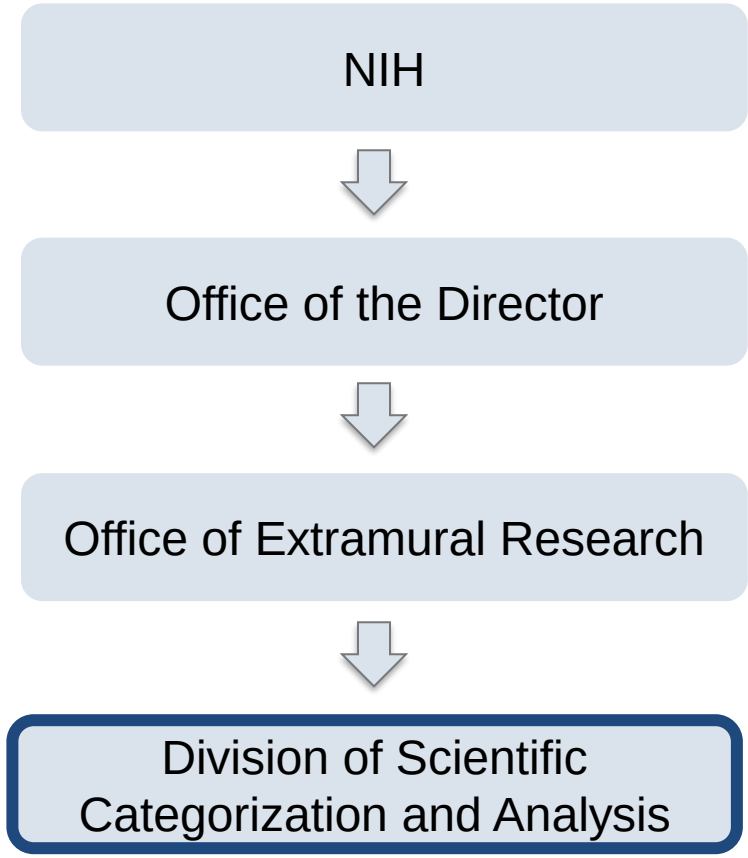
NIH's Research, Condition, and Disease Categorization (RCDC) System

Division of Scientific Categorization and Analysis, Office of Extramural Research, Office of the Director, NIH
Evelina Cebotari and Nancy Praskievicz
January 25, 2024

Agenda

- What is RCDC?
- History and purpose of the RCDC system
- RCDC methodology
- Reporting considerations
- Public tools

Division of Scientific Categorization and Analysis



- The Division of Scientific Categorization and Analysis (DSCA) curates and maintains the research, condition, and disease categorization (RCDC) research portfolios for official public reporting and analysis.

- Scientific information analysts (SIAs), data analysts and computational linguists make up DSCA.

Research, Condition, and Disease Categorization

NIH CATEGORICAL SPENDING

Fiscal Year	Public Categories
2022	315
2023	+10

Annual support level for these categories: NIH RePORT Categorical Spending website

- Developed as a requirement of the NIH Reform Act of 2006 to uniformly code research grants and activities at all Institutes and Centers (ICs)
- DSCA works with IC stakeholders to create and maintain categories for public consumption and internal NIH analysis
- Categories are not mutually exclusive, and projects can fall into multiple categories

RCDC Reporting Benefits

- Official source of categorical reporting
  - NIH-wide consensus on scientific project listings
  - Consistent budgetary methodology
  - Transparent public source that reflects NIH appropriations
- Automated process
  - Curated, centralized thesaurus describing a hierarchy of concepts
  - Text mining, indexing, and categorization
- Standardized project level reporting
  - Grant, Intramural, Inter-Agency Agreement, Contract projects, Inter Departmental Delegations of Authority (IDDA) and Other Transactions of Authority (OTA)
  - Multi-component grants reported at subproject level (e.g., P01s and P50s)

How do we define RCDC categories?

