Uses and Validity of Primary Care Database studies

May 2013

David Springate, Evan Kontopantelis, Ivan Olier, David Reeves

Outline

1. Use of text-mining to explore the scientific literature

Outline

1. Use of text-mining to explore the scientific literature
2. Text-mining the PCD literature

Outline

1. Use of text-mining to explore the scientific literature
2. Text-mining the PCD literature
   - What is being studied using PCD's?

Outline

1. Use of text-mining to explore the scientific literature
2. Text-mining the PCD literature
   - What is being studied using PCD's?
   - Changes in topics of investigation over time

Outline

1. Use of text-mining to explore the scientific literature
2. Text-mining the PCD literature
   - What is being studied using PCD's?
   - Changes in topics of investigation over time
3. Validity of Clinical coding

Outline

1. Use of text-mining to explore the scientific literature
2. Text-mining the PCD literature
   - What is being studied using PCD's?
   - Changes in topics of investigation over time
3. Validity of Clinical coding
4. ClinicalCodes.org : A new repository for clinical code lists

Text mining

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction
    - Automatically extracting structured information from unstructured text

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction
   - Automatically extracting structured information from unstructured text
2. Semantic searching

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction
   - Automatically extracting structured information from unstructured text
2. Semantic searching
   - Improves search accuracy by including context into a search

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction
   - Automatically extracting structured information from unstructured text
2. Semantic searching
   - Improves search accuracy by including context into a search
3. Knowledge discovery

What is it?

- The process of extracting high-quality structured information from unstructured text (e.g. Scientific literature).
- Uses a variety of computational and statistical methods to find patterns and trends in text

Text mining consists of:

1. Information extraction
   - Automatically extracting structured information from unstructured text
2. Semantic searching
   - Improves search accuracy by including context into a search
3. Knowledge discovery
   - Identifying relationships in extracted data

Why do we need it?

- The scientific literature is rapidly increasing in size

Why do we need it?

- The scientific literature is rapidly increasing in size
- Humans can't keep up to date with the literature

Why do we need it?

- The scientific literature is rapidly increasing in size
- Humans can't keep up to date with the literature
  - 75 trials and 11 Systematic reviews published per day! Bastian et al. (2010) PLoS Medicine

Why do we need it?

- The scientific literature is rapidly increasing in size
- Humans can't keep up to date with the literature
  - 75 trials and 11 Systematic reviews published per day! Bastian et al. (2010) PLoS Medicine
- It is increasingly difficult to hone in on relevant papers

Why do we need it?

- The scientific literature is rapidly increasing in size
- Humans can't keep up to date with the literature
  - 75 trials and 11 Systematic reviews published per day! Bastian et al. (2010) PLoS Medicine
- It is increasingly difficult to hone in on relevant papers
- More of the literature is being held online in machine-readable archives

Why do we need it?

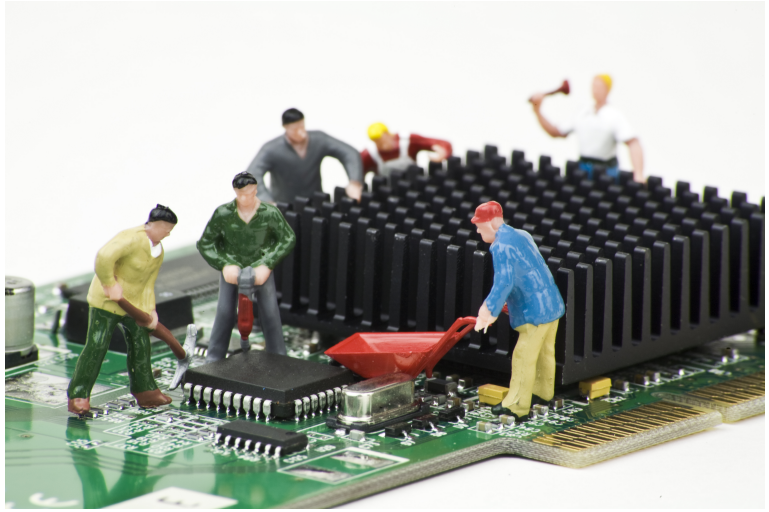
- The scientific literature is rapidly increasing in size
- Humans can't keep up to date with the literature
  - 75 trials and 11 Systematic reviews published per day! Bastian et al. (2010) PLoS Medicine
- It is increasingly difficult to hone in on relevant papers
- More of the literature is being held online in machine-readable archives
- TM can reduce processing time for systematic reviews by 80% (NCTM)

