Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement

Nigel Field, Ted Cohen, Marc J Struelens, Daniel Palm, Barry Cookson, Judith R Glynn, Valentina Gallo, Mary Ramsay, Pam Sonnenberg, Duncan MacCannell, Andre Charlett, Matthias Egger, Jonathan Green, Paolo Vineis, Ibrahim Abubakar

Molecular data are now widely used in epidemiological studies to investigate the transmission, distribution, biology, and diversity of pathogens. Our objective was to establish recommendations to support good scientifi c reporting of molecular epidemiological studies to encourage authors to consider specifi c threats to valid inference. The statement Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID) builds upon the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) initiative. The STROME-ID statement was developed by a working group of epidemiologists, statisticians, bioinformaticians, virologists, and microbiologists with expertise in control of infection and communicable diseases. The statement focuses on issues relating to the reporting of epidemiological studies of infectious diseases using molecular data that were not addressed by STROBE. STROME-ID addresses terminology, measures of genetic diversity within pathogen populations, laboratory methods, sample collection, use of molecular markers, molecular clocks, timeframe, multiple-strain infections, non-independence of infectious-disease data, missing data, ascertainment bias, consistency between molecular and epidemiological data, and ethical considerations with respect to infectious-disease research. In total, 20 items were added to the 22 item STROBE checklist. When used, the STROME-ID recommendations should advance the quality and transparency of scientifi c reporting, with clear benefits for evidence reviews and health-policy decision making.

Department of Infection and Population Health, University College London, London, UK (N Field PhD, P Sonnenberg PhD, Prof I Abubakar FRCP); **Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA, USA** (T Cohen DPH); **Microbiology Coordination Section, European Centre for Disease Prevention and Control, Stockholm, Sweden** (Prof M J Struelens PhD, D Palm PhD); **Laboratory of Healthcare Associated Infection** (B Cookson FRCPath), **Immunisation, Hepatitis, and Blood Safety** (M Ramsay MRCP), **Modelling and Economics Department** (A Charlett PhD), **and Bioinformatics Unit** (J Green PhD), **Centre for Infectious Disease Surveillance and Control, Public Health England, London, UK; Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK** (Prof J R Glynn PhD); **Centre for Primary Care and Public Health, Queen Mary's University, London, UK** (V Gallo PhD); **National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA** (D MacCannell PhD); **Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland** (Prof M Egger MD); **and School of Public Health, Imperial College London, London, UK** (Prof P Vineis PhD)

Correspondence to:
Prof Ibrahim Abubakar, Institute of Epidemiology and Health, Faculty of Population Health Sciences, University College London, London WC1E 6JB, UK
i.abubakar@ucl.ac.uk

Background

Through the synthesis of epidemiology and molecular biology, a powerful set of scientifi c methods have been developed to investigate the transmission, distribution, biology, and diversity of infectious organisms.[1] The term molecular epidemiology emerged to describe this field in 1973, in relation to infl uenza virus strains.[2] Many technological advances have been made since that time, with typing methods undergoing a process of near-continuous evolution. Molecular epidemiology has been variously defined[3–6] and applied to many non-infectious settings.[7–9] To set the focus for this statement and to be inclusive of past, present, and future technologies, we have selected the following defi nition for infectious-disease molecular epidemiology:[10] the use of molecular-typing methods for infectious agents in the study of the distribution, dynamics, and determinants of health and disease in human populations.

Pathogen typing was once mainly phenotypic,[1,11] but this approach has been largely superseded by four broad types of genetic analysis:[11–13] direct comparison of DNA or RNA sequence data (including whole-genome sequencing); use of gel electrophoresis to compare band patterns, which can derive from pathogen-specific PCR amplifi cation products or DNA fragmentation by restriction endonuclease enzymes that cut DNA at specifi c recognition sequences; use of fluorescently labelled nucleotides to measure the accumulation of PCR products; and use of hybridisation, in which short nucleic-acid recognition sequences (probes) are attached to a matrix to compare gene expression or different genotypes of pathogens.

The basic assumption underlying most molecular epidemiological studies is that pathogens with similar profiles (as measured by molecular typing) are probably related, and that the degree of similarity between strains can be used to infer the time since their divergence.[10] Molecular-typing data (informed by appropriate epidemiological data relating to infected hosts or a point source) can thus be used to test epidemiological hypotheses as to whether cases are linked by recent transmission events or not.[11] Epidemiological studies increasingly use such data, for example from routine surveillance or outbreak investigations, to increase the precision of case definitions and explore the determinants of disease transmission.

Applications of molecular epidemiology include measurement of the geographical distribution of pathogen strains over time, place, and person to inform outbreak investigations and assess prevention and control interventions; contribution to surveillance programmes (eg, to identify the emergence of clinically important phenotypes); support for investigations into transmission dynamics and clustering of infections; understanding the evolution of pathogens; to inform the classification and identification of new species; vaccine design and monitoring for escape mutants; and understanding why infections recur. Recent specific examples with direct applicability to public health include the use of whole-genome sequencing to verify an outbreak of meticillin-resistant *Staphylococcus aureus* in a special-care baby unit and to trace the source to a health-care worker carrier;[14,15] characterisation and linking of cases of infection with the Shiga-toxin-

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

producing *Escherichia coli* 0104:H4 strain by real-time whole-genome sequencing during an outbreak in Germany in 2011;[16,17] and the finding of rapid global transmission of influenza A H1N1 during the 2009 pandemic.[13,18]

In some studies, functional genomic differences within a pathogen species are of interest in their own right. These studies use epidemiological observational designs to assess associations between functional characteristics (eg, pathogenesis, clinical symptoms, and drug resistance) and genetic characteristics. Technology for whole-genome sequencing will lead to these studies becoming increasingly common in the future.

