Phenotype algorithm and data source reporting in top clinical journals: where we are and where should we go?

PRESENTER: **Anna Ostropolets**

BACKGROUND

Despite multiple quality standards existing out there, we know little about the real requirements for phenotype standards and data source reporting in clinical journals.

METHODS

We searched the top clinical journals (Lancet, BMJ, JAMA and JAMA Internal Medicine, NEMJ, Circulation, Nature Medicine and Annals of Internal Medicine) for recent observational studies.
We extracted and analyzed 5 papers per journal.

RESULTS

57.5% of papers relied on previously published phenotypes even if the latter were not validated.
Only 11% of original created de-novo phenotypes were validated.

Table 1. Distribution of papers based on implemented phenotype algorithm

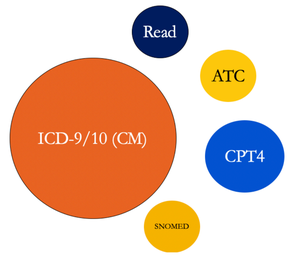
	Validated	Non-validated	Total
Original	2	15	17
Re-used	18	5	23
Total	20	20	40

Most of the published studies produced non-validated phenotypes and inconsistently described data sources

Take a picture to download the full paper

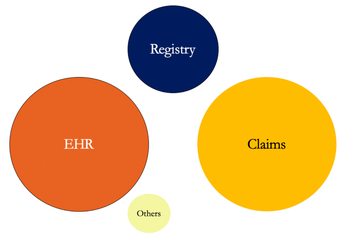
Most of the studies used ICD-9/10 codes along with free text medication records

.Figure 1. Ontologies used in phenotypes.



EHR and administrative claims were the most common data source types. Only 5% of the studies were performed on multiple data sources.

Figure 2. Distribution of types of data sources



The description of the data sources lacked structure and varied in the level of detail.

Only new data sources had rigorous descriptions, including information about data gathering process, data content and quality assurance.

Anna Ostropolets, RuiJun Chen, Matthew Spotnitz, Runsheng Wang, Patrick Ryan, George Hripcsak

COLUMBIA
COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

OHDSI