- NIH-wide subject matter experts (SMEs) discuss scientific areas of inclusion, achieving consensus on what is reported in each category

- SMEs also provide project-level validation permitting analysts to monitor category performance and to further refine the project listing

RCDC NIH-wide Category Development and Maintenance Process

DSCA Review

Analyze new and existing categories

In-depth analysis by DSCA staff

Thesaurus updates

From 1 day to 4 weeks

Sessions

Conduct Sessions as needed w/ ICs

DSCA facilitates NIH-wide discussions w/ SMEs

1 – 4 Meetings

SME and DSCA Review

ICs and DSCA Review Data

ICs review individual projects within categories

Review for glaring errors

April - August

Final Review

End of Year Data Review

Review project list and total expenditures

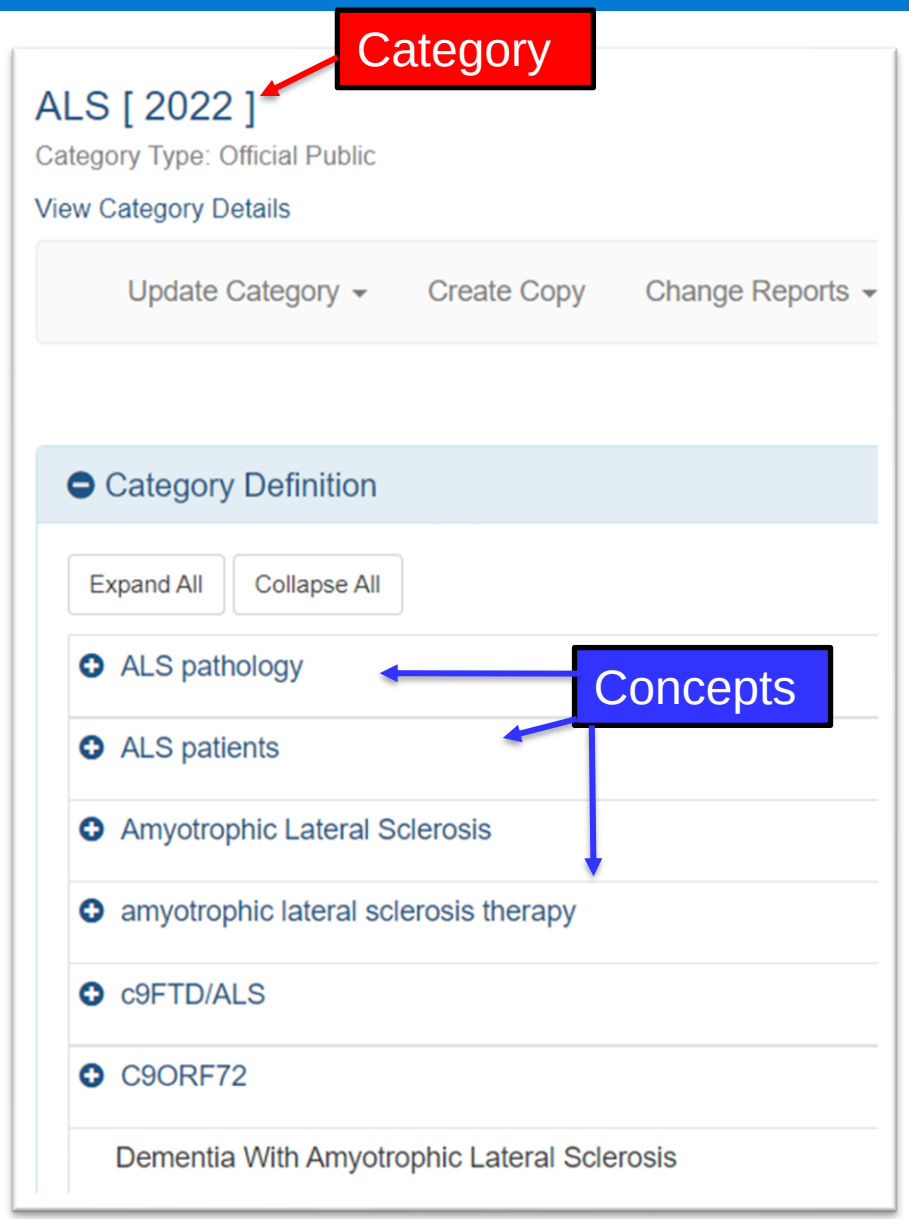
Documentation and final DSCA review

Post-Validation and Data Freeze

Concepts vs. Categories

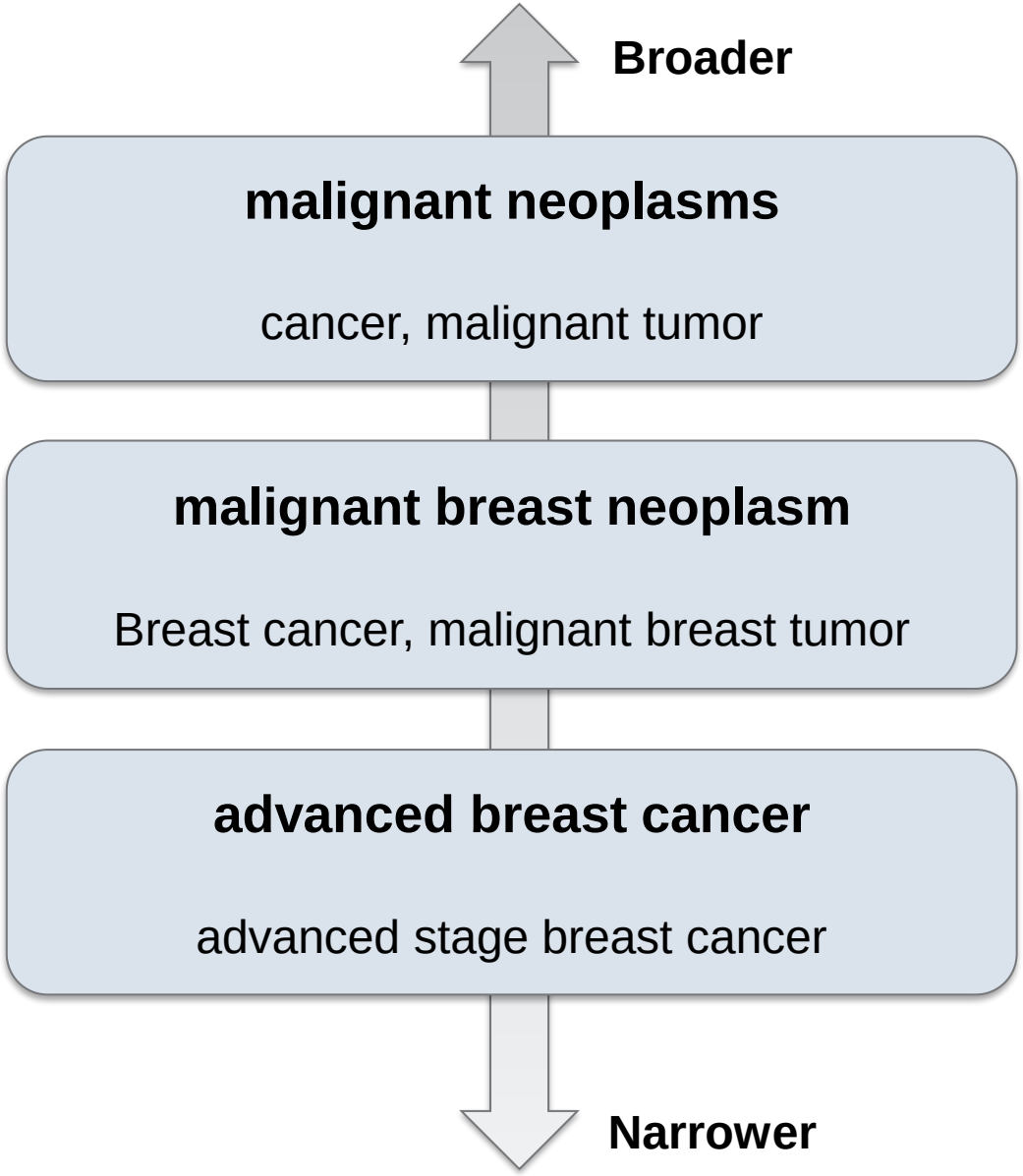
- **Categories** produce project listings for reporting NIH expenditures on:
  - Diseases
  - Conditions
  - Research Areas (scientific or budgetary)

