dataset, and $n_{11}$ are the matched cases [13]. A $2 \times 2$ contingency table was used to summarize the presence or absence of cases in one or both lists.

Ninety-five percent confidence intervals were calculated by fitting a loglinear model. This gives an identical point estimate of unknown cases to the Petersen estimator. The loglinear model assumes that the observed and unobserved counts ($n_{11}$, $n_{01}$, $n_{10}$ and $n_{00}$) are subject to Poisson variability, from which confidence intervals can be calculated. Confidence intervals can be estimated either on the log scale or on the count scale. For this study, the log scale was used as this gives asymmetrical intervals that are close to those obtained by other methods. The interval for $n_{00}$ is from a linear combination of estimates on the log scale after fitting the model. The intervals for total cases and ascertainment percentages are from non-linear combinations using the delta method.

Sensitivity analyses were carried out and the estimates were also recalculated after stratification of the cases by sex (92 unmatched records did not specify gender and were assigned on a two males to one female ratio, based on findings from this study and other national data [11]), reporting quarter, species (calculated both by excluding unspecified malaria cases, and also by assigning species based on findings for unmatched HES cases from this study: 77% *P. falciparum*, 18% *P. vivax* and 5% other/mixed species), and either London or non-London residence.

## RESULTS

### Matching

After de-duplication of the datasets, 3267 cases of malaria were reported to the MRL surveillance system between 1 April 2003 and 31 March 2005, and 3699 hospital admissions for malaria were recorded during the same period.

After running the matching queries in Access, 743 pairs were matched automatically and a further 945 pairs were matched following visual inspection and majority decision between reviewers. After removal of records with dates outside the core analysis period, the number of matched pairs recorded between 1 July 2003 and 31 December 2004 inclusive was 1632. The number of unmatched records remaining was 956 (MRL) and 1273 (HES). There were 68 possible pairs of records where a decision could not be made on whether they were a match due to lack of information; these were included in the unmatched totals. The main discrepancy between the MRL and HES datasets was that the reference laboratory system consistently identified more specified malaria infections with *P. falciparum* (2508 compared to 2319), *P. vivax* (464 compared to 371), and mixed infections (464 compared to 371). The HES dataset included 831 cases of unspecified malaria whilst the MRL database had only one such case.

### CRC estimate

The total number of malaria cases diagnosed in England between 1 July 2003 and 31 December 2004 was estimated as 4607 [95% confidence interval (CI) 4446–4767] (Table 1*a*). The number of malaria cases not captured by either MRL or HES systems was estimated to be 746 (95% CI 677–822). Eighty-four percent (95% CI 83–85) of all estimated cases were captured by one or both datasets. The MRL surveillance system captured 56% (95% CI 54–58) of the estimated total cases and 63% (95% CI 61–65) were captured by the HES database.

https://doi.org/10.1017/S0950268809991130 Published online by Cambridge University Press

Table 1*a*. *Estimate of unreported malaria cases in England*

| | | Malaria reference laboratory reported | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hospital Episode Statistics reported | Yes | 1632 | 1273 | 2905 |
| | No | 956 | 746 | 1702 |
| | Total | 2588 | 2019 | 4607 |

Table 1*b*. *Estimate of unreported malaria cases in London*

| | | Malaria reference laboratory reported | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hospital Episode Statistics reported | Yes | 996 | 598 | 1594 |
| | No | 590 | 354 | 944 |
| | Total | 1586 | 952 | 2538 |

Table 1*c*. *Estimate of unreported falciparum malaria cases in England*

| | | Malaria reference laboratory reported | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Hospital Episode Statistics reported | Yes | 1309 | 667 | 1976 |
| | No | 658 | 335 | 993 |
| | Total | 1967 | 1002 | 2969 |

When only cases resident in London were considered, capture was higher for MRL (62%, 95% CI 60–65), with smaller increases in capture for HES and for both datasets together: 63% (95% CI 60–65) and 86% (95% CI 85–88), respectively (Table 1*b*). Similarly, when only *falciparum* cases were considered, capture was higher for MRL: 66% (95% CI 64–68), with smaller increases for capture for HES and for both datasets together: 67% (95% CI 64–69) and 89% (95% CI 88–90), respectively (Table 1*c*).

### Sensitivity analyses

By including records that had insufficient data for matching as 'matched' ($n=68$) the estimate for total cases fell to 4422, giving an ascertainment of 86% for both datasets together, 59% for MRL only and 66% for HES: a small but significant difference to the original estimate. Excluding these 68 records from the analysis altogether altered the estimate for the total cases to 4381, giving an ascertainment of 85% for both datasets together, 58% for the MRL and 65% for HES, not significantly different to the original estimate.

A third of the unmatched cases in the HES dataset (406) were classified as unspecified malaria. These corresponded to cases allocated with the ICD-10 code B54.X ('clinically diagnosed malaria without parasitological confirmation'), and it is possible that some of these were not true malaria cases. After excluding these cases from the matching process, the estimate for total cases decreased to 3963 (95% CI 3822–4104), and overall ascertainment increased significantly: 87% (95% CI 86–88) for both datasets together and 65% (95% CI 63–67) for the MRL surveillance system.

