
C. Jennison, I. Roddick, A. Deas, L. Emmett, S. Bracebridge

HPA East of England Regional Epidemiology Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK

Address correspondence to C. Jennison, E-mail: charles.jennison@hpa.org.uk

**ABSTRACT**

**Background** Widespread testing for chlamydia is expected to result in a reduction in prevalence. In 2008, coverage indicators introduced by the Department of Health (DH) required collection and submission of all tests performed outside of genitourinary medicine clinics. No mechanism existed to collect community-based tests conducted outside of the National Chlamydia Screening Programme. The Health Protection Agency Regional Epidemiology Unit in the East of England (EoE) set up a new system to routinely collect and submit these tests on behalf of the regional Primary Care Organizations (PCOs).

**Methods** Testing data were requested from all laboratories commissioned to undertake chlamydia testing by EoE PCOs. Data were imported into a bespoke Structured Query Language server database and automated data processing routines were run. Data fulfilling national criteria were submitted for inclusion in the DH indicators.

**Results** High-quality data were submitted to set deadlines with minimum impact on laboratories. Completeness of data variables varied by laboratory and by variable type. After complex data processing, 96% of laboratory reported tests in the 15–24 year age range were eligible for submission.

**Conclusions** This centralized method of data collection provides high-quality data, allowing for further analysis, which can be used to inform improvements in health care. These methods could be transferred to any of the hundreds of organisms for which similar laboratory data exist.

**Keywords** *Chlamydia trachomatis*, genitourinary diseases, health services, laboratory, public health, surveillance

**Introduction**

*Chlamydia trachomatis* is the most commonly diagnosed sexually transmitted infection in genitourinary medicine (GUM) clinics in England, with the majority of infections in young adults. Left untreated infection can have serious consequences, such as ectopic pregnancy and infertility in women and epididymitis and reactive arthritis in men. However, because the majority of infections are asymptomatic, the identification of infected individuals is a major challenge. In 2003, implementation of the National Chlamydia Screening Programme (NCSP) in England began, with the aim to identify, test and treat sexually active young adults who are infected. Chlamydia screening in the community was introduced as a ‘Tier 2 vital signs indicator’ by the Department of Health (DH) in England, with targets of 17% of the eligible population (15–24 years) to be tested during 2008/09, 25% during 2009/10 and 35% during 2010/11.

The East of England (EoE) has a population of ~5.6 million and is one of the nine healthcare regions in England. The region is broken down into a number (14 until the beginning of the 2010 fiscal year) of Primary Care Organizations (PCOs), which administrate the healthcare...
services within them. In the EoE these contain populations from 164 000 to 756 000. Smaller geographic areas within a PCO include local authorities (LA), and smaller still, lower super output areas (LSOAs), which in the EoE have average populations of \( \approx 1600 \).

Prior to the introduction of the vital signs indicator, the local delivery plan data monitoring line (2007/2008) asked all PCOs to test at least 15% of men and women aged 15–24 years, and included tests reported through the NCSP and the Boots Pathfinder Project.\(^6\) However, the chlamydia testing activity survey (CTAS) conducted in 2007 showed that a significant proportion of tests (30% nationally and 23% in the EoE) occurred in the community, that is, testing outside of GUM clinics occurring primarily in general practices (GPs), hospitals, contraception and sexual health services, abortion clinics, prisons and military settings, etc., not commissioned by the NCSP.\(^7\)

The new indicator thus required collection and submission of all tests performed outside of GUM clinics. This presented a challenge to all PCOs in England since community testing data not commissioned by the NCSP was not currently being collected, and no national solution was provided. Existing national laboratory surveillance undertaken by the Health Protection Agency (HPA) only collects laboratory-confirmed positive results and therefore, HPA methods for collecting chlamydia testing data at the beginning of the 2008 fiscal year could only meet GUM data requirements.\(^8\) NCSP data collection relies on programme areas collecting data from the laboratories they commission and submitting it centrally themselves. Technical guidance relating to the chlamydia vital signs indicator suggested obtaining this additional source of data directly from laboratories, and the previous CTAS and Avon surveillance system for sexually transmitted infections (ASSIST) studies had demonstrated that this was feasible.\(^5,7,9\) Indeed, this method of data collection and handling is not just limited to sexual health and could be used for any of the hundreds of organisms for which similar laboratory data exist, such as tuberculosis or healthcare-associated infections.