- Categories contain a list of weighted **concepts** from the RCDC Thesaurus that are highly relevant to that category.
  - Known as a "**category fingerprint**"



RCDC Thesaurus

- Backbone of the RCDC system
- Concepts added, removed, and modified as science and language evolves
- Can have synonyms
  - "cancer", "malignant tumor" are synonymous with "malignant neoplasm"
- Adjustments analyzed for potential impact across all categories



RCDC Thesaurus

- Natural Language Processing (NLP) logic applied to concepts based on context
  Some examples include:
  - Word order – “nursing home” vs. “home nursing”
  - Punctuation – “birth control” vs. “birth, control”
  - Case sensitivity – “AIDS” vs. “aids”

- Normalization of text
  - Unicode (α = alpha)
  - Spelling differences (tumour = tumor)
  - Removing suffixes (researching -> research)

- Concept ambiguity
  - Add specific descriptive words to supplement intent and provide context
    - Operation: *surgical* operation; *tactical* operation; *mathematical* operation
    - Nursing: Nursing *student*, nursing *degree*, *infant* nursing

RCDC Indexing Process

Project Text

- Title
- Abstract
- Public Health Relevance
- Specific Aims (not public)

Text Mining & Processing → Concept Identification → Calculate Frequency

Thesaurus Concepts

Project Index

Concept Name	Project Index
Alzheimer's Disease	89.872%
abeta oligomer	39.223%
Apolipoprotein E	100.000%
Alzheimer's disease risk	27.735%
Amyloid	70.711%
abeta accumulation abeta aggregation amyloid beta accumulation amyloid beta aggregation amyloid β accumulation amyloid β aggregation aβ accumulation aβ aggregation	19.612%
abeta toxicity	19.612%
Amyloid beta-Protein	19.612%
Senile Plaques amyloid beta plaque Amyloid Plaques amyloid-b plaque cored plaque diffuse plaque Neuritic Plaques	33.968%

Automated Categorization

Categorization compares the **project index** to the **category fingerprint**

Category Fingerprint

2017 Category Definition [ALS]	View Category Details
Amyotrophic Lateral Sclerosis	100
C9orf72	100
Dementia With Amyotrophic Lateral Sclerosis	100
Familial Amyotrophic Lateral Sclerosis	100
frontotemporal lobar dementia-amyotrophic lateral sclerosis	100
Motor Neuron Disease	65

Project Matching

Concept Name	Category	Project Index	Match
Amyotrophic Lateral Sclerosis	100 %	100.000 %	100.000 %
C9ORF72	100 %	100.000 %	100.000 %
Familial Amyotrophic Lateral Sclerosis	100 %	21.567 %	21.567 %
Motor Neuron Disease	65 %	30.500 %	19.825 %

Project IS categorized

Category Threshold

Project is NOT categorized

Match Score: Calculated by comparing project index to category fingerprint

Categorized projects: Projects with match scores above the threshold

Reporting Methods

Automated

Standard categorization method that uses text mining, category logic, and metadata business rules

Nearly all RCDC categories

Subject Matter Experts validate to help DSCA staff curate project listings

Dollars are reported at the full project amount (100%)

Manual

Only for special appropriations or policy requirements

Only 3 RCDC categories

Subject Matter Experts determine relevancy of projects

Dollars can be prorated (1-100%)

RCDC Over Time

- Underlying methodology has not changed
- Standardized processes and increased data quality checks
- Incorporated advancements in machine learning and natural language processing
- Overall increased specificity and accuracy

Reporting a Project in Multiple RCDC Categories

- RCDC categories are not meant to be a sum of the total NIH expenditure

- Example: A project conducting a clinical trial to treat cachexia in breast cancer patients would be reported in the following RCDC categories:
  - *Clinical Trials and Supportive Activities*
  - *Cachexia*
  - *Breast Cancer*
  - *Cancer*
  - *Women's Health*

- How could one go about splitting up funding in these situations in a consistent and reproducible manner? What happens to a multi-year project when a new related RCDC category is created?