The number of published articles about molecular epidemiology in infectious disease has increased rapidly (figure). In a search of PubMed with the terms "infection" AND "molecular epidemiology", we identified more than 10 000 articles published between 1975 and 2010, with more than 1000 added in 2010. However, scientific publications are of varying quality,[19] have not consistently included true epidemiological applications (the term molecular epidemiology is sometimes used to describe studies that were never intended to be epidemiological).[20] Epidemiological studies are subject to several threats to valid inference (eg, selection bias, uncontrolled confounding, and failure to identify and report stratified results in the presence of effect modification),[20] and reporting practices can be poor.[21–23]

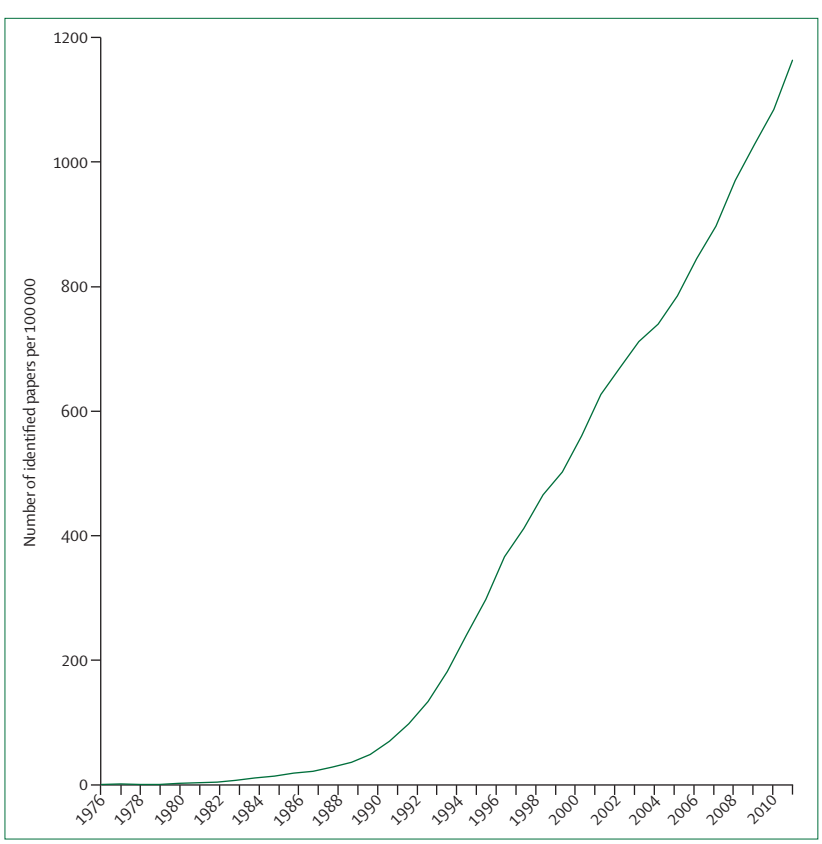

Figure: **Number of papers per 100 000 published per year identified with the search terms "infection" AND "molecular epidemiology" in the Medline database in that year**

In other epidemiological subspecialties, prominent statements setting out detailed recommendations aim to provide guidance to authors and journals for the reporting of research. These statements include the Consolidating Standards of Reporting Trials (CONSORT) statement,[24] the Outbreak Reports and Intervention Studies of Nosocomial Infection (ORION) statement,[25] the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) initiative,[21,26] the Strengthening the Reporting of Genetic Association Studies (STREGA) statement,[27,28] and the Reporting of Tumour Marker Studies (REMARK) statement.[29] When adopted, such statements improve the quality of reporting in research, inform the use of appropriate terminology, and set minimum standards for comparison, with clear benefits for meta-analyses, evidence reviews, and decision making in public health. A recent STROBE extension, STROBE for Molecular Epidemiology (STROBE-ME), addressed reporting issues associated with the use of human biomarkers in epidemiological studies but did not deal with issues specific to infectious diseases.[30] So far, no statements (with the exception of ORION) have mentioned infectious diseases. Although other authors have identified reporting issues that arise in bacterial molecular epidemiology,[31] no comprehensive statement that systematically addresses the full range of issues relating to the reporting of the design, collection, collation, and analysis of molecular epidemiological data for infectious diseases has been published.

Aims and use of the STROME-ID statement

We aim to extend the STROBE statement, with reference to other statements when appropriate, to improve and standardise the reporting of infectious-disease molecular data in epidemiological research. Our intention is to explain how to report research comprehensively, not to dictate how research should be done. The new statement extension, Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID), provides a checklist of items that should be read in conjunction with the STROBE statement (table). We specifically excluded guidance for studies about the genetics of human hosts and human biomarkers, for which detailed guidance is provided by STREGA and STROBE-ME, respectively.[27,28,30] The STROME-ID statement therefore focuses on aspects of reporting for which epidemiological studies that include infectious-disease molecular data differ from epidemiological studies without such data. We give detailed descriptions and provide background information in the appendix. The explicit aim of this statement is to improve the reporting of studies and, in turn, to assist interpretation of the data and increase understanding of what was actually done by researchers.

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

Development of standards

The STROME-ID statement was developed by a core working group, which met and communicated to agree the structure and content of the statement. Consensus was sought and reached by circulating iterative versions and teleconferences to discuss and to agree content. We intentionally sought a broad authorship group to include several institutions, countries, and specialties. We included epidemiologists, statisticians, bioinformaticians, virologists, and microbiologists who were specialists in infectious diseases, molecular epidemiology, public health, and control of communicable diseases.

Overall, the structure of the STROME-ID statement follows the model of STROBE and other similar statements. Specifi cally, STROME-ID has been developed as an extension to the STROBE statement and the points should be read in addition to the 22 elements of STROBE (table).

STROME-ID standards

Title and abstract

Introduction

The term molecular epidemiology should be applied to the study in the title or abstract and the keywords when molecular and epidemiological methods contribute substantially to the study (STROME-ID 1.1). Appropriate use of terminology is essential to enable identification of relevant research information. Studies can be tagged with inaccurate medical subject heading (MeSH) labels, which reduces their usefulness as search terms.[32,33] The term molecular epidemiology should therefore be used to describe infectious-disease research with true molecular and analytical epidemiological components.[20] Authors reporting findings from studies of pathogen typing without epidemiological analyses should avoid use of the term.

Background rationale

Provide background information about the pathogen population and the distribution of pathogen strains within the host population at risk (STROME-ID 2.1). Phylogeographical differences between regions and populations in the distribution of strains can be substantial. For this reason, the distribution of pathogen strains within the host population should be stated, when available, to provide context for the study and to assist interpretation of findings. The level of detail should be appropriate to the study setting and the research question. An evolving example is that of measles.[34] Measles genotyping for surveillance uses a small sequencing window of the measles *N* gene (N-450).[35] However, because transmission chains are increasingly interrupted by vaccination the genetic variability of strains in some areas has decreased, which limits the discriminatory ability of genotyping the N-450 sequence.[34] As such, the variability within the population under study is needed to enable assessment of whether the sequencing window is sufficiently wide to maintain sensitivity in areas approaching measles elimination.[34]

Objectives

State the epidemiological objectives of using molecular typing (STROME-ID 3.1). The study introduction should provide a clear rationale for the methods used and inform the reader of the epidemiological objectives of molecular typing in the context of the study design. In doing so, citation of any previous methodological studies that justify the use of the molecular techniques for the aims described and stating of their performance characteristics (eg, discriminatory power and marker stability) is helpful (appendix).