Text-mining is not a magic bullet

- Many publications are not open access
  - Often need to rely on abstracts
  - Grey literature is often inaccessable

Text-mining is not a magic bullet

- Many publications are not open access
  - Often need to rely on abstracts
  - Grey literature is often inaccessable
- Still need plenty of human input!
- TM algorithms can be very complex
- Breadth at the expense of depth



Text mining the PCD literature

UK Primary Care Databases

GPRD / CPRD

The General Practice Research Database / The Clinical Practice Research Datalink

- ~ 900 papers

THIN

The Health Improvement Network

- ~ 360 papers

QResearch

- ~ 75 papers

The Dataset

- All articles reported by CPRD, THIN, QResearch in Pubmed

The Dataset

- All articles reported by CPRD, THIN, QResearch in Pubmed
- 1185 Abstracts with metadata

The Dataset

- All articles reported by CPRD, THIN, QResearch in Pubmed
- 1185 Abstracts with metadata
- 141 full-text articles for validation

The Dataset

- All articles reported by CPRD, THIN, QResearch in Pubmed
- 1185 Abstracts with metadata
- 141 full-text articles for validation

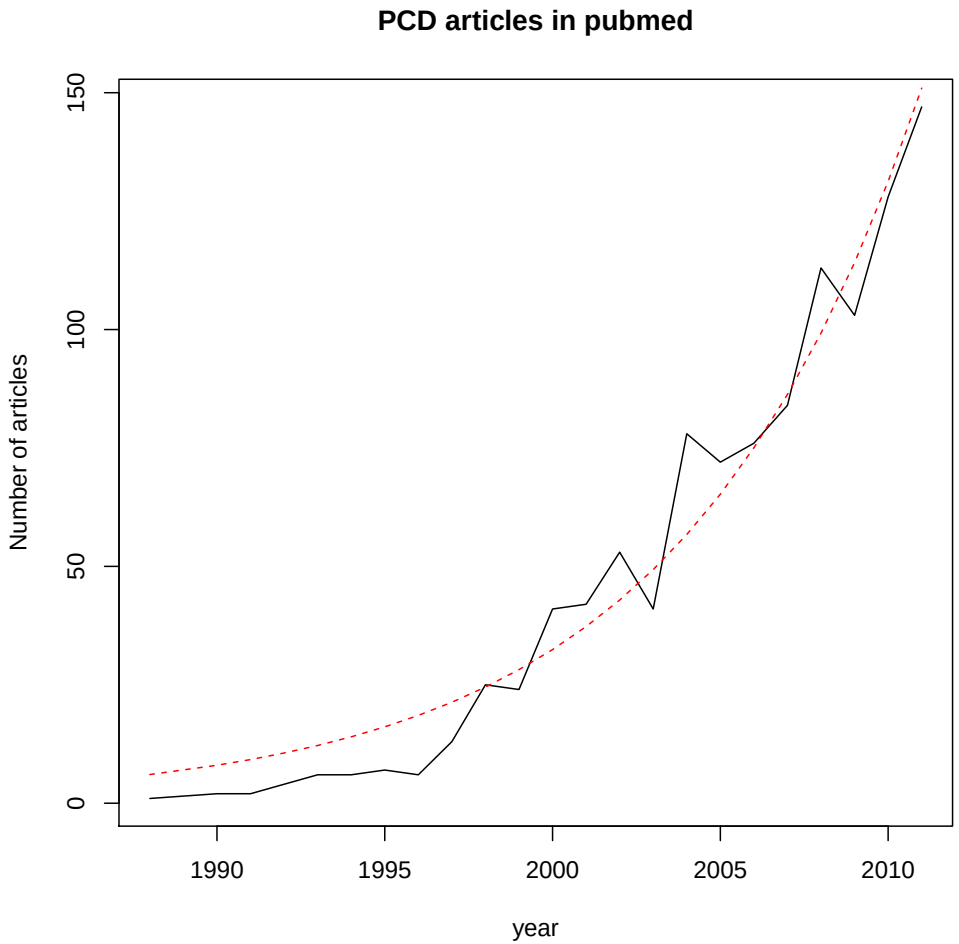
The Dataset

- All articles reported by CPRD, THIN, QResearch in Pubmed
- 1185 Abstracts with metadata
- 141 full-text articles for validation

How are PCD's being used by researchers?