Why don't we report on topic X?

- Topics reported on the categorical spending page do not represent the extent of NIH-funded research

- New RCDC categories are created as they are requested based on public interest and evolving research
  - e.g., *Coronaviruses, Machine Learning and Artificial Intelligence*

- We are continuously adding new reporting categories
  - *Polycystic Ovary Syndrome (PCOS)*: **New for FY22 dataset**
  - *Menopause*: **New for FY23 dataset (forthcoming)**

How is an RCDC category requested?

- Reach out to the related NIH IC stating why a category of interest should go through the process
- Support from an NIH IC Director or Congressional language helps justify the request and a working group within NIH decides if it should proceed though the official process
- Approved categories are prioritized based on urgency and number of requests

RCDC Tools

- RCDC has several internal tools ICs can use to further analyze their NIH appropriation portfolio (*e.g.*, identify trends and gaps in research funding)
  - Clustering by topics, organizations, project metadata
  - Intersect and non-intersect reports
  - Multi-year trend analysis for one or more categories
  - Creating custom categories for internal use

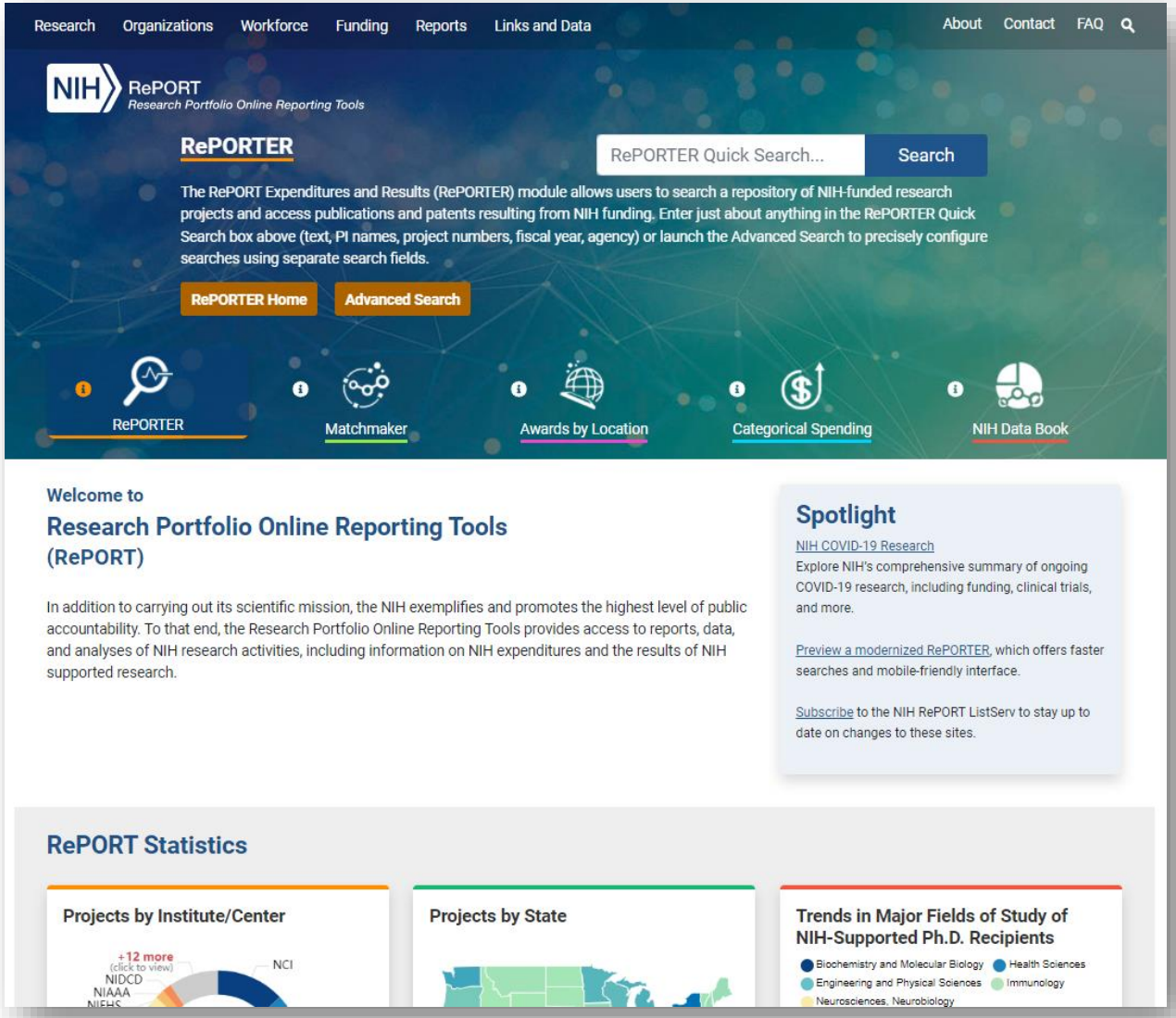
Public Tools: RePORT and RePORTER

- The Research Portfolio Online Reporting Tools (RePORT) website is a one-stop shop for reports, data, and analyses of NIH research activities.

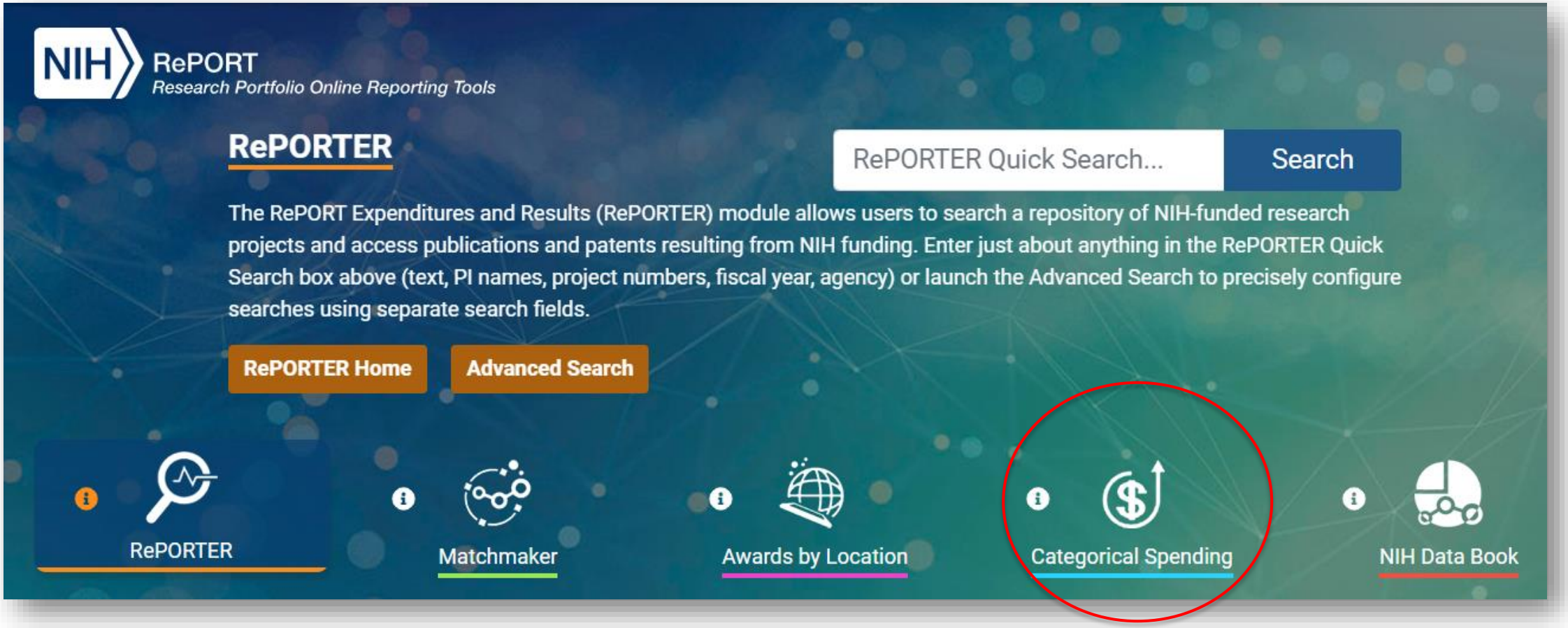
- Provides public data at multiple levels of complexity.