Methods

Study design

Define or cite definitions for key molecular terms used within the study (eg, strain, isolate, and clone; STROME-ID 4.1). The precise meaning of some terms—such as strain, isolate, clone, and clade—depends on the study question, the molecular techniques used, and the pathogen.[36] As for many aspects of the STROME-ID statement, definitions are important for reasons of consistency and comparability, and we recommend, as an overriding general principle, that authors provide or cite the intended definitions of key terms.

We do not provide a full glossary of terms here because these are often context dependent. An exception is the term strain, which is widely used in this statement and is essential to most molecular epidemiological studies, partly because strains can have very different functional characteristics.[37–39] Most molecular methods define strains using a small number of markers, which by definition can fail to take into account potential differences in other parts of the pathogen genome. With the exception of whole-genome sequencing, strain identity is therefore inferred rather than proven in many cases. We suggest the following, intentionally broad, definition as a guide: a group of pathogens that share an indistinguishable genome sequence by descent,[36] but which, by the molecular technique or techniques used, are measurably distinct from other pathogens of the same species.[10] However, we recognise that this definition might not be appropriate to all studies or pathogens, and that a more functional definition (provided or referenced by authors) might be needed for some publications.

Clearly define the molecular markers that were used with a standard nomenclature (STROME-ID 4.2). A molecular marker is a phenotypic, proteomic, or genetic characteristic (in this case of a pathogen) that can be used to differentiate between isolates within a species. The level of resolution of the molecular markers used in the study should be sufficient to satisfy the molecular epidemiological objectives (appendix).

Clearly state the infectious-disease case definitions (STROME-ID 4.3). The infectious-disease case definition

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

	Item number	STROBE items	STROME-ID items
Title and abstract			
Introduction	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	STROME-ID 1.1: the term molecular epidemiology should be applied to the study in the title or abstract and the keywords when molecular and epidemiological methods contribute substantially to the study
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	STROME-ID 2.1: provide background information about the pathogen population and the distribution of pathogen strains within the host population at risk
Objectives	3	State specific objectives, including any prespecified hypotheses	STROME-ID 3.1: state the epidemiological objectives of using molecular typing
Methods			
Study design	4	Present key elements of study design early in the paper	..
Molecular terminology		..	STROME-ID 4.1: define or cite definitions for key molecular terms used within the study (eg, strain, isolate, and clone)
Molecular markers		..	STROME-ID 4.2: clearly define the molecular markers that were used with a standard nomenclature
Infectious disease case definition		..	STROME-ID 4.3: clearly state the infectious-disease case definitions
Laboratory methodology		..	STROME-ID 4.4: describe sample collection and laboratory methods, including any methods used to minimise and measure cross-contamination, and give the criteria used to interpret strain classification
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	STROME-ID 5.1: clearly state the timeframe of the study; consider and appropriately reference the molecular clock of markers if known, and the natural history of the infection
Participants	6	(a) *Cohort study*—give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—give the eligibility criteria, and the sources and methods of selection of participants (b) *Cohort study*—for matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—for matched studies, give matching criteria and the number of controls per case	STROME-ID 6.1: state the source of participants and clinical specimens, and clearly describe sampling frame and strategy
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	..
Data sources/ measurement	8*	For each variable of interest give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	..
Multiple-strain infections		..	STROME-ID 8.1: describe any methods used to detect multiple-strain infections and measure their effect on the study findings
Bias	9	Describe any efforts to address potential sources of bias	STROME-ID 9.1: describe any efforts made to address discovery or ascertainment bias
Study size	10	Explain how the study size was arrived at	STROME-ID 10.1: describe any unique restrictions placed on the study sample size
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	..
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study*—if applicable, explain how loss to follow-up was addressed *Case-control study*—if applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—if applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	STROME-ID 12.1: state how the study took account of the non-independence of sample data, if appropriate STROME-ID 12.2: state how the study dealt with missing data
Results			
Participants	13*	(a) Report the numbers of individuals at each stage of the study (eg, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	STROME-ID 13.1: Report numbers of participants and samples at each stage of the study, including the number of samples obtained, the number typed, and the number yielding data STROME-ID 13.2: if the study investigates groups of genetically indistinguishable pathogens (molecular clusters), state the sampling fraction, the distribution of cluster sizes, and the study population turnover, if known

(Table continues on next page)

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

	Item number	STROBE items	STROME-ID items
(Continued from previous page)			
Descriptive data	14*	(a) Give characteristics of study participants (eg, demographic, clinical, social) and information on exposures and potential confounders b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study*—summarise follow-up time (eg, average and total amount)	STROME-ID 14.1: give information by strain type if appropriate, with use of standardised nomenclature
Outcome data	15*	*Cohort study*—report numbers of outcome events or summary measures over time *Case-control study*—report numbers in each exposure category, or summary measures of exposure *Cross-sectional study*—report numbers of outcome events or summary measures	..
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorised (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	STROME-ID 16.1: consider showing molecular relatedness of strain types by means of a dendrogram or phylogenetic tree
Other analyses	17	Report other analyses done (eg, analyses of subgroups and interactions, and sensitivity analyses)	..
Discussion			
Key results	18	Summarise key results with reference to study objectives	..
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	STROME-ID 19.1: consider alternative explanations for findings when transmission chains are being investigated, and report the consistency between molecular and epidemiological evidence
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	..
Generalisability	21	Discuss the generalisability (external validity) of the study results	..
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	..
Ethics	23	..	STROME-ID 23.1: report any ethical considerations with specific implications for infectious-disease molecular epidemiology

STROBE=Strengthening the Reporting of Observational Studies in Epidemiology. STROME-ID=Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases. *Give such information separately for cases and controls in case-control studies, and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Table: **The STROBE checklist and additional STROME-ID items**

should be stated with a level of detail appropriate to the study question. The case definition can include symptoms, anatomical site of sample collection, typing information, methods to distinguish between carriers and cases (eg, to distinguish between contacts who are asymptomatically colonised with *Neisseria meningitidis* and cases of invasive disease[40]), identification of chronic or recent infections (eg, strict serological definitions are needed to distinguish between acute and chronic infections with viral hepatitis[41,42]), and relevant exclusions (eg, smear-negative or extrapulmonary *Mycobacterium tuberculosis*).[43] When the study objective is to investigate transmission within a population, care should be taken that several isolates from an individual patient do not bias measures of relatedness in a sample.