PCD studies are a growth area!

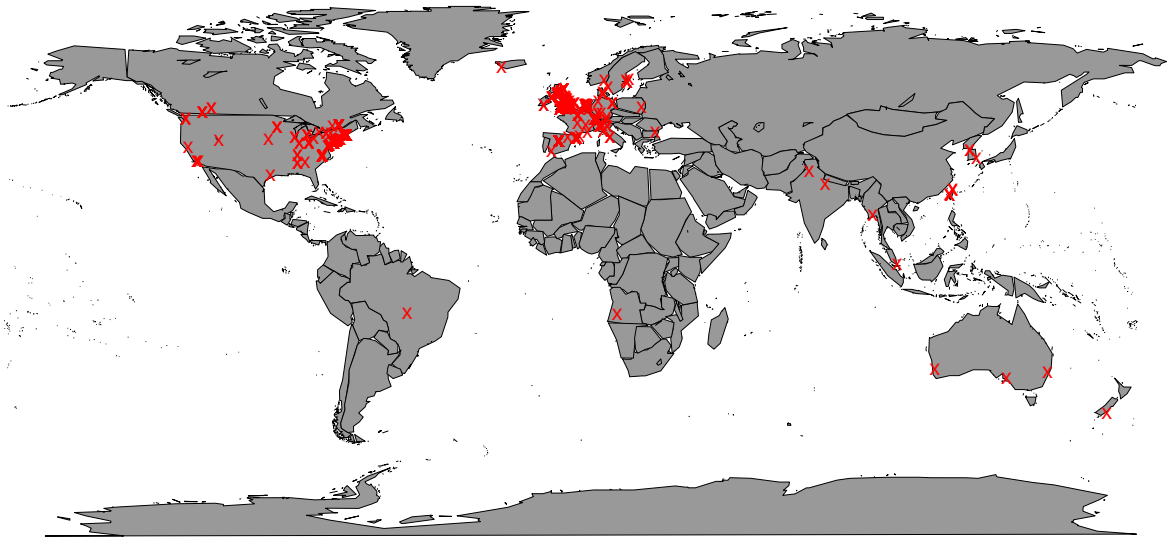
Number of publications is rapidly increasing. . .



PCD studies are a growth area!

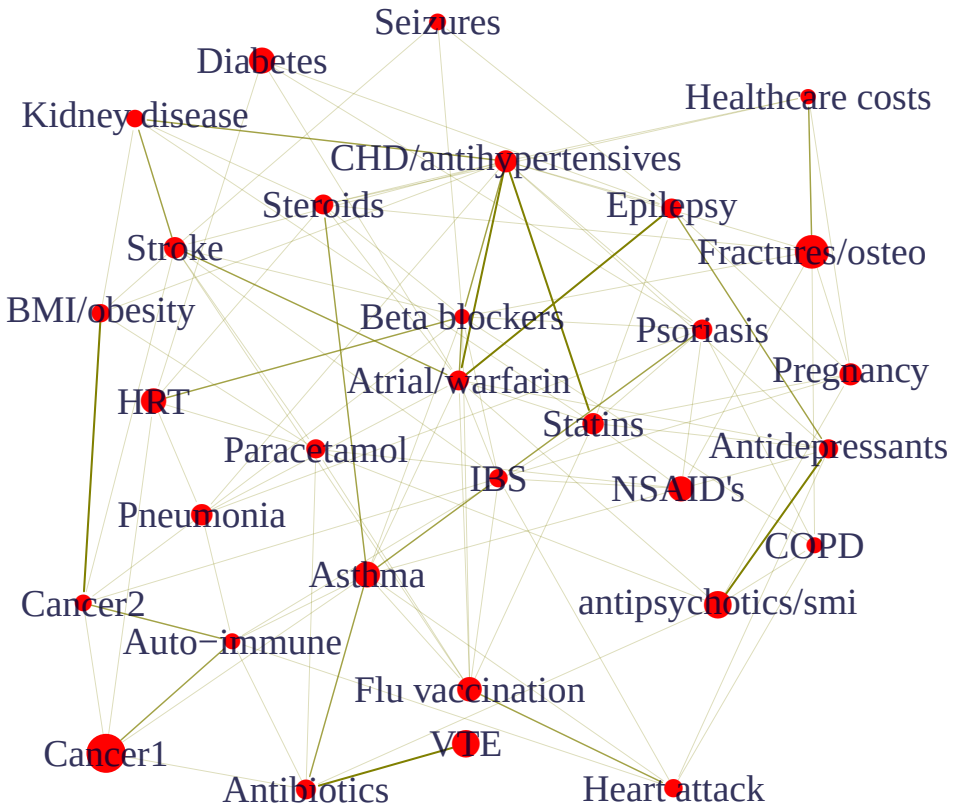
. . . and there is global interest in UK PCD research

