# ClinicalCodes: An online clinical codes repository to improve the validity and reproducability of research using electronic medical records

David A Springate[1,2,*], Evangelos Kontopantelis[1,3], Darren M Ashcroft[4], Ivan Olier[5], Rosa Parisi[4], Edmore Chamapiwa[1], David Reeves[1,2]

**1 Centre for Primary Care, Institute for Population health, University of Manchester, Manchester, UK**

**2 Centre for Biostatistics, Institute for Population health, University of Manchester, Manchester, UK**

**3 Centre for Health Informatics, Institute for Population health, University of Manchester, Manchester, UK**

**4 Centre for Pharmacoepidemiology and Drug Safety Research, Manchester Pharmacy School, University of Manchester, Manchester, UK**

**5 Manchester Institute for Biotechnology, University of Manchester, Manchester, UK**

∗ **E-mail: Corresponding david.springate@manchester.ac.uk**

## Abstract

Lists of clinical codes are the foundation for research undertaken using electronic medical records and without access to them, reviewers are unable to determine the validity of research, true replication is impossible and researchers are unable to make effective comparisons between studies and are subject to much duplication of effort in building new code lists. Despite this, publication of clinical codes is rarely if ever a requirement for obtaining grants, validating protocols or publishing research. We have built a centralised online repository where electronic medical records researchers can upload and download lists of clinical codes to help address these problems. The repository will enable clinical researchers to better validate electronic medical records studies, build on previous code lists and compare disease definitions across studies. It will also assist health informaticians in replicating database studies, tracking changes in disease definitions or clinical coding practice through time and sharing clinical code information across platforms and data sources as research objects.

## Introduction

Over the last 20 years, increasing numbers of general practitioners have used computers to store patients medical records for various administrative functions [1]. Hospitals are also beginning to store their records electronically, though electronic records are far less prevalent than in primary care [2]. These electronic medical records (EMRs) offer great potential for research, enabling the rapid identification of patients for inclusion in intervention and observational studies. As their use becomes more widespread, it is becoming more important to have better means for ensuring and evaluating the validity of these studies. EMRs are being used by researchers to address important questions in healthcare that would be difficult or impossible to address using randomised controlled trials, because of the costs involved, the low prevalence of conditions or because a condition may occur in a subgroup such as children or pregnant women. In UK primary care in particular, the annual number of research outputs based on the three main UK primary care databases (The Clinical Practice Research Datalink (CPRD, formerly the General Practice Research Database, GPRD), The Health Improvement Network (THIN) and QResearch) appears to be increasing at an exponential rate (figure 1).

Much research has been done into establishing the internal and external validity of EMR studies [3], particularly from the point of view of data quality, data completeness and confounding. There has also been work replicating studies from one EMR database in another to assess their external validity; i.e. the

generalisability of their findings to other populations [4–6]. Where possible, PCD studies are also used to replicate randomised controlled trials and re-assess the effectiveness of interventions, in particular to patient groups that had been excluded in the original trials [7, 8]. Even if all of these issues are adequately addressed, most EMR studies still assume that the underlying definitions of clinical entities (such as conditions, treatments and monitoring tests) are valid. However, these clinical entities are defined through lists of 'clinical codes' and the process of preparing these code lists is rarely straightforward and can often lack rigor [9]. Despite the importance of clinical code list validity, and calls for sharing of code lists and greater transparency in the selection of code lists and reporting of sensitivity analyses using different sets of codes [10, 11], code lists are still seldom reported in published papers [3]. There is also currently no obligation on researchers to publish their code lists by funding bodies, journals or regulators. Furthermore, there is no centralised repository to hold lists of clinical codes. In most cases it is impossible to assess the validity of code lists used in EMR research.

There is also a gradual movement towards greater transparency and openness in academic research [12–14], sometimes driven by learned societies [15], particularly in disciplines where there is high computational load. There is also growing pressure from governmental organisations to share publicly funded research data [16, 17].

To address this transition towards full transparency, we developed www.ClinicalCodes.org, a web repository for EMR researchers to freely upload and download clinical code lists. Below we give an overview of the role and use of clinical codes in EMR research and provide details on the features of the ClinicalCodes repository.