Inclusion of molecular-typing characteristics within the case definition can increase the precision of the definition and reduce the likelihood of misclassification. Improved precision will tend to reduce the number of cases needed to establish an association and improve accuracy of statistical estimations in molecular epidemiological studies. When molecular characteristics are included within the case definition, particular care should be taken to ensure that comparable definitions have been used.[44]

Describe sample collection and laboratory methods, including any methods used to minimise and measure cross-contamination, and give the criteria used to interpret strain classification (STROME-ID 4.4). Authors should include detailed microbiological and practical methodological descriptions of sample collection and processing, both to inform the reader if they might affect interpretation, and to ensure reproducibility of the study. Setting (eg, community *vs* health-care, how the samples were collected, or the laboratories used), sample type,

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

timing of collection in relation to symptoms, use of antimicrobials at the time of collection, handling and storage of the sample before processing, timing of different steps in sample processing, temperature, length of culture, and effectiveness of culture recovery can all affect the successful isolation and typing of an organism, and should be reported.[30] Specimen workflow can also be an important variable, dependent on the study type.

For well established methods, authors should provide appropriate citations supporting the validity of the method, and if necessary explain how molecular measurements were interpreted.[31] Standardisation of procedures across laboratories should be stated and referenced if done. The criteria used to identify strains according to molecular markers should be described in detail with reference to standard nomenclature, if available. Any variation from published interpretation protocols or deviation from standard nomenclature should be explicitly stated.

If the study describes specific methods or markers for specific epidemiological objectives or time-population frames, authors should report data for the validation of markers and the resolution capacity of these data within this context.[44] Performance characteristics of molecular markers and analytical methods for molecular epidemiological studies include typability, reproducibility, discriminatory power, stability, and concordance with epidemiological evidence.[31] Dependent on study design and typing method, portability of data between databases might be important.[31] A description of reporting practice for assay type, detection limit, reliability of measurement, and calibration procedures has been documented elsewhere.[30]

Description of the molecular-typing methods used should be clearly linked to the epidemiological objectives, scope of the study, and description of molecular markers analysed. Investigators should ensure that the combined discriminatory power of the methods used is sufficient to achieve the objectives of the study. The discriminatory ability of the typing system used should therefore be reported whenever possible (appendix).

Cross-contamination is a particular difficulty for infectious-disease research, and usually results in false-positive cases. For example, findings from several studies of tuberculosis have shown cross-contamination rates of 2–4%,[45,46] and rates can be higher if laboratory standards are poor. Authors should therefore describe any methods used to minimise cross-contamination, such as those described by the Guidelines for Australian Mycobacterial Laboratories,[47] and report cross-contamination rates if measured.

Setting

Clearly state the timeframe of the study; consider and appropriately reference the molecular clock of markers if known, and the natural history of the infection (STROME-ID 5.1). The timeframe of a study has particular importance in investigations of pathogen relatedness and transmission. Any epidemiological inferences made about transmission events on the basis of molecular differences between pathogens need to be consistent with the pathogen, the timeframe of the study, and the molecular clock of the marker used (appendix). The study timeframe should therefore be clearly stated, taking into account other factors such as spatial context, social networks, and background prevalences.

The term clustering is used to denote groups of individuals that are more similar to each other than to members of another group. When clustering refers to the grouping of pathogens by molecular markers, the proportion clustered tends to increase with time, reaching a plateau that depends on the molecular clock of the marker.[48]

For transmission studies, key characteristics of a pathogen's natural history should also be considered with respect to the timeframe, and reported accordingly. These characteristics include serial intervals and infectious and incubation periods. For example, if two patients are infected with a similar strain, but have disease symptoms within less than the shortest possible serial interval for the infection, direct transmission could not have occurred.

Participants

State the source of participants and clinical specimens, and clearly describe sampling frame and strategy (STROME-ID 6.1). Selection of participants and samples is a potential source of bias for infectious-disease molecular epidemiology, as for many other epidemiological studies. Infectious-disease clinical research frequently includes selection methods that need specific reporting considerations. These methods include case finding, contract tracing, surveillance programmes, and convenience sampling, which can affect representativeness of samples and generalisability of the results.

A key variable is whether all consecutive incident cases in an exposed population, or only a selected subset, are sampled. To increase the proportion of cases identified, a study can include active case finding or contact tracing; investigators of a case-control study[49] about infection with hepatitis A used sequence data to link cases in several US states to a single food source. In some instances, such case finding can affect the validity of assumptions about the independence of cases and increase the likelihood of identification of related strains. Whether this effect of case finding affects valid inference from study findings depends on the study question. STROME-ID 12.1 describes in more detail how statistical methods may be used to take account of this issue.

For studies using surveillance programmes and organisation networks for typing to obtain samples, authors should provide information about the representativeness of samples and generalisability of results, in addition to information about the underlying

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

population from which samples are obtained (STROME-ID 2.1). The European Centre for Disease Prevention and Control has proposed key design features for effective and consistent networks that collect representative isolates.[13] These lend support to several STROME-ID components including agreement about a minimum dataset, a common typing approach, a suitable timeframe, rules about sample submission, and regular calibration of the structure and molecular diversity of the baseline population.[13] Authors should also consider and comment on bias associated with reference strains and database collections, which are more likely than samples from the whole population to contain pathogenic, clinically relevant strains and opportunistic samples.[50]

Use of convenience sampling (eg, a sample defined by a geographical location or clinical service) can reduce the representativeness of samples and generalisability of a study, and tends to overestimate effects measured by clustering (STROME-ID 13.2).

Data source and measurement

Describe any methods used to detect multiple-strain infections and measure their effect on the study findings (STROME-ID 8.1). Several distinct strains of a pathogen can simultaneously infect individual hosts, which can substantially complicate the design and analysis of findings from molecular epidemiological studies. Multiple-strain infections occur with a wide range of human pathogens (eg, viruses, bacteria, fungi, protozoa, and helminths) and the prevalence of detectable multiple-strain infections is sometimes high (median 11·3%, mean 21·7% of infections), even with use of fairly insensitive detection methods.[36] Multiple-strain infections of congenital cytomegalovirus were isolated from a single anatomical site and from different sites in the same host.[51] For some pathogens (eg, RNA viruses such as dengue and HIV) in-host variability is a mandatory event driven by low proofreading fidelity of the viral RNA polymerase, which leads to the accumulation of variants known as quasispecies.[52,53] This differentiation makes detection of multiple-strain infections more difficult.