Institutions affiliated with UK PCD publications

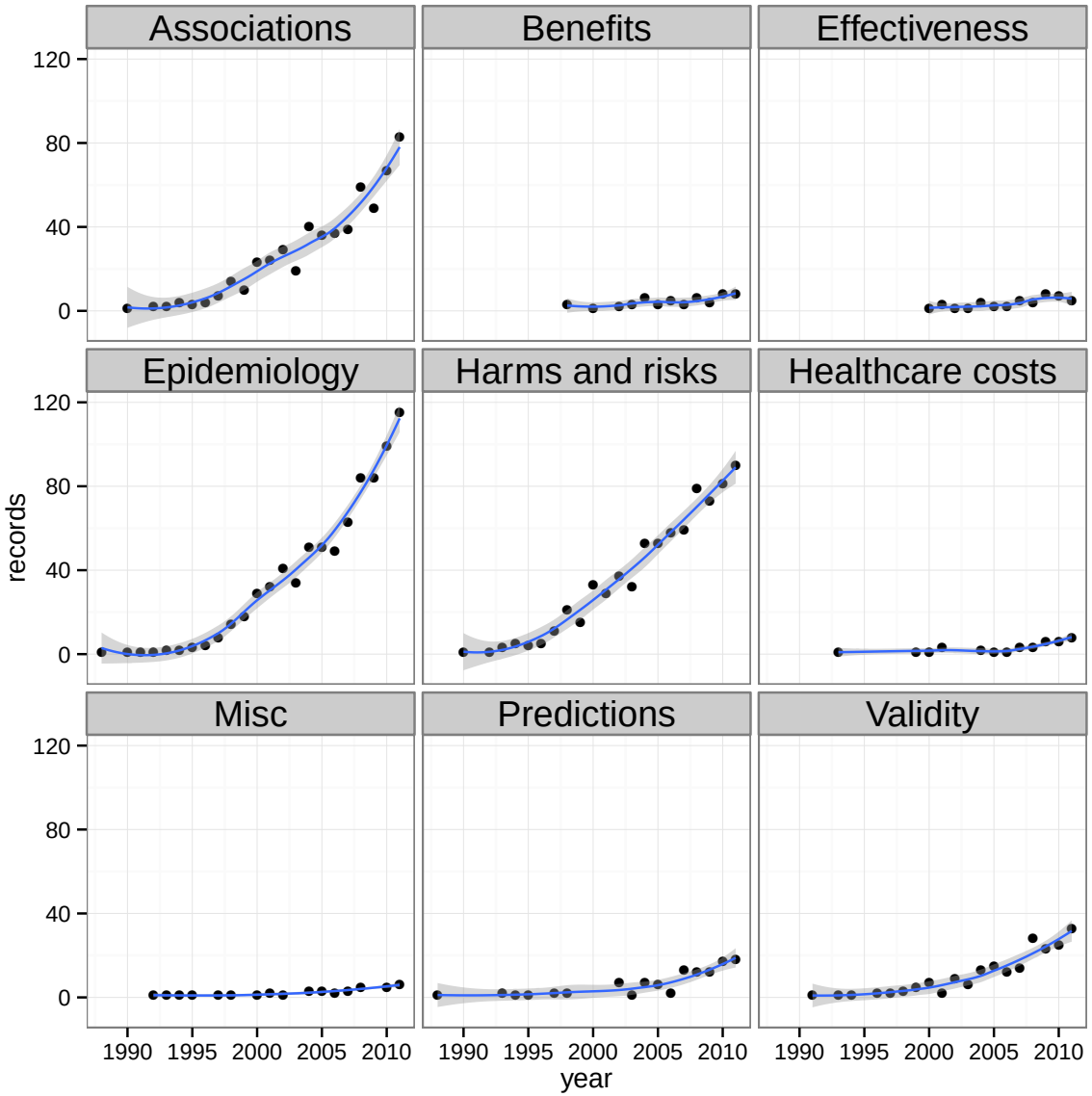


Broad scope of