The majority of PCOs in England chose to liaise on an individual/small group basis directly with laboratories to gather this non-NCSP non-GUM information in an aggregated format. However, this approach may suffer from several drawbacks.

(i) A single laboratory might test samples for more than one PCO resulting in contacts from several PCOs to that laboratory for data.

(ii) National requirements outlined specific data processing requirements, which may require additional resources at each laboratory.

(iii) PCOs may request testing data just for their own residents, with tests pertaining to other PCOs potentially being lost.

(iv) Laboratories may process their data differently, potentially making regional comparisons less valid.

This paper describes the development and implementation of a new regional laboratory-based surveillance system for genital chlamydial infection in the EoE.

**Methods**

Minimum data collection, processing and reporting requirements were set out by the NCSP national team (commissioned by the DH to collect this additional source of data).\(^5\) These requirements stipulated that data should be reported on a quarterly basis and only include:

- tests for chlamydia in those aged 15–24 years,
- specimen types that were either genital, urinary or rectal specimens,
- results that were positive or negative,
- tests that were not commissioned by the NCSP and did not come from GUM clinics,
- tests with specimen dates after the 1 April 2008,
- geographic information including patient postcode, registered GP postcode, test service postcode.

Using these criteria, the HPA EoE regional epidemiology unit (REU) developed a data specification for laboratories, which included the list of required data fields and a schedule for data returns. A database was developed to securely store this information, and a series of complex automated data processing procedures were written. These steps are examined in more detail below.

**Data collection**

Laboratories commissioned by EoE PCOs to undertake chlamydia testing were identified through liaison with individual PCOs. A total of 16 laboratories in the EoE were identified as well as 4 (2 private) out of region laboratories.

Once identified, the commissioned laboratories were sent the data specification and contacted to discuss routine reporting of this information. To minimize the impact of data processing at the laboratory end, requested data included all chlamydia tests on individuals of all ages. Disaggregate data fields included: date of birth, sex, specimen date, test result, postcode of residence, registered GP postcode, postcode of specimen service,
specimen service code, specimen number, hospital number and patient name.

One laboratory in the EoE was unable to provide disaggregate data and was therefore asked to supply aggregated data which would fulfil the minimum national data requirements. The two out of region private laboratories were unable to provide either aggregated or disaggregated returns. Data were therefore obtained directly from the two commissioned provider clinics; several Marie Stopes clinics, via their data team in London, and the Brook clinic, Luton. The term data providers will be used in the rest of the paper to denote both the laboratories and the two clinics.

Data returns were transferred to the REU via Voltage SecureMail® (a commercial high-level encryption e-mail system).

**Database development and data processing**

A bespoke Structured Query Language (SQL) Server database was built with import tables for each data provider, a core data table and code translation tables. SQL Server database software was chosen because this data set was going to require computationally intensive data processing and a high-level of data security.

Processing of the data received from the data providers involved a number of stages, which are shown in Fig. 1 and described briefly below.

**Stage 1.** Before the data were imported into the SQL Server import tables, they were briefly checked in Microsoft Excel for data completeness and accuracy. They were then formatted and imported into their respective SQL Server import tables.

**Stage 2.** The data from each of the SQL Server import tables were moved into the core data table using SQL routines in SQL Server Integration Services. These routines mapped the supplied data fields to each corresponding field in the core data table, resulting in a consistent structure for all the supplied data from across the region.

Each data provider uses their own code set to describe their data. To convert uniquely coded data into a standardized format, a number of translation tables, relating locally supplied codes to a regionally consistent code set were constructed. These were continually updated as new codes were received. These translation tables were used to translate locally supplied codes for sex, specimen type and test result to standardized codes. A similar table was constructed to link local codes for test service locations to their postcodes. Additional information such as the type of service (GP, pharmacy, etc.) was later included in this table to facilitate analysis by service type.

Using SQL queries, duplicate records from the same data provider, falling within the nationally assigned period of 1 week, were flagged in an additional data field. Duplicates were identified using patient name, date of birth, NHS number, hospital number and specimen collection date. Postcodes were parsed to ensure they were in a consistent 7-character format allowing them to be easily referenced against the Office of National Statistics Gridlink® postcode file to assign PCO, LA and LSOA.