The probability of detection of multiple-strain infections, when present, depends on the design, execution, and analysis of the study. Factors described in STROME-ID 4.1–4.4, including sample collection (eg, number and type), sample transportation and storage, laboratory processing (eg, decontamination, culture, and subculture), selection of isolates for molecular typing, and the molecular-typing method used, all affect the likelihood of detection of multiple-strain infections and the ability of researchers to interpret findings in the presence of this type of heterogeneity.

Molecular-typing methods that can identify more alleles at a single locus than are possible within a single strain (eg, mycobacterial interspersed repetitive unit-variable number tandem repeat typing[54]) might be the most useful to identify the presence of multiple-strain infections. By contrast, typing methods that assess the presence or absence of specific markers (eg, spacer oligonucleotide typing[55]) can produce results that are difficult to interpret in the presence of multiple-strain infections. Some computational methods to identify patterns that are consistent with multiple-strain infections have been suggested.[56]

When reporting results, authors should consider, and report if appropriate, to what extent multiple-strain infections could alter interpretations of their findings.[57] If multiple-strain infections have been detected, authors should state this finding and clearly describe how their analysis has accounted for such complex infections. In reports of studies for which multiple-strain infections might be present but have not been detected, authors should discuss how infection with several strains might affect the interpretation of their results.

Bias

Describe any efforts made to address discovery or ascertainment bias (STROME-ID 9.1). Single nucleotide polymorphisms (SNPs) are widely used to identify sequence variation within pathogen populations.[58] This method usually needs a set of wholly sequenced strains that constitute a panel of genetic variation (identified by SNPs) across a pathogen's genome, which is used to define sequence variation in new strains.[59] Use of a subset of sequences to define the whole population with a set of SNPs can introduce bias, because subgroups are defined by the strains included. This issue has been called discovery or ascertainment bias, and can lead to so-called branch collapse that was first used to describe the bamboo-like evolutionary trees sometimes resulting from the use of SNPs to explore evolution of *M tuberculosis*.[60] Pearson and colleagues[61] subsequently reported methods to define and control for ascertainment bias in modelling studies of *Bacillus anthracis* genomes. Authors should mention whether their results could be subject to this form of bias, and describe whether additional analyses have been done to take account of these effects.

Study size

Describe any unique restrictions placed on the study sample size (STROME-ID 10.1). Technological advances and the falling unit cost of molecular assays tend to increase the ability of researchers to process large numbers of samples.[62] However, molecular epidemiological studies can still be associated with high laboratory costs and logistic troubles. This difficulty can place greater limits on the ability of researchers to obtain adequate sample sizes for molecular epidemiological studies than for other epidemiological studies.[63] When the sample size is smaller than ideal, type 2 errors will be more common, such that the study findings might be insufficient to justify a conclusion. Authors should report when the power of the study has been limited by technical, financial, or other such constraints.

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

Statistical methods

State how the study took account of the non-independence of sample data, if appropriate (STROME-ID 12.1). Non-independence of molecular data needs careful consideration in the reporting of infectious-disease molecular epidemiological studies. This difficulty can be caused by the fact that samples tend to be collected from predetermined groups of patients (eg, wards or hospitals) and that individuals within these groups are more likely to be similar to each other than to unrelated individuals. Furthermore, the very nature of infections—ie, transmission from one person to another—also introduces grouping of events in both time and space. For example, one of the main reasons to analyse surveillance data is to identify unusual clusters of events.

The statistical methods used should address the lack of independence between individuals. For some study findings, the aggregation of individual data into larger units can be dealt with by directly accounting for the structure of the dataset during statistical analysis—eg, by introduction of random effects in a statistical regression model.

State how the study dealt with missing data (STROME-ID 12.2). Systematic differences between cases with and without missing molecular data can introduce bias (appendix). As such, we suggest that authors should avoid using the availability of data as an inclusion criterion. As for other studies, the level of completeness should always be declared and the implications of the level discussed. Authors should provide information about the type of missingness—ie, missing wholly at random, missing at random, or missing not at random. Statistical methods used to handle missing data should be stated and justified.[64] Missing data can be a particular difficulty when molecular data are used to track and link transmissions, because parts of the transmission chain will be unidentifiable. For example, in analyses of data for *M tuberculosis* outbreaks, recent transmission is often inferred from the proportion of individuals with shared strains, but this outcome is underestimated when data are incomplete.[22,64] Further detailed descriptions about potential bias introduced by missing data are mentioned in the STROBE statement and supporting publications.[21,26]

Results

Participants

Report numbers of participants and samples at each stage of the study, including the number of samples obtained, the number typed, and the number yielding data (STROME-ID 13.1). In addition to STROBE 13,[21,26] authors should consider specific factors that might contribute to missing data for molecular typing for pathogens (appendix), and report numbers of participants (and numbers of samples if different) at each stage of sample processing, including at isolation, culture, and typing stages as appropriate.

If the study investigates groups of genetically indistinguishable pathogens (molecular clusters), state the sampling fraction, the distribution of cluster sizes, and the study population turnover, if known (STROME-ID 13.2). In studies of groups of genetically indistinguishable pathogens (molecular clusters), analyses can be subject to several biases. Reporting of the sampling fraction, the distribution of cluster sizes, and the study population turnover helps readers interpret the extent to which these biases exist in such cases.

The sampling fraction is the number of cases in the study as a proportion of all cases occurring in the population (including those not detected). Generally, increasing the sampling fraction tends to increase the proportion of samples that are clustered, with diminishing returns towards a plateau representing the true proportion clustering in a population.[48,65] For *M tuberculosis*, the proportion of clustered *M tuberculosis* strains detected in a population provides a proxy measure of recent infections.[66] Methods used to estimate the proportion of clustered and unique cases can be biased if the sample does not include all the cases in the population.[22,48] Reporting of the sampling fraction therefore allows assessment of the potential for bias in this estimate.[67] If the population size is not known, the likelihood of clustering bias can be inferred from the distribution of cluster sizes, which also shows the presence of dominant strains in the population studied.[44]

Several methods are available to estimate the proportion of cases clustered.[64] For example, the *N* method sums all cases reported in a cluster, and is often used in investigation of transmission chains. The $N - 1$ method assumes one case in each cluster is the index and adjusts the total accordingly. The method used should be stated clearly.

