Data & Sample Repositories
How are archived data sets being used?

Elizabeth C. Wright,
Kenneth J. Wilkins, Rebekah S. Rasooly
NIDDK/NIH
SCT May 18, 2015

Outline

- Data sharing and data repositories
  - Institute of Medicine report on data sharing
  - Models for data sharing
- The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) data and sample repository
- Review of requests for data from the NIDDK repository
  - Methods
  - Results
  - Conclusions

IOM Data Sharing Report (Jan 2015)

- Investigators, sponsors, and other stakeholders in clinical research should “foster a culture in which data sharing is the expected norm”
- Benefits:
  - Advance scientific discovery and improve clinical care
  - Maximize knowledge gained from data collected in clinical trials
  - Stimulate new ideas for research
  - Avoid unnecessarily duplicative trials
- Risks, burdens and challenges:
  - Protect the privacy of participants
  - Protect the investment of funders, sponsors and investigators
  - Protect the validity of analyses

Models for data sharing

- NIH – Grant applications with a direct cost of ≥$500,000 per year must have a data sharing policy
  - NIDDK and NHLBI have data repositories that receive de-identified data from DCCs and release data to approved investigators, subject to signed data use agreements
- Industry – Data are made available on a password protected web site. Users sign a data use agreement and have a private work space. Patient level data may not be downloaded. Statistical programs (SAS, R) are made available.
  - ClinicalStudyDataRequest.com – Bayer, Boehringer Ingelheim, GSK, Lilly, Novartis, Roche, Sanofi, Takeda, UCB, ViiV Healthcare
  - Yale University Open Data Access (YODA) Project – a partnership between Yale and pharmaceutical companies – Medtronic, Janssen/Johnson & Johnson

NIDDK Central Repositories

- 3 repositories have been funded by contract since 2003:
  - Data Repository: Information Management Services, Calverton, MD; Previously RTI, Research Triangle Park, NC,2003-June 2013
  - Biosample Repository: Fisher BioServices, Rockville, MD
  - Genetics Repository: Rutgers, The State University of New Jersey, Piscataway, NJ
- According to NIDDK policy, the schedule for data availability is for:
  - Major publications: within 6 months of publication
  - Baseline data: within 2 years after enrollment is completed
  - Entire study dataset: 2 years after the data are locked for analysis at the completion of an intervention.