Nationally devised algorithms outlined how the allocation of geography should take place. In accordance with this, the residential postcode was used whenever possible to assign patient PCO for each record. When this was unavailable, that of the patient’s registered GP was used, and if this was unavailable the postcode of the test service location was used. Test service location codes were often supplied instead of test service postcodes but a location look up table was populated with their associated postcodes.

![Fig. 1 The collection and importing of data.](image-url)
Tests were classified into residential PCO/LA or attributed PCO/LA (registered GP or test service) categories depending on which geographical data were available. Where no geographical information was supplied, tests were excluded as they could not be reliably assigned to a PCO/LA (PCOs and LAs in the EoE are all co-terminous).

**Data protection**
The data files received from the data providers were stored in a restricted access folder accessible only to people directly involved with the work. The SQL Server database was protected by personal passwords with individual permissions limited to each user as required. Patient identifiable information was collected to enable deduplication within the data set as described above. Patient name, date of birth, residential postcode, NHS number and hospital number were deleted once the data had been validated by the national data management team at HPA centre for infections (CfI), and in accordance with the HPA record retention schedule and Caldicott guidance.

**Quality assurance**
The SQL Server database was queried to identify any inconsistencies and inaccuracies according to a pre-defined checklist. The numbers of tests were compared with previous returns, and those with missing geographical information were scrutinized and compared across quarters to establish any potential problems. Each quarter, missing specimen types, result types and test service codes were queried with the data providers and updated in the relevant translation tables, enabling the accuracy of the data set to be continually improved.

**Data submission**
Approximately 3 working days before each data submission to HPA CfI, individual reports were provided to each of the 14 PCOs in the EoE for their information and to check for any anomalies. Returns were subsequently submitted to HPA CfI in their specified format and to their specified deadlines.

The extract of data used in the analyses presented in this paper was taken on 10 June 2010. Consequently, the information presented may differ slightly from that actually submitted to HPA CfI for inclusion in the chlamydia vital signs indicator.

**Data quality and completeness**
Analysis of the data set for completeness of individual data fields by data provider was conducted to assess data quality. This particular analysis was performed solely on data from laboratories within the EoE. As described earlier, a full extract of all chlamydia testing data was requested. As well as minimizing the impact on data providers, this also allowed all cleaning and processing of the data to be conducted at the REU. Data received from data providers outside of the EoE, one laboratory in the EoE region and the Luton Brook and Marie Stopes clinics were either received in an aggregate form or underwent varying degrees of pre-processing (e.g. to identify EoE tests) before transfer to the REU and therefore were not amenable to analysis of data quality.

The nationally devised algorithm to assign patient geography uses registered GP or test service where residential postcode is missing. This assumes that these two alternatives are good predictors of a patient’s PCO. This hypothesis was tested using records which had both residential PCO and attributed PCO (based on registered GP or test service location) data fields complete.

**Results**
Data were reported by 20 data providers. This included 18 hospital laboratories (2 external to the EoE region) and 2 clinics, Marie Stopes and Luton Brook.

Between 1 April 2008 and 31 March 2010, the REU received a total of 229,242 non-NCSP non-GUM tests. Of these, 225,566 (98.4%) were in disaggregate form and 3,676 (1.6%) were aggregated. Luton Brook clinic initially provided aggregate data, but provided disaggregate data midway through 2009/10 after IT improvements.

The completeness of the data fields requested from laboratories is shown in Table 1. This data excludes the two laboratories external to the EoE region, one laboratory in the EoE region and the Luton Brook and Marie Stopes clinics for the reasons described above. Therefore, this leaves 86% (197,486/229,242) of tests. Completeness of specimen date was 100%, with date of birth, sex and test result fields all close to 100% complete. The postcode of the test service was only received from four laboratories, however after using supplied test service location codes to translate to postcodes, mean completeness rose from 54 to 97%. Two laboratories were unable to provide residential postcode and there was considerable variation in completeness of this field between laboratories when this was supplied. Registered GP postcode was only provided as a separate data field by two laboratories.

The data described above were processed according to the national minimum data collection and processing requirements. Of the 197,486 disaggregate tests on all age-groups provided by the 15 laboratories in the EoE, 24% (47,338) met the criteria for submission (Fig. 2).
Approximately three-quarters of the tests received were outside of the required age range (15–24 years) and therefore excluded. Another 1855 tests were either duplicates, had insufficient geographical information to allocate either residential or attributed PCO, were not genital, urinary or rectal specimens and/or had test results other than positive or negative.