topics in PCD studies

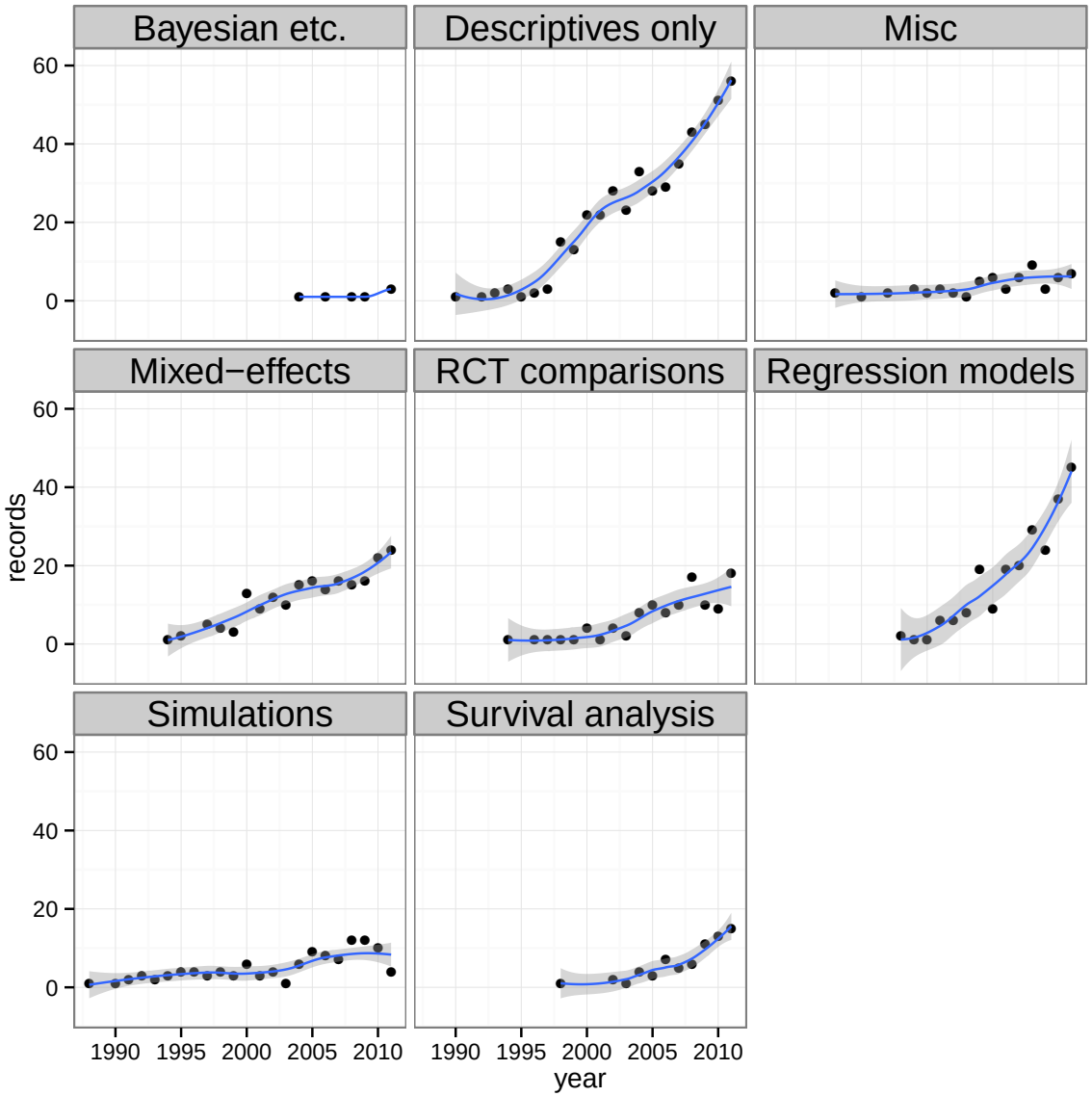
A network graph of PCD topics of investigation