Population turnover will also affect detection of molecular clustering. For example, in a population with high immigration rates, more unique cases than would be expected will be detected if compared with a closed population. Selective emigration of infected or non-infected people in the study population will also affect clustering measures.[64] The turnover of the study population and the migration status of participants should therefore be reported if known.[44]

Another consideration for measures of molecular clustering is the establishment of latent infection by organisms such as *M tuberculosis* and some herpes viruses. Reactivation in individuals infected in large historical clusters could lead to the identification of clustering for later cases with no immediately obvious epidemiological links, which could lead to misinterpretation.[68]

Descriptive data

Give information by strain type if appropriate, with use of standardised nomenclature (STROME-ID 14.1). The analysis and reporting by strain type of epidemiological information (demographic, clinical, and behavioural) can be informative in some studies. Statistical models might need stratification by pathogen strains, and authors

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

should therefore report sufficient detail for readers to understand the appropriateness of decisions about model construction. For example, a study[69] of meticillin-resistant *S aureus* compared characteristics of patients infected with the USA100 strain with those of patients infected with the USA300 strain, and assessed the antibiotic resistance profile of both strains. Whenever possible, standardised nomenclature, for example using publicly available multilocus sequence typing databases,[1] should be used to facilitate comparisons between studies.

Main results

Consider showing molecular relatedness of strain types by means of a dendrogram or phylogenetic tree (STROME-ID 16.1). Dendrograms or other graphical representational methods can be used to depict quantitative estimates of divergence within and between strains, typically done in reference to an unrelated control strain.[70] If the typing methods used provide valid estimates for the genetic distance between strains, phylogenetic trees can be computed with use of appropriate algorithms and bootstrap analysis, and the results can be displayed as a tree indicating statistical CIs for the branch nodes.[1,70]

Discussion

Limitations

Consider alternative explanations for findings when transmission chains are being investigated, and report the consistency between molecular and epidemiological evidence (STROME-ID 19.1). For any putative transmission chain from person A to person B, several diff erent scenarios might reasonably account for two individuals being infected with identical or closely related pathogens.[71] These scenarios include different directions of transmission (A to B or B to A), intervening cases (A to C to B), and a common source (C to A and C to B). The direction of transmission, and whether intervening (possibly undiagnosed) cases have not occurred, can be particularly diffi cult to establish. These difficulties have been well described when molecular-typing evidence has been used in legal settings.[72,73] Such cases show that transmission is much easier to disprove than to prove beyond doubt. Furthermore, litigation cases show that molecular data are of little value without linked epidemiological data, which provide essential supportive evidence for any conclusions drawn.[74] Epidemiological and molecular evidence therefore need to be considered together and checked for consistency.[12]

At a population level, the infection network can be far wider than the cases studied, with many unknown and unknowable transmissions occurring.[3] Pathogens can be transmitted from the environment, animals, or other individuals, and through many different transmission routes (eg, airborne, ingested, direct contact, sexual contact, vertical, and others). For these reasons, authors should consider reporting on the one hand whether distinct transmission chains might show evolutionary convergence, and on the other whether infections with several strains can occur simultaneously (appendix).

Other information

Ethics

Report any ethical considerations with specific implications for infectious-disease molecular epidemiology (STROME-ID 23.1). Bioethical principles, such as those set out by Beauchamp and Childress,[75] should guide researchers of all studies involving human participants, and studies should be reviewed by accredited research ethics committees or their equivalent. For studies in which direct transmission of infections is investigated, fully informed consent and confidentiality are of paramount importance.[76] In extreme cases, studies can have legislative implications in countries where individuals have been prosecuted under criminal law for exposing others to infections such as HIV,[71,77] hepatitis B virus,[78] and herpes simplex virus.[79] The decision as to whether or not to return results to participants should also be presented and justified by authors,[80] taking into account the public health implications for people who have been in contact with those diagnosed with (possibly asymptomatic) infections. Authors should therefore report ethical approval.

Conclusions

The STROME-ID statement aims to support transparent reporting of research (rather than to guide research methods) and emphasises specific examples for which epidemiological research into infectious diseases can be subject to interpretation errors because of misleading reporting of molecular data. We have given particular emphasis to the interpretation of findings and the transparent description of limitations in this context. This statement has been designed to be directly useful to authors and we expect it to be highly relevant to scientists, clinicians, public health specialists, journal staff, and grant reviewers with an interest in infectious diseases. Its longevity and effect will depend on wide dissemination and systematic use. To this end, we have developed the statement following the example of the widely applied STROBE collaboration.

We acknowledge several limitations in this statement. We have reached consensus through iterative revisions by contributors mostly from Europe and North America who were identified through scientific networks. The statement is intentionally written from an epidemiological perspective (rather than with a laboratory focus), and is probably affected by the composition of this group. In common with other similar statements, we have not done a formal systematic review for each item of the checklist.[21,30] Reporting guidelines can improve the quality of published articles,[81] but we recognise that authors are often limited by word count and that many high-quality studies are published in the form of short reports;

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

Search strategy and selection criteria

We identified relevant articles with searches of Medline, PubMed, and references from articles, with the terms "molecular epidemiology" AND "infection". To support the specific examples provided, further searches were undertaken combining search terms such as "multiple-strain", "molecular typing", "molecular marker", "molecular clock", "bias", "methods", "clustering", and "missing data", with searches for articles describing specific pathogens. Only articles published in English were included. We used no date restrictions. The last search was done in July, 2012.

compliance with every item of this statement might not be possible. Many journals increasingly provide opportunities to publish additional material online, and such options should be used to ensure full reporting of high-quality research without limiting publication.

Finally, we recognise that STROME-ID must form the basis for a work in progress, and we expect and welcome comments relating to structure, content, and use of this statement. In the near future, whole-genome sequencing and proteomics applications will provide additional typing resolution that is not feasible at present, and we expect the emergence of these and other technologies with new epidemiological applications that might necessitate future revisions. Likewise, the use of molecular data might be extended beyond purely observational studies and used in trials and interventions, which could warrant further reporting considerations. Nevertheless, we expect that many of the underlying principles, as described, will remain highly relevant.