## The role of clinical codes in EMR databases and research

Clinical entities in EMR databases are entered by medical professionals as clinical codes. In UK primary care, Read codes (named after Dr. James Read) are the most commonly used, while the ICD-9/10 system, which was adopted by the World Health Organisation, is more popular in UK hospital settings and primary care in North America and mainland Europe. Clinical codes form a hierarchical classification system for reporting and research purposes and can be used to define symptoms, signs and diagnoses, referrals to hospitals and clinics, immunisation records, prescribed medications and diagnostic test results.

The process of drawing up code lists to identify clinical entities of interest (e.g patients with a given clinical condition, patients on particular medications, patients with certain diagnostic test, smoking status etc.) is a critical step in setting up EMR studies and multiple code lists will often be required within one study to define multiple conditions, covariates, confounders and outcomes. This is often a complicated and time-consuming process that involves defining the clinical entity of interest and iteratively searching for codes in lookup tables, running searches for codes in different sections of the database, collating the results and classifying them (generally by clinically trained investigators) [9, 18]. The built in flexibility and redundancy of clinical coding systems allows practitioners to use a variety of codes to describe a given condition and minimises their time spent searching for codes, but it presents a challenge to researchers using these codes to effectively define a clinical entity. A patient may receive one of several possible diagnostic codes as well as, or instead of, one or more codes describing symptoms or investigations. This flexibility in the coding structure facilitates the clinical use of these codes and minimises the time spent searching for codes by practitioners. However, the multitude of codes for a given condition presents a challenge when data need to be aggregated. For example, it may be necessary to combine codes representing diagnoses, symptoms, prescribed drugs and diagnostic tests in order to accurately identify patients with a certain complicated condition. On the other hand, some entities can be identified with a very simple code list, or even a single clinical code [19].

The selection of codes used to identify conditions, comorbidities or other entities will vary according to the particular question being asked, partly reflecting the sensitivity (or specificity) required of the definition. Sometimes it is important to identify all possible cases but at other times more stringent

criteria may be needed to be adaopted in order to avoid misclassification of cases. This variability in code-lists may have major implications for results of all studies using EMRs [20]. For example, a sevenfold variation in estimates of incidence of rheumatoid arthritis can be largely explained by differences in code-lists [21, 22]. To account for such variation some studies have used different subsets of code-lists in sensitivity analyses [3, 23]. Furthermore, and in particular for uncommon diseases, small errors in code selection can result in large numbers of misclassified patients, leading to biased results and classification errors affecting conclusions in unpredictable ways [24]. Clinical definitions may also change over time, resulting in a need to revise the corresponding code list [10], a good example being a change in the UK Quality and Outcomes Framework (QOF) Business rules in 2006. When QOF was first introduced, people with diabetes were identified on the basis of any diabetes code, including non-specific diabetic type codes. From April 2006, the case definition for diabetes was restricted to only include codes that specified type I or type II diabetes [25]. In practice this meant that about 170 previously used Read codes were no longer being used to identify the condition. Finally, different researchers may have different interpretations of the relevance of particular codes.

## Reporting of codes in the current literature

A large component of total EMR research is made up by primary care database (PCD) studies and UK PCDs are among the most researched in the world. Figure 1 shows that research outputs with UK PCDs appears to be increasing at an exponential rate, while figure 2 shows that research using UK PCDs is being conducted in universities, pharmaceutical companies and research hospitals around the world, rather than being limited to the UK. As one of the largest and most important resources for EMR-based research, it seems reasonable to expect reporting of code lists in UK PCD-based studies to be at least as comprehensive as in other EMR studies. To evaluate levels of transparency in the reporting of clinical code lists, we took a representative sample of UK PCD studies and assessed each study on its extent of reporting of the clinical codes used.

We took a sample of 450 papers from the original 1359 identified from a PubMed search. Of these, 392 (87%) had both the full text accessible to the University of Manchester library and were examples of primary PCD research. Only 35 (9%) studies published the entire set of clinical codes needed to reproduce the study in an online appendix, while an additional 47 (12%) stated explicitly that the clinical codes are available upon request 1.