Study types are changing. . .



. . . as are analysis methods



PCD validity

Threats to validity

- Unmeasured confounding

Threats to validity

- Unmeasured confounding
- Correlation does not equal causation

Threats to validity

- Unmeasured confounding
- Correlation does not equal causation
- GP recording

Threats to validity

- Unmeasured confounding
- Correlation does not equal causation
- GP recording
- Clinical coding

Clinical Coding in PCD's

- All clinical events are entered by GP's as clinical codes:

Clinical Coding in PCD's

- All clinical events are entered by GP's as clinical codes:

- Symptoms, signs & diagnoses (READ codes)
- Referrals to external care centres
- Immunisation records
- Prescription information
- Diagnostic test records and results

Clinical Coding in PCD's

- All clinical events are entered by GP's as clinical codes:

- Symptoms, signs & diagnoses (READ codes)
- Referrals to external care centres
- Immunisation records
- Prescription information
- Diagnostic test records and results

- Everything recorded by a GP can be identified (if you know which codes to look for and where to look for them!)

Clinical Coding in PCD's

- All clinical events are entered by GP's as clinical codes:

- Symptoms, signs & diagnoses (READ codes)
- Referrals to external care centres
- Immunisation records
- Prescription information
- Diagnostic test records and results

- Everything recorded by a GP can be identified (if you know which codes to look for and where to look for them!)

e.g.

- H331.00 - Asthma diagnosis
- H33z011 - Severe asthma attack
- 33G1 - Spirometry testing

Clinical codes in PCD studies

Diagnoses are made by reference to a set of clinical codes

Workflow

1. Researchers decide on a rough set of codes for a condition
   - By searching lookup tables for matching terms
   - By reference to an external source (e.g. QOF)
2. Clinicians go through this draft list by hand and select the relevant codes
3. The database is searched for events matching the finalised code list
4. The correct combination of events in the timeframe of interest gives a diagnosis
   - e.g. For Asthma: Need at least 1+ clinical event 1+ drug event in the last year to qualify

Code list? What code list?

- Currently no obligation to publish code lists
- No centralised repository for clinical codes
- The vast majority of PCD studies do not publish their codes
- No way of knowing if a condition diagnosis is valid
- No way to replicate the research

For example. . .

In 45 UK case-control PCD studies (diabetes):

Code list? What code list?

- Currently no obligation to publish code lists
- No centralised repository for clinical codes
- The vast majority of PCD studies do not publish their codes
- No way of knowing if a condition diagnosis is valid
- No way to replicate the research

For example...

In 45 UK case-control PCD studies (diabetes):

- Only 5 reported ANY clinical codes...

Code list? What code list?

- Currently no obligation to publish code lists
- No centralised repository for clinical codes
- The vast majority of PCD studies do not publish their codes
- No way of knowing if a condition diagnosis is valid
- No way to replicate the research

For example...

In 45 UK case-control PCD studies (diabetes):

- Only 5 reported ANY clinical codes...
- Only 2 of these published codes in appendix

Code list? What code list?

- Currently no obligation to publish code lists
- No centralised repository for clinical codes
- The vast majority of PCD studies do not publish their codes
- No way of knowing if a condition diagnosis is valid
- No way to replicate the research

For example...