Contributors

NF, TC, MJS, DP, BC, JRG, MR, PS, DM, and IA conceived of and designed the structure and content of the statement. NF and IA did the literature review. NF, TC, MJS, DP, JRG, and IA wrote the first draft of the statement. All authors contributed to the writing of the statement and agree with its content and conclusions.

Declaration of interests

We declare that we have no competing interests.

Acknowledgments

We thank Mike Catchpole (Health Protection Agency, Centre for Infections, UK), who set up the Centre for Infections, Epidemiology and Surveillance Governance Group, through which this work was initiated. NF is supported by a National Institute for Health Research Academic Clinical Lectureship. IA is supported by a National Institute for Health Research Senior Research Fellowship.

References

1 Morand S, Beaudeau F, Cabaret J. New frontiers of molecular epidemiology of infectious diseases. Berlin: Springer, 2012.
2 Kilbourne ED. The molecular epidemiology of influenza. *J Infect Dis* 1973; **127:** 478–87.
3 Besser J. Use of molecular epidemiology in infectious disease surveillance. In: M'ikanatha NM, Lynfield R, Van Beneden CA, de Valk H, eds. Infectious disease surveillance, 1st edn. Hoboken, NJ: Wiley-Blackwell, 2008: 393–407.
4 Traub RJ, Monis PT, Robertson ID. Molecular epidemiology: a multidisciplinary approach to understanding parasitic zoonoses. *Int J Parasitol* 2005; **35:** 1295–307.
5 Monis PT, Andrews RH. Molecular epidemiology: assumptions and limitations of commonly applied methods. *Int J Parasitol* 1998; **28:** 981–87.
6 Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. *Clin Microbiol Rev* 2006; **19:** 658–85.
7 Vineis P, McMichael AJ. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* 1998; **19:** 2063–67.
8 Ioannidis JPA. Molecular bias. *Eur J Epidemiol* 2005; **20:** 739–45.
9 Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol Biomarkers Prev* 2007; **16:** 1954–65.
10 Hall A. What is molecular epidemiology? *Trop Med Int Health* 1996; **1:** 407–08.
11 Riley LW. Molecular epidemiology of infectious disease: principles and practices. Washington, DC: ASM Press, 2004.
12 Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF. Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol Perspect Innov* 2005; **2:** 10.
13 Palm D, Johansson A, Ozin A, Friedrich AW, Larsson JT, Struelens MJ. Molecular epidemiology of human pathogens: how to translate breakthroughs into public health practice, Stockholm, November 2011. *Euro Surveill* 2012; **17:** 2.
14 Köser CU, Holden MTG, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012; **366:** 2267–75.
15 Harris SR, Cartwright EJP, Török ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis* 2013; **13:** 130–36.
16 Bielaszewska M, Mellmann A, Zhang W, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect Dis* 2011; **11:** 671–76.
17 Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 2011; **365:** 709–17.
18 Lemey P, Suchard M, Rambaut A. Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr* 2009; **1:** RRN1031.
19 Porta M, Malats N, Vioque J, et al. Incomplete overlapping of biological, clinical, and environmental information in molecular epidemiological studies: a variety of causes and a cascade of consequences. *J Epidemiol Community Health* 2002; **56:** 734–38.
20 Foxman B, Riley L. Molecular epidemiology: focus on infection. *Am J Epidemiol* 2001; **153:** 1135–41.
21 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, and the STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007; **4:** e296.
22 Glynn JR, Vynnycky E, Fine PEM. Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. *Am J Epidemiol* 1999; **149:** 366–71.
23 Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* 2009; **4:** e7815.
24 Moher D, Hopewell S, Schulz KF, et al, and the Consolidated Standards of Reporting Trials Group. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010; **63:** e1–37.
25 Stone SP, Cooper BS, Kibbler CC, et al. The ORION statement: guidelines for transparent reporting of outbreak reports and intervention studies of nosocomial infection. *J Antimicrob Chemother* 2007; **59:** 833–40.
26 Vandenbroucke JP, von Elm E, Altman DG, et al, and the STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007; **4:** e297.

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

27 Little J, Higgins JPT, Ioannidis JPA, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. *Eur J Epidemiol* 2009; **24:** 37–55.

28 von Elm E, Moher D, Little J, for the STREGA collaboration. Reporting genetic association studies: the STREGA statement. *Lancet* 2009; **374:** 98–100.

29 McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, and the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor MARKer prognostic studies (REMARK). *Nat Clin Pract Oncol* 2005; **2:** 416–22.

30 Gallo V, Egger M, McCormack V, et al, and the STROBE Statement. STrengthening the Reporting of OBservational studies in Epidemiology—Molecular Epidemiology (STROBE-ME): an extension of the STROBE Statement. *PLoS Med* 2011; **8:** e1001117.

31 van Belkum A, Tassios PT, Dijkshoorn L, et al, and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin Microbiol Infect* 2007; **13** (suppl 3)**:** 1–46.

32 Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 1994; **271:** 1103–08.

33 O'Rourke A, Booth A, Ford N. Another fine MeSH: clinical medicine meets information science. *J Inf Sci* 1999; **25:** 275–81.

34 Featherstone DA, Rota PA, Icenogle J, et al. Expansion of the global measles and rubella laboratory network 2005–09. *J Infect Dis* 2011; **204** (suppl 1)**:** S491–98.

35 WHO. Manual for the laboratory diagnosis of measles and rubella virus infection, 2nd ed. http://www.who.int/ihr/elibrary/manual_diagn_lab_mea_rub_en.pdf (accessed Nov 12, 2013).

36 Balmer O, Tanner M. Prevalence and implications of multiple-strain infections. *Lancet Infect Dis* 2011; **11:** 868–78.

37 de Roode JC, Pansini R, Cheesman SJ, et al. Virulence and competitive ability in genetically diverse malaria infections. *Proc Natl Acad Sci USA* 2005; **102:** 7624–28.

38 Turner CMR, Aslam N, Dye C. Replication, differentiation, growth and the virulence of *Trypanosoma brucei* infections. *Parasitology* 1995; **111:** 289–300.

39 Anderson RM. The pandemic of antibiotic resistance. *Nat Med* 1999; **5:** 147–49.

40 Vogel U, Morelli G, Zurth K, et al. Necessity of molecular techniques to distinguish between *Neisseria meningitidis* strains isolated from patients with meningococcal disease and from their healthy contacts. *J Clin Microbiol* 1998; **36:** 2465–70.

41 Andersson MI, Low N, Irish CJ, et al. Molecular epidemiology of a large community-based outbreak of hepatitis B in Bristol, U.K. *J Clin Virol* 2012; **53:** 125–29.