### Stratification

Table 2 shows the cumulative estimates of the unknown population, and newly calculated ascertainment values for MRL and both systems together after stratification of the cases by sex, reporting quarter, species, and region of residence. No significant differences from the original estimates were observed.

## DISCUSSION

The UK malaria surveillance system by the national MRL is one of the most long established and comprehensive surveillance systems in the world [6]. Its longitudinal database, over 20 years, drawing on laboratory reporting nationwide, is highly specific, and has formed a sound basis for guiding public health interventions and public and professional advice [13]. However, like all passive surveillance systems, it is likely to underestimate cases and this study has confirmed that malaria is significantly underreported in England.

Using two-list CRC analysis, the MRL system identified 56% of an estimated total of 4607 malaria cases during an 18-month period. The MRL system was more sensitive for cases resident in London, where most cases in England reside, and for *falciparum* cases (62% and 66%, respectively). These findings are similar to a comparable study in The Netherlands [7]. Higher sensitivities have been observed for

https://doi.org/10.1017/S0950268809991130 Published online by Cambridge University Press

Table 2. *Stratification analysis of estimate to test for heterogeneity in reporting*

| | $n_{11}$ | $n_{01}$ | $n_{10}$ | $n_{00}$ | Total | Both systems (%) | MRL (%) |
|---|---|---|---|---|---|---|---|
| All cases stratified by | | | | | | | |
| Sex | 1632 | 1273 | 956 | 746 | 4607 | 84 | 56 |
| Reporting quarter | 1632 | 1273 | 956 | 757 | 4618 | 84 | 56 |
| Region of residence | 1632 | 1273 | 956 | 739 | 4600 | 84 | 56 |
| Species (PUNS redistributed) | 1632 | 1273 | 956 | 758 | 4619 | 84 | 56 |
| Species (excluding PUNS) | 1631 | 867 | 956 | 517 | 3971 | 87 | 65 |
| *Falciparum* cases only stratified by | | | | | | | |
| Region of residence | 1309 | 667 | 658 | 334 | 2968 | 89 | 66 |

MRL, Malaria Reference Laboratory; PUNS, unspecified malaria.
$n_{11}$, Matched cases; $n_{10}$, unmatched cases in the MRL dataset; $n_{01}$, unmatched cases in the Hospital Episode Statistics dataset; $n_{00}$, unknown cases.

surveillance systems in special situations, such as in certain US states [9] and by the French military [8].

The validity of our total case estimate is dependent on compliance with assumptions of the CRC methodology. First, a significant minority of HES cases may not be true cases. One third of cases in this database were classified as unspecified malaria (PUNS), for which clinical suspicion is sufficient. It is not known what proportion of these cases are true malaria cases. Some may be true cases, even though they may not have had parasitological confirmation performed or noted by the hospital recording the case. However, even under the extreme assumption that none of these cases are true malaria, capture by the MRL system is 65%. MRL data has very high specificity, as a minimum of 80% of cases are parasitologically confirmed by the national reference laboratory, and almost all cases are documented as having been confirmed by the reporting laboratory.

The second assumption made is of independence of the datasets. Two source CRC estimates are unadjusted for dependence between lists, which can result in underestimation of unknown cases if dependence is positive, or overestimation if dependence is negative. Positive list dependence would be expected for malaria reporting in the UK, as there is a tendency for one system (MRL or HES) to alert another to the existence of a malaria case. It is possible therefore that our figure for surveillance coverage by the MRL system is an overestimate. For example, if we suppose that the HES system forwards details of 10% of all malaria cases to MRL, it can be calculated from the observed counts in Table 1*a* that the point estimate of total incidence is 4767 instead of 4607. This reduces the ascertainment by both systems (Table 2) from 84% to 81% and the ascertainment by MRL from 56% to 54%.

Using a third data source, such as clinician notifications (Notification of Infectious Disease Surveillance: NOIDS) would have allowed modelling to correct for any dependence [12]. However, the notification system lacked sufficient identifiers for matching cases in the database readily available at national level. (Further, NOIDS is known to be subject to substantial under-reporting at national level; during the study period, there were 1017 malaria notifications through NOIDS, representing 39% of cases reported to the MRL.) Despite this, use of NOIDS as a third dataset would have been desirable if practicable, as it may include cases unknown to either MRL or HES systems.

The validity of CRC estimates is also dependent on accurate matching of records, which in turn requires correct and complete identifiers. Data in both HES and MRL datasets were missing or incomplete in a significant number of cases and it was often necessary to make value judgements in matching records. Using three independent reviewers helped to reduce observer bias in this, but some unmatched cases may still be the same patients reported in different ways; this would lead to an underestimation of the sensitivity of MRL reporting. Reporting dates from centres were sometimes batched and therefore difficult to match exactly. Using overlapping dates ('soft matching') helped to correct for this problem.