In 45 UK case-control PCD studies (diabetes):

- Only 5 reported ANY clinical codes...
- Only 2 of these published codes in appendix
- Only 1 provided full set of code lists

Validity of Clinical coding

Clinical codes should be held to scrutiny and peer-review (either pre- or post-publication)

This would allow for:

- replication of studies

Validity of Clinical coding

Clinical codes should be held to scrutiny and peer-review (either pre- or post-publication)

This would allow for:

- replication of studies
- validation of diagnoses

Validity of Clinical coding

Clinical codes should be held to scrutiny and peer-review (either pre- or post-publication)

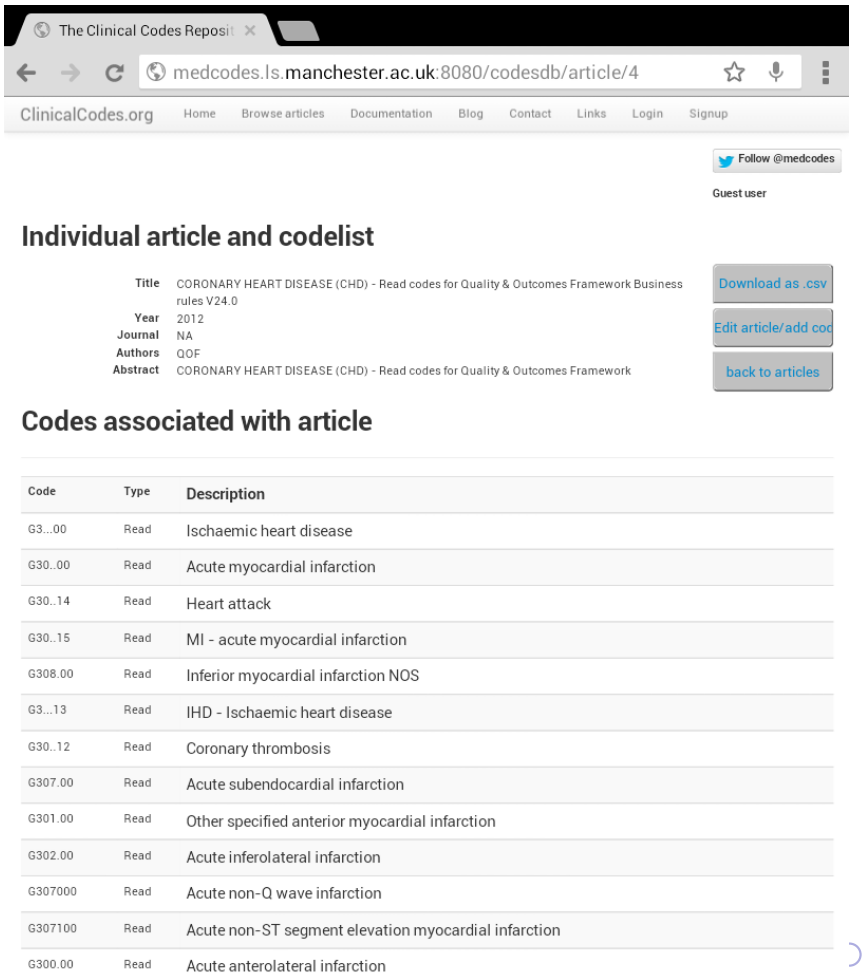
This would allow for:

- replication of studies
- validation of diagnoses
- incremental improvements to clinical definitions

ClinicalCodes.org

. . . Is an online repository for PCD researchers to upload their codes upon publication.

- Deposit code-lists for published studies
- Download historical code-lists
- Archive for all Quality and Outcomes Framework business rules (2004 - current)
- Database-specific information (e.g. consultation types)



ClinicalCodes.org

- Allows for validation / replication of PCD studies
- Tracking of disease definitions through time
- Comparitive studies of clinical codes

Don't reinvent the wheel!

Currently in development on campus:

medcodes.ls.manchester.ac.uk:8080/codesdb

Summary

- Publish open access!

Summary

- Publish open access!
- Upload your codes!

Summary

- Publish open access!
- Upload your codes!
- Thank you