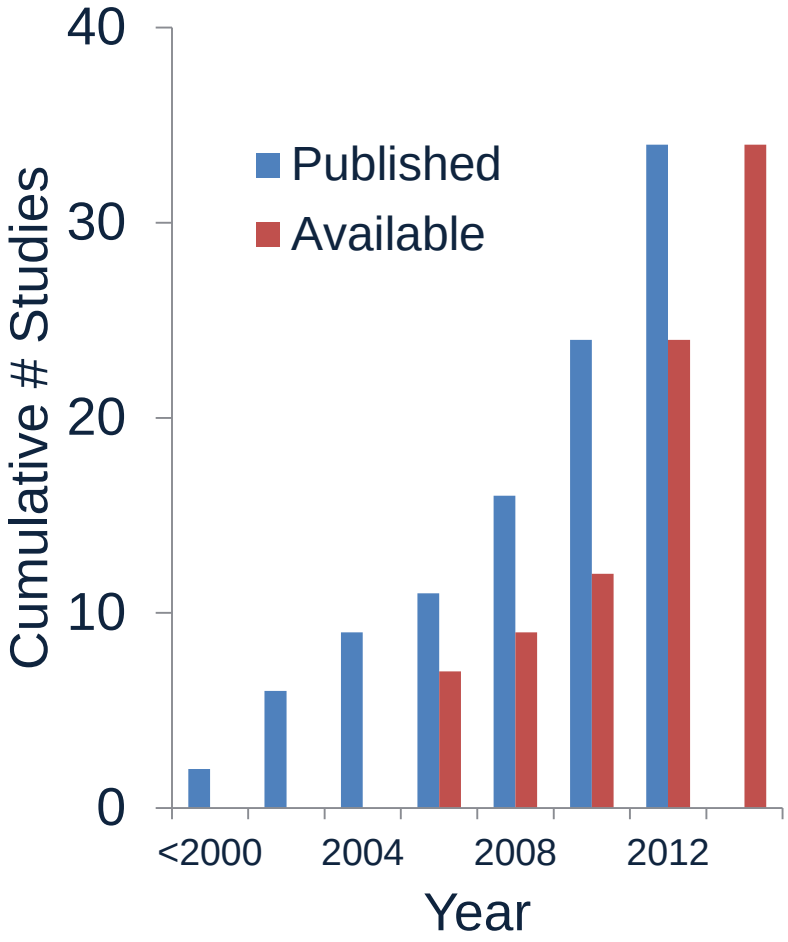
NIDDK data curation procedures

- Study DCC creates a de-identified data set and sends it to the repository
  - Includes both form data and analysis data sets
  - Data are de-identified by assigning new ID variables, removing any other ID variables, and converting dates to days from an index date
- Forms, MOP, data set documentation, and relevant publications are sent to the repository.
- Data Set Integrity Check
  - Repository verifies that published results from the study can be reproduced using the archived data sets
  - Performs a small number of analyses to provide confidence that the data set is a true copy of the study data
  - Does not attempt to resolve minor discrepancies with published results

Methods for this review

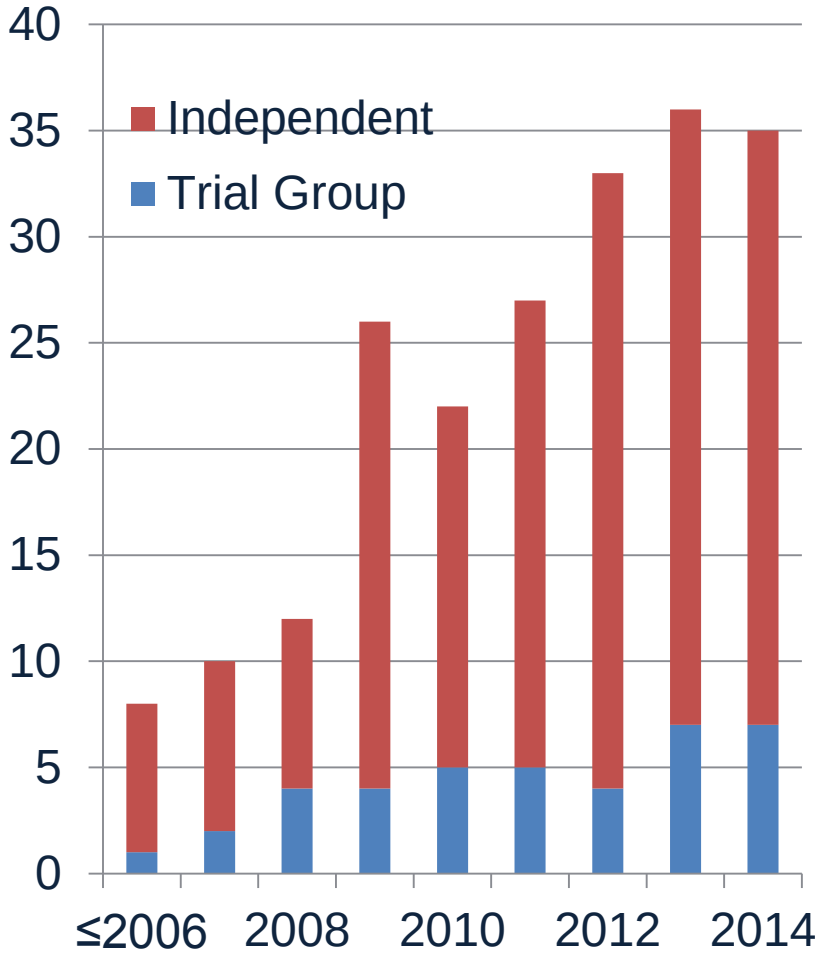
- Lists of available data sets and requests for data were obtained from the NIDDK Central Repository website
- Executive summaries were obtained from the website if available or directly from NIDDK.
- The primary aims and the relationship between the requester and study investigators were coded.
- Requests for multiple studies by a single investigator were counted separately (requests) and then combined (projects).
- PUBMED was searched for publications resulting from these projects.