## The need for transparency in clinical code usage

We identify four main consequences of lack of transparency of clinical code lists. First, if code lists are not available and not expected to be published alongside the primary research using them, they represent an important part of a study methodology that is not subject to scrutiny or peer review. In the extreme case, there is no way to determine if a condition diagnosis in a study is valid and clinical decisions could be based on the invalid assumptions drawn from an invalid diagnosis. This could happen despite rigorous downstream statistical analysis. Second, the effective replication of EMR studies is dependent on the availability of the clinical codes in the original study. If all of the codes are not available, it is impossible to tell if differences found in study replications are due to artefactual differences in code lists or if they are genuine. Third, if code-lists are unknown, comparisons between studies on the same condition are potentially invalidated. Condition definitions change over time and GP coding practice may also change with respect to regulations and incentives [26]. Also, different studies may use different types of codes for a condition; some studies, for example, include medication and monitoring codes as part of their definition of a patient with diabetes (e.g. [27]) while others do not (e.g. [28]). Not having access to code-lists means that it is difficult to know whether fair comparisons are being made between studies. Fourth, building

code lists is a time consuming process; having access to historical code-lists for a condition would mean that new lists could be built incrementally and iteratively, saving much 'reinvention of the wheel' while increasing consistency, and potentially accuracy, of definitions across studies.

# The ClinicalCodes online repository

The main ClinicalCodes database consists of Articles containing metadata such as citation details and abstracts. Code lists are associated with these and individual clinical codes with the code lists. All individual clinical codes are assigned a code name, coding system (Read, OXMIS, SNOMED, CPRD_product_code, BNF_code, OXMIS, ICD-9, ICD-10), description and entity type (diagnostic, drug, test, clinical_sign, administrative, demographic, observation, immunisation). Users are able to upload supplementary fields for individual codes or add comments at the code list or article level. Code lists can be downloaded by any user but an account must be created to upload article metadata or code lists or to leave comments. Code lists can be downloaded individually as csv files. If a code lists from a previous article has been used verbatim in a new study, the ClinicalCodes entry for the new study can link to the previous code list. This reduces workload in uploading lists that are unchanged from previous studies while retaining information on the origin of code lists. At the time of publication, the complete code lists used for three papers from our group [6,23,28] as well as codes from the UK Quality and Outcomes Framework Business rules versions 5 and 24 - a total of 15193 clinical codes across 105 code lists covering medical conditions, lifestyle variables such as smoking status, physical observations such as BMI and testing (for example for retinal screening and blood sugar levels).

We have endeavored to make the upload and download processes as straightforward as possible. In particular, download of individual code lists is a one-click operation requiring no log in or provision of user information. The comments feature, which is available for articles and code lists, enables both the study authors to add extra methodological information and for other researchers to raise questions and issues on the code lists which could further assist the development of future code lists.

We have also developed an open-source R package [29] to automate the downloading and importing of clinical code lists from ClinicalCodes.

## Clinical Codes as research objects

Research objects are annotated aggregations of data often associated with a scientific publication that facilitate reuse and reproducibility of scientific research [30]. Following this model, metadata and links to code lists for articles are available as research objects that can be shared across platforms in machine readable form. In practice, this means that a JSON manifest file is available for each article containing: Article metadata (title, author, abstract, reference, link, doi), article level comments, code list level comments and links to the individual code list files. These research object files are available directly by adding a '/ro' to the URI for an article (e.g. www.clinicalcodes.org/medcodes/article/5/ro). The research object format is designed to be available without getting in the way of the main method of download that will be required by most users. The rClinicalCodes R package [29] enables the automated download of code lists and metadata via the research object file. An example JSON manifest file is shown in the online appendix.