**PCO allocation**

Of the 47 338 eligible tests, 2% (951) were attributed to PCOs outside of the EoE. In contrast only 645 tests were attributed to the EoE PCOs from other regions. This represents 1.4% of non-NCSP non-GUM testing for the EoE (197 tests in 2008/09 and 448 in 2009/10).

The chlamydia vital signs indicator measures the coverage of chlamydia testing in 15–24 year olds resident in each PCO. If an individual’s test is conducted in a service location outside the PCO boundary, and if the residential postcode of that individual is not known, the test will then be allocated to the PCO of the test service provider. This may lead to an underestimate of testing coverage for the true PCO of residence.

Of the non-NCSP non-GUM tests reported from laboratories in those aged 15–24 years, 40% (19 456) had both valid residential and test service PCOs for the 2008/09 and 2009/10 fiscal years. Of these, 6.9% (n = 1342) did not have matching residential and test service PCOs.

**Discussion**

**Main finding of this study**

The EoE data collection methodology has been demonstrated to be a feasible and effective way to meet national reporting requirements regarding the chlamydia vital signs indicator.

As predicted, data providers often performed tests for several PCOs. Using this system, high-quality data were submitted to set deadlines with impact on data providers kept to a minimum by only requiring one single contact with each data provider during each data reporting period. Correspondence with data providers has shown that they are satisfied with the current data collection routine, and logic would suggest that this method of data capture is more efficient, since much of the data handling is automated at the REU rather than being done individually by the data providers.

The results on data completeness (Table 1) show that there were considerable differences between laboratories and it is therefore reasonable to assume that processing of these data for submission to CfI would also vary. Data processing at the REU was standardized, conforming to national guidance, and therefore data were as consistent as was possible and allowed for reliable comparisons to be made between PCOs in the region.

All non-NCSP non-GUM chlamydia testing performed in the 2008/09 and 2009/10 fiscal years within the EoE was reported to the REU. Contrary to what commuting patterns would suggest (2001 Census, CommuterView. Crown copyright 2008. Crown copyright material is reproduced with the permission of the Controller of HMSO and the Queen’s Printer for Scotland),

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Completeness of disaggregate chlamydia testing data provided by laboratories situated in the EoE, April 2008–March 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of EoE laboratories that supplied the specified data field (n = 15)</td>
<td>Mean field completeness when provided (%)</td>
</tr>
<tr>
<td>Residential postcode</td>
<td>13</td>
</tr>
<tr>
<td>Registered General practice postcode</td>
<td>2</td>
</tr>
<tr>
<td>Test service postcode received directly</td>
<td>4</td>
</tr>
<tr>
<td>Test service postcode available after data processing</td>
<td>15</td>
</tr>
<tr>
<td>Date of birth</td>
<td>15</td>
</tr>
<tr>
<td>Specimen date</td>
<td>15</td>
</tr>
<tr>
<td>Specimen type clearly stated</td>
<td>14</td>
</tr>
<tr>
<td>Sex (male or female)</td>
<td>15</td>
</tr>
<tr>
<td>Result</td>
<td>15</td>
</tr>
</tbody>
</table>

*aExcludes one laboratory in the EoE region that only supplied aggregate data.

bThis is based on test records that had a valid PCO for the test service location after having been processed in the database.

*cWhere specimen type provided was translated into the, ‘MISSING’ or ‘OTHER’ categories, they were classified as being ‘not clearly stated’.
communication with the data management team at CfI suggest the former.

What is already known on this topic?
Surveillance of infectious diseases using data collected directly from laboratories has been undertaken by the HPA for many years (via the CoSurv system). This data source is an invaluable public health resource allowing for the monitoring of disease patterns, trends and the identification of potential outbreaks and new emerging threats. However, CoSurv only collects laboratory-confirmed positive results and thus was unable to provide the data required for the chlamydia vital signs indicator.

The CTAS and ASSIST studies showed that it was feasible to collect chlamydia testing data from all test sources (NCSP, GUM, and non-NCSP non-GUM), although the return rate from laboratories varied considerably between regions from 17 to 83%. In the CTAS survey, the EoE had the highest return rate from laboratories, which provided an encouraging basis for the development of the surveillance system.