Inclusion of pairs of records where insufficient identifiers were available to match as 'matched'

https://doi.org/10.1017/S0950268809991130 Published online by Cambridge University Press

($n=68$) made a small difference to the total case estimate, but complete exclusion did not make any significant difference to the estimate or ascertainment figures. These cases therefore remained in the dataset as unmatched, for subsequent analyses. Stratification made little difference to the overall estimate, suggesting that there is little heterogeneity in malaria case ascertainment, and that all cases were as likely as each other to be reported to both surveillance systems.

Malaria surveillance in England is well established and its coverage is comparable with other international surveillance systems [7–9]. Our study suggests that a third or more cases diagnosed in England may not be reported to the official surveillance system. It is not known whether, and in what respects, these cases differ from those routinely captured. It is possible that these 'unknown' cases are not admitted to hospital nor laboratory diagnosed: they may be treated as outpatients, perhaps self-treat with drugs acquired overseas or on the black market, or even recover spontaneously.

Malaria is a notifiable disease, and steps should be taken to increase reporting by doctors, including reporting of suspected cases, although evidence relating to how this may be achieved in practice is sparse. Proposed changes to the Health Protection (Notification) Regulations will place an obligation on diagnostic laboratories to notify the proper officer (Health Protection Agency) of a confirmed case of malaria. There is a suggestion that mandatory laboratory notification improves reporting rates [9].

Reasonably complete surveillance is desirable for monitoring incidence trends, identifying groups at risk, implementing public health and health service interventions and, by analysis of treatment and prophylaxis failures, obtaining accurate data on drug resistance patterns to update chemoprophylaxis policy. The accuracy of surveillance data, and its repeatability, allows the UK to track the success or otherwise of preventive measures in travellers. By identifying spikes in malaria incidence it also can act as an early warning system for outbreaks, which are likely to become more unpredictable as malaria control moves to elimination in some geographical areas. Finally, the first imported case of *Plasmodium knowlesi*, now regarded as the fifth species of human malaria parasites, was first diagnosed in the MRL, reflecting its importance as a national reference facility [14].

## ACKNOWLEDGEMENTS

We acknowledge Northgate Information Solutions for providing data from the Hospital Episode Statistics database. Thanks are due to Marie Blaze, Valerie Smith (MRL) and Margaret Armstrong (Hospital for Tropical Diseases) for providing data and assisting with queries relating to the data. P.L.C. is supported by the UCL Hospitals Comprehensive Biomedical Research Centre Infection Theme.

## DECLARATION OF INTEREST

None.

## REFERENCES

1. **World Health Organization.** Malaria Factsheet No. 94. Updated May 2007. Published online http://www.who.int/mediacentre/factsheets/fs094/en/index.html. Accessed 16 April 2009.
2. **Smith AD, *et al.*** Imported malaria and high risk groups: observational study using UK surveillance data 1987–2006. *British Medical Journal* 2008; **337**: a120.
3. **Health Protection Agency.** Malaria imported into the United Kingdom in 2007: Implications for those advising travellers. *Health Protection Report* 2008; **2** (17). (http://www.hpa.org.uk/hpr/archives/2008/hpr1708.pdf).
4. **Health Protection Agency.** Foreign travel-associated illness – a focus on those visiting friends and relatives, 2008 report. London: Health Protection Agency, 2008 (http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1231419800356).
5. **Leder K, *et al.* (for the GeoSentinel Surveillance Network)**. Illness in travelers visiting friends and relatives: a review of the GeoSentinel Surveillance Network. *Communicable Infectious Disease* 2006; **43**: 1185–1193.
6. **Davidson RN, *et al.*** Under-reporting of malaria, a notifiable disease, in Britain. *Journal of Infection* 1993; **26**: 348–349.
7. **van Hest NAH, Smit F, Verhave JP.** Underreporting of malaria incidence in the Netherlands: results from a capture-recapture study. *Epidemiology and Infection* 2002; **129**: 371–377.
8. **Millar T, *et al.*** Glossary of terms relating to capture-recapture methods. *Journal of Epidemiology and Community Health* 2008; **62**: 677–681.
9. **Hook EB, Regal RR.** Capture–recapture methods in epidemiology: methods and limitations. *Epidemiology Review* 1995; **17**: 243–264.
10. **Domingo-Salvany A.** Estimating the prevalence of drug use using the capture-recapture method: an overview. In: *Estimating the Prevalence of Problem Drug Use in Europe*. EMCDDA Scientific Monograph Series: Luxembourg, 1997.

11. **World Health Organization.** International Classification of Diseases (ICD) (http://www.who.int/classifications/icd/en/). Accessed 16 April 2009.
12. **Van Hest NA, *et al.*** Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999–2002. *Epidemiology and Infection* 2008; **136**: 1606–1616.
13. **Chiodini P, *et al.*** Guidelines for malaria prevention in travellers from the United Kingdom. London: Health Protection Agency; 2007 (http://www.hpa.org.uk).
14. **Health Protection Agency.** Imported malaria cases and deaths, United Kingdom: 1989–2008. Data from the HPA Malaria Reference Laboratory, July 2009 (http://www.hpa.org.uk/webw/HPAweb&HPAwebStandard/HPAweb_C/1195733773780?p=1191942128262).