# Conclusions

Large electronic medical datasets, including medical records datasets are already playing an important role in clinical research and this role is set to grow in the era of big data in healthcare [31]. The continuing success of big data in healthcare will depend on the ability of researchers to access and validate that data and then combine it with other sources [32]. We have developed a repository for clinical codes that will

be of great use to two groups of researchers. First, clinical researchers using primary care and other medical databases will be able to more effectively validate their research, build upon previous code lists and match appropriate disease definitions through time. Second, health informaticians will more easily be able to produce study replications (e.g. replications across databases such as [6]), share clinical code data as research objects across platforms and data sources and use the ClinicalCodes database as a research resource in its own right (e.g. to track changes in disease definitions and clinical coding practice through time).

Researchers using the ClinicalCodes repository can benefit from faster and more consistent development of new code lists, improvements in research quality associated with better scrutiny of lists of clinical codes, greater exposure and potential for studies with uploaded codes to be more highly cited and also from discovering other researchers working in the same area.

Despite these motivations, the success of this project will depend on its widespread adoption by the electronic medical records research community. Although ClinicalCodes solves the problem of having a centralised repository for holding codes, the problem remains that there are few, if any, requirements for researchers to make clinical code lists accessible. We believe that adoption and support of a centralised clinical codes repository by regulators, funding bodies and publishers of electronic medical records research will be of great benefit to the electronic medical records research community.

### Availability and Requirements

ClinicalCodes is freely accessable at http://www.clinicalcodes.org.

## Materials and Methods

### Article Classification

To get an estimate of the extent of the problem of lack of transparency in clinical code-lists in EMR studies, we collected articles conducting primary research using the three major UK-wide Primary care databases (PCDs) (The Clinical Practice Research Datalink, formerly the General Practice Research Database; The Health Improvement Network; QResearch). The UK has one of the most extensive and longest running systems of collection of EMRs and The main UK PCDs are the subject of considerable research interest. A Search was made on Pubmed for articles with the following terms: “CPRD”, “Clinical Practice Research Datalink”, “GPRD”, “General Practice Research Database”, “The Health improvement Network”, “QResearch” up until September 2013, returning 1359 articles. A random sample of 450 articles from this Pubmed search was taken for further analysis. From this sample, all articles were identified that were both primary EMR research and had their full text accessable via the University of Manchester library (392 articles). We then scored each paper as belonging or not to the following categories:

1. Any clinical codes listed in the methods section
2. At least one full code list provided in the paper or in an appendix
3. All code lists provided to enable replication of the study
4. States that “Code lists are available on request”

Analyses were performed using R v2.15.2 [33]. Article counts over time and geo-coded article affiliations (Via the Google Geocoder API) were aggregated using the R package rpubmed (https://github.com/ropensci/rpubmed).

### Database Architecture and Web Interface

The repository data is stored in a relational database called PostgreSQL (http://www.postgresql.org). Server-side web programming was done in Python v2.7.5 (http://www.python.org) using the Django v1.5 web framework (https://www.djangoproject.com). The client side scripting was done in JavaScript and HTML5 and used Twitter Bootstrap v3 (http://getbootstrap.com) as a front-end framework. The dynamic parts of the site were served using Gunicorn v18.0 (http://gunicorn.org) and static parts with Nginx v1.0.15 (http://nginx.org). Cacheing and sessions are handled by a Redis v2.4.10 NoSQL database (http://redis.io). The repository is hosted on a 64 bit Red Hat Enterprise Linux server release 6.4 virtual machine at the University of Manchester.

## Acknowledgments

We are thankful to Matt Ford for extensive technical support. Thanks to the Research team at CPRD for fruitful discussions in the development stage.

## Author Contributions

Conceived, designed and built the repository and software: DAS. Data collection: DAS, DR, EK, IO, RP, DA, EC. Data Analysis: DAS. Wrote the manuscript DAS. Edited the manuscript DAS, DR, EK, IO, RP, DA, EC.

## References

1. Purves IN (1996) The paperless general practice. BMJ 312: 1112-1113.

2. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, et al. (2009) Use of electronic health records in u.s. hospitals. New England Journal of Medicine 360: 1628-1638.

3. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ (2010) Validation and validity of diagnoses in the general practice research database: a systematic review. British Journal of Clinical Pharmacology 69: 4–14.

4. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P (2008) Performance of the qrisk cardiovascular risk prediction algorithm in an independent uk sample of patients from general practice: a validation study. Heart 94: 34-39.

5. Vinogradova Y, Coupland C, Hippisley-Cox J (2013) Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with qresearch and cprd data. BMJ 346.

6. Reeves D, Springate D, Ashcroft D, Ryan R, Doran T, et al. (2014) Can analyses of electronic patient records be independently and externally validated? the effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case-control analysis. BMJ open - submitted .

7. Tannen RL, Weiner MG, Xie D (2008) Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the prior event rate ratio to adjust for unmeasured confounding by indication. Pharmacoepidemiology and Drug Safety 17: 671-685.

8. Tannen RL, Weiner MG, Xie D (2009) Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ 338.

9. Davé S, Petersen I (2009) Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiology and Drug Safety 18: 704-707.

10. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, et al. (2009) Selection of medical diagnostic codes for analysis of electronic patient records. application to stroke in a primary care database. PLoS ONE 4: e7168.

11. Bhattarai N, Charlton J, Rudisill C, Gulliford MC (2012) Coding, recording and incidence of different forms of coronary heart disease in primary care. PLoS ONE 7: e29776.

12. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, et al. (2013) Why linked data is not enough for scientists. Future Generation Computer Systems 29: 599-611.

13. Stodden V, Guo P, Ma Z (2013) Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. PLoS ONE 8: e67111.

14. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, et al. (2013) Making research data repositories visible: The re3data.org registry. PLoS ONE 8: e78080.

15. The Royal Society (2012). Science as an open enterprise: The royal society science policy centre report 20/12. [Accessed 25 Nov. 2013].

16. The European Commission (2012). Commission recommendation on access to and preservation of scientific information. URL `http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf`. [Accessed 25 Nov. 2013].

17. Office of Science and Technology Policy (2013). Increasing access to the results of federally funded scientific research. URL `http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf`. [Accessed 25 Nov. 2013].

18. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, et al. (2013) Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: A strategy for developing code lists. PLoS ONE 8: e54878.

19. Kotz D, Simpson CR, Sheikh A (2011) Incidence, prevalence, and trends of general practitionerrecorded diagnosis of peanut allergy in england, 2001 to 2005. Journal of Allergy and Clinical Immunology 127: 623 - 630.e1.

20. Nicholson A, Tate AR, Koeling R, Cassell JA (2011) What does validation of cases in electronic record databases mean? the potential contribution of free text. Pharmacoepidemiology and Drug Safety 20: 321–324.

21. Garca Rodrguez LA, Tolosa LB, Ruigmez A, Johansson S, Wallander M (2009) Rheumatoid arthritis in uk primary care: incidence and prior morbidity. Scandinavian Journal of Rheumatology 38: 173-177.

22. Watson DJ, Rhodes T, Guess HA (2003) All-cause mortality and vascular events among patients with rheumatoid arthritis, osteoarthritis, or no arthritis in the uk general practice research database. The Journal of Rheumatology 30: 1196-1202.

23. Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, et al. (2011) Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the uk quality and outcomes framework. BMJ 342.

24. Manuel DG, Rosella LC, Stukel TA (2010) Importance of accurately identifying disease in studies using electronic health records. BMJ 341.

25. Hippisley-Cox J, O'Hanlon S (2006). Identifying patients with diabetes in the qof - two steps forward one step back. URL `www.bmj.com/cgi/eletters/333/7570/672-a`. Response to: Tanne J. Diabetes, not obesity, increases risk of death in middle age. BMJ 2006; 333: 672.

26. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N (2009) Effect of the quality and outcomes framework on diabetes care in the united kingdom: retrospective cohort study. BMJ 338.

27. Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, et al. (2006) Mortality in people with type-2 diabetes in the uk. Diabetic Medicine 23: 516-521.

28. Kontopantelis E, Springate D, Reeves D, Ashcroft DM, Valderas JM, et al. (2014) Withdrawing performance indicators: retrospective analysis of general practice performance under uk quality and outcomes framework. BMJ 348.