The ASSIST study, using a similar methodology for chlamydia data collection, was published in 2007. One of the key messages from the ASSIST study was that the system developed could be a useful model for surveillance if it could become timely and sustainable. In the EoE, a routine reporting system with data collected every quarter (not yearly as in the ASSIST study) was set up and used to report to set deadlines for 2 years (this work is continuing into a third year). In addition, some data providers in the region chose to report on a monthly basis, rather than quarterly, meaning there is scope for even more timely data collection using the methodology described in this paper. The current study has also demonstrated the feasibility of the approach on a much larger scale. All laboratories in the EoE were included covering a population of 5.6 million. In contrast the ASSIST study covered just 984 000 and involved far fewer laboratories.

What this study adds
There are several studies2,14–21 that have attained good chlamydia testing data sets; however the methodology used in the EoE region is unique in that it is the first to collect these data in a consistent and routine manner from data providers.

The EoE system collected disaggregate data, including both residential and test service postcodes. It was therefore possible to assign PCO for the vast majority (96%) of tests. In regions where data are collected from laboratories in aggregate form, it is difficult to validate the process by which PCO has been assigned, and therefore difficult to estimate the degree to which tests have not been attributed to the actual PCO of residence.

Where both residential and test service postcodes were available for the same test, approximately 7% of these were undertaken by a test service, which was not allocated in the same PCO as that of the patient’s residence. Research suggests that populations generally commute to urban centres for work and other reasons.11–13 Misallocation of tests could therefore lead to underestimation of testing coverage in some

Fig. 2 Reasons for exclusion in disaggregate chlamydia testing data provided by laboratories situated in the EoE, April 2008–March 2010.

aOne test may fulfil several exclusion criteria. The total excluded at this stage was 1855 tests.
PCOs if a test undertaken in an urban centre is not attributed back to the PCO of residence. This may be a particular issue for the EoE region which borders London.

A complete non-NCSP non-GUM data set has several benefits. The data set can be queried to answer questions from commissioning PCOs, which may be difficult or impossible to answer in other regions, such as which venues are performing testing, and test numbers broken down by month. The data set can be combined with existing national data sets that collect chlamydia testing data from other sources (NCSP and GUM clinic activity data set) to produce the most complete data set in England providing rich opportunities for epidemiological analyses. The first stage of this has already been completed whereby the non-NCSP non-GUM and NCSP data sets were combined for the region. Using these data, reports detailing all chlamydia testing in the community were produced for sexual health commissioners and included local coverage maps, test service locations and coverage by age, sex and quarter. Initial feedback from PCOs has been extremely positive with some using these reports to guide their chlamydia testing strategies. A full evaluation of this work is currently underway.

The EoE region has provided support to both the West Midlands and London regions who have adopted the methodology developed. Personal communication with PCO commissioners, and the renewal of the data collection contract for the 2009/10 and 2010/11 fiscal years, has demonstrated that this service was both desirable and effective. Nationally the HPA, informed by the work undertaken by the EoE, is now working with a number of laboratories to develop laboratory reporting of chlamydia testing data, known as the chlamydia testing activity data set (CTAD).

Limitations of this study

Where there is low completion of residential postcode data this will have a small impact on the quality of the data when assigning PCOs. Our results showed that where both residential and test service postcodes were available 93% matched. However, the level of error is likely to increase when analysing the data using smaller geographical units.

The non-NCSP non-GUM tests are a small proportion of the total chlamydia tests undertaken which can be included in the vital signs target (in the 2009 fiscal year it was just under 20%). Completeness of residential information in the NCSP data over the 2 years is >99% complete, and therefore when combined, the impact of missing residential information for non-NCSP non-GUM tests will be reduced.

At present, there is no accurate way of determining how much additional testing data are captured with this centralized method. It is also unknown how many work hours are saved through centralized data collection. Without a disaggregate data set from neighbouring regions, London in particular, it will be very difficult to quantify the test misallocation previously mentioned.

Conclusions

This centralized method of data collection provides high-quality data, allowing for further analysis, which has been used to inform improvements in health care. These methods would be transferable across a variety of infections.

Acknowledgements

All data providers including the laboratories, Marie Stopes International and the Luton Brook clinic for providing the data extracts. The NCSP Information Management team for providing the NCSP data extract.

Funding

This work was supported by the PCOs in the EoE.

References