42 Chaudhuri S, Das S, Chowdhury A, Santra A, Bhattacharya SK, Naik TN. Molecular epidemiology of HCV infection among acute and chronic liver disease patients in Kolkata, India. *J Clin Virol* 2005; **32:** 38–46.

43 Bromham L, Penny D. The modern molecular clock. *Nat Rev Genet* 2003; **4:** 216–24.

44 Glynn JR, Bauer J, de Boer AS, et al, for the European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis* 1999; **3:** 1055–60.

45 Glynn JR, Yates MD, Crampin AC, et al. DNA fingerprint changes in tuberculosis: reinfection, evolution, or laboratory error? *J Infect Dis* 2004; **190:** 1158–66.

46 Burman WJ, Stone BL, Reves RR, et al. The incidence of false-positive cultures for *Mycobacterium tuberculosis*. *Am J Respir Crit Care Med* 1997; **155:** 321–26.

47 National Tuberculosis Advisory Committee. Guidelines for Australian mycobacteriology laboratories. *Commun Dis Intell* 2006; **30:** 116–28.

48 Murray M. Sampling bias in the molecular epidemiology of tuberculosis. *Emerg Infect Dis* 2002; **8:** 363–69.

49 Bialek SR, George PA, Xia G-L, et al. Use of molecular epidemiology to confirm a multistate outbreak of hepatitis A caused by consumption of oysters. *Clin Infect Dis* 2007; **44:** 838–40.

50 Pfaller MA, Acar J, Jones RN, Verhoef J, Turnidge J, Sader HS. Integration of molecular characterization of microorganisms in a global antimicrobial resistance surveillance program. *Clin Infect Dis* 2001; **32** (suppl 2)**:** S156–67.

51 Ross SA, Novak Z, Pati S, et al. Mixed infection and strain diversity in congenital cytomegalovirus infection. *J Infect Dis* 2011; **204:** 1003–07.

52 Wang W-K, Lin S-R, Lee C-M, King C-C, Chang S-C. Dengue type 3 virus in plasma is a population of closely related genomes: quasispecies. *J Virol* 2002; **76:** 4662–65.

53 Chin-inmanu K, Suttitheptumrong A, Sangsrakru D, et al. Feasibility of using 454 pyrosequencing for studying quasispecies of the whole dengue viral genome. *BMC Genomics* 2012; **13** (suppl 7)**:** S7.

54 Supply P, Magdalena J, Himpens S, Locht C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* 1997; **26:** 991–1003.

55 Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 1997; **35:** 907–14.

56 Lazzarini LCO, Rosenfeld J, Huard RC, et al. *Mycobacterium tuberculosis* spoligotypes that may derive from mixed strain infections are revealed by a novel computational approach. *Infect Genet Evol* 2012; **12:** 798–806.

57 Cohen T, van Helden PD, Wilson D, et al. Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. *Clin Microbiol Rev* 2012; **25:** 708–19.

58 Morin PA, Luikart G, Wayne RK, for the SNP workshop group. SNPs in ecology, evolution and conservation. *Trends Ecol Evol* 2004; **19:** 208–16.

59 Worobey M. Anthrax and the art of war (against ascertainment bias). *Heredity (Edinb)* 2005; **94:** 459–60.

60 Alland D, Whittam TS, Murray MB, et al. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J Bacteriol* 2003; **185:** 3392–99.

61 Pearson T, Busch JD, Ravel J, et al. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA* 2004; **101:** 13536–41.

62 Bonassi S, Au WW. Biomarkers in molecular epidemiology studies for health risk prediction. *Mutat Res* 2002; **511:** 73–86.

63 Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004; **96:** 434–42.

64 Murray M, Alland D. Methodological problems in the molecular epidemiology of tuberculosis. *Am J Epidemiol* 2002; **155:** 565–71.

65 Borgdorff MW, van den Hof S, Kalisvaart N, Kremer K, van Soolingen D. Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. *Am J Epidemiol* 2011; **174:** 243–51.

66 Alland D, Kalkut GE, Moss AR, et al. Transmission of tuberculosis in New York City: an analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med* 1994; **330:** 1710–16.

67 Houben RMGJ, Glynn JR. A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation. *Trop Med Int Health* 2009; **14:** 892–909.

68 Braden CR, Templeton GL, Cave MD, et al. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. *J Infect Dis* 1997; **175:** 1446–52.

69 Chua T, Moore CL, Perri MB, et al. Molecular epidemiology of methicillin-resistant *Staphylococcus aureus* bloodstream isolates in urban Detroit. *J Clin Microbiol* 2008; **46:** 2345–52.

70 Foxman B. Molecular tools and infectious disease epidemiology. San Diego, CA: Academic Press, 2012.

71 Bernard EJ, Azad Y, Vandamme AM, Weait M, Geretti AM. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med* 2007; **8:** 382–87.

Published in final edited form as: Lancet Infect Dis. 2014 Apr;14(4):341-52. doi: 10.1016/S1473-3099(13)70324-4

72 Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci USA* 2010; **107:** 21242–47.

73 Abecasis AB, Geretti AM, Albert J, Power L, Weait M, Vandamme A-M. Science in court: the myth of HIV fingerprinting. *Lancet Infect Dis* 2011; **11:** 78–79.

74 Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. *J Virol* 1994; **68:** 5918–24.

75 Beauchamp TL, Childress JF. Principles of biomedical ethics, 5th edn. Oxford: Oxford University Press, 2001.

76 UNAIDS. Policy brief: criminalization of HIV transmission. 2008. http://data.unaids.org/pub/basedocument/2008/20080731_jc1513_policy_criminalization_en.pdf (accessed Nov 12, 2013).

77 Chalmers J. The criminalization of HIV transmission. *Sex Transm Infect* 2002; **78:** 448–51.

78 Mohanty K. The first case of criminalization of transmission of hepatitis B in the UK: defendant sentenced to two years' imprisonment on the grounds of hepatitis B deoxyribonucleic acid sequencing. *Int J STD AIDS* 2009; **20:** 587–89.

79 Gurnham D. What role should criminal justice play in the fight against STIs? *Sex Transm Infect* 2012; **88:** 4–5.

80 Lévesque E, Joly Y, Simard J. Return of research results: general principles and international perspectives. *J Law Med Ethics* 2011; **39:** 583–92.

81 Cobo E, Cortés J, Ribera JM, et al. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. *BMJ* 2011; **343:** d6783.