29. Springate DA (2014) rClinicalCodes: R tools for integrating with the www.clinicalcodes.org repository. Institute for Population Health, University of Manchester. URL `https://github.com/rOpenHealth/rClinicalCodes`.

30. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I (2010) Research objects: Towards exchange and reuse of digital knowledge. Nature Preceedings .

31. Wang SD (2013) Opportunities and challenges of clinical research in the big-data era: from rct to bct. Journal of Thoracic Disease 5.

32. Murdoch T, Detsky A (2013) The inevitable application of big data to health care. JAMA 309: 1351-1352.

33. R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

## Tables

## Figure Legends

**Table 1. Percentages of a random sample of UK primary care database studies with details of code lists**

| | Number of articles | Percentage |
|---|---|---|
| All UK PCD articles | 1359 | — |
| In random sample | 450 | — |
| Full-text available | 417 | — |
| Primary PCD research | 392 | 100 |
| Any code in methods | 104 | 26.5 |
| Any code list in study | 74 | 18.9 |
| All relevant code-lists | 34 | 8.7 |
| Any codes in paper | 117 | 29.8 |
| Codes available on request | 48 | 12.2 |
| Any codes or available | 138 | 35.2 |

Percentages are relative to the number of primary PCD research studies

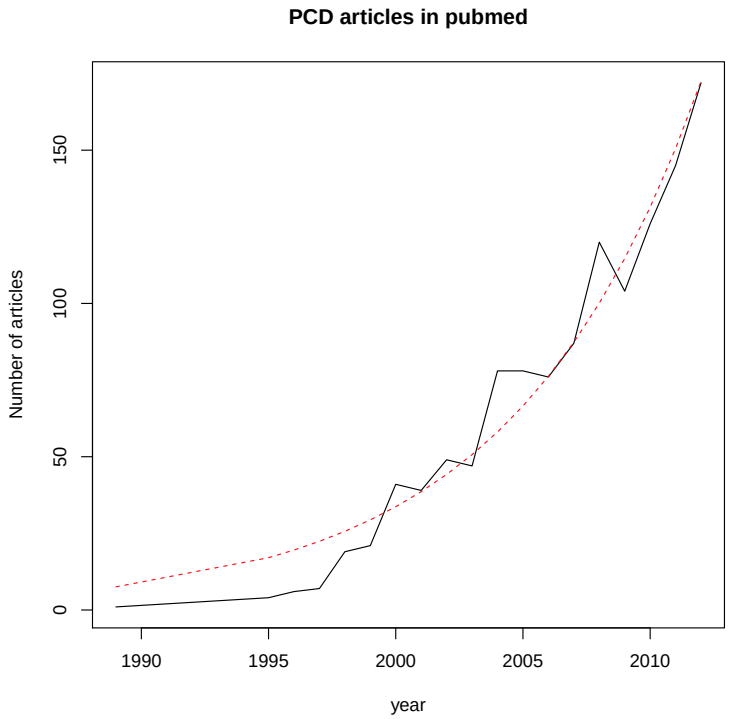


**Figure 1. Number of UK Primary Care Database publications.**

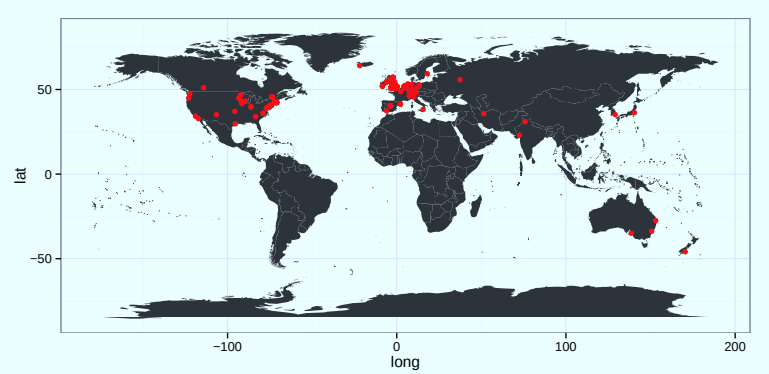


**Figure 2. Locations of primary affiliated departments.**