34 trials with data as of November 2014



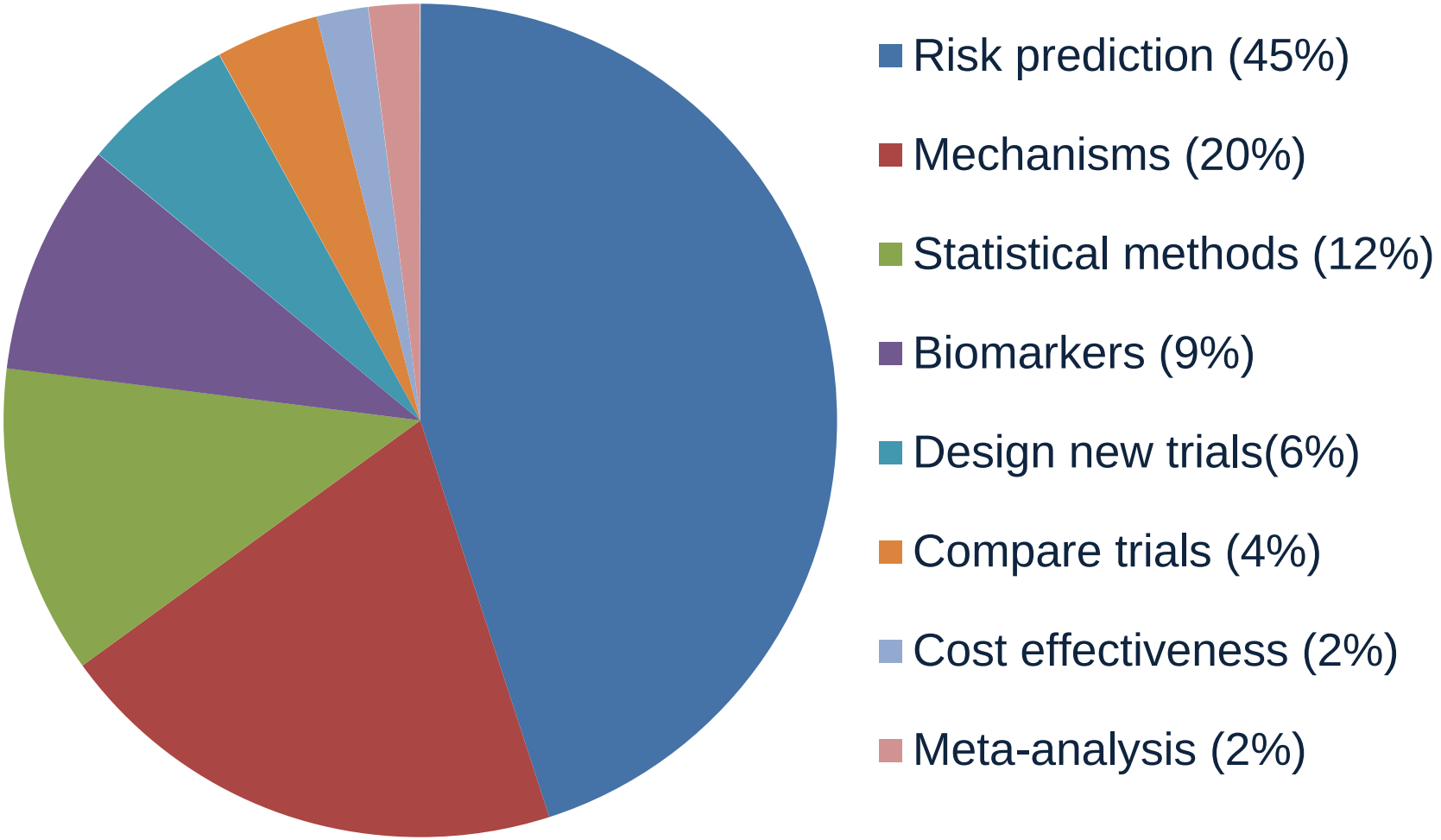
- Data are available for 34
  - 5 have ≥20 requests (DCCT/EDIC, DPP, AASK, MDRD, HEMO)
  - 5 have 5-19 requests
  - 15 have 1-4 requests
  - 9 have had no requests
- Data were available within 2 years for 58% of trials published after 2005
- 12 more trials are expected to have data available in 2015

209 requests as of January 2015



- 81% independent investigators
- 87% from medical center or university
- 85% located in USA
- 22% requested data from multiple studies
- 12% requested both data and samples

Primary aims of 209 requests



Conclusions

- Although data should be available within 6 months of the primary publication, most were not available within this time limit
- Time to release is decreasing for newer trials
- Preparing data sets for release requires considerable work by DCC and the repository
- Most data requests were from independent investigators
- Almost all were analyses of data not included in the original study publications
- Publications were identified for 29% of requests made before 2012; some have resulted in multiple publications