- RePORTER is a publicly-available tool within RePORT, listing awards from NIH and other agencies* and data on news, publications, patents, and clinical trials associated with NIH funding.

** NIH, ACF, AHRQ, CDC, FDA, VA*

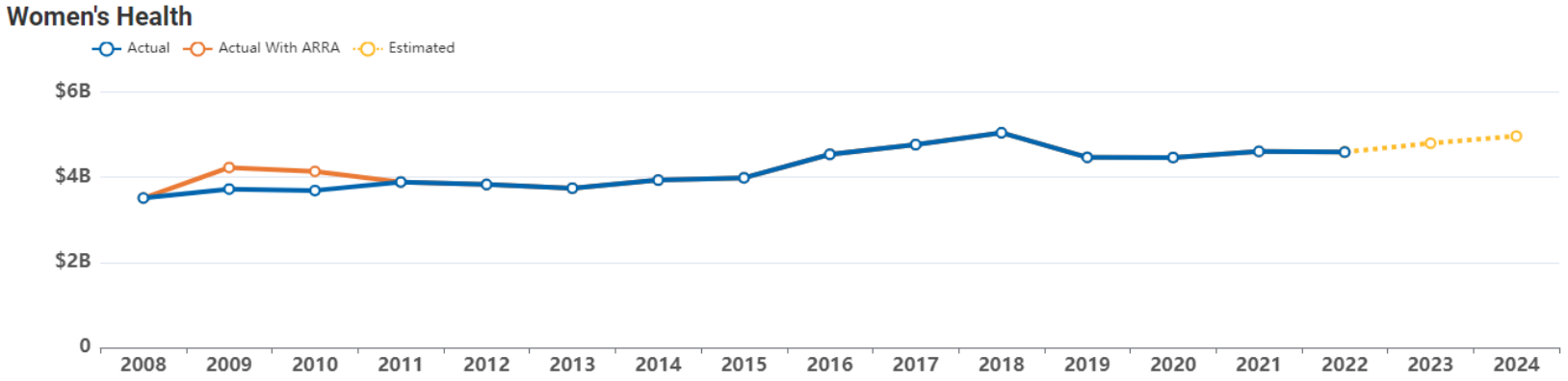


Accessing Public *Women's Health* Data



- Prior year data for RCDC categories can be found by going to Categorical Spending

Official *Women's Health* Report



Data for Fiscal Year: 2022

Back | Export | Showing 100 of 13740 projects | « ‹ 1 2 3 4 ... › »

Funding IC	Project Number	Sub Project #	Project Title	PI Name	Org Name	State / Country	Amount
NIA	5R01AG051647-04	-	Combining Testosterone Therapy and Exercise to Improve Function Post Hip Fracture	BINDER, ELLEN	WASHINGTON UNIVERSITY	MO	3021234
NIAMS	5R01AR070806-04	-	Role and Mechanism of Claudin-11 Action and Signaling in Bone	MOHAN, SUBBURAMAN	LOMA LINDA VETERANS ASSN RESEARCH & EDUC	CA	207400
NCI	5R01CA172145-08	-	Very-long Term Neurocognitive Outcomes in Breast Cancer Survivors	PALESH, OXANA	STANFORD UNIVERSITY	CA	135815
NIDA	5R01DA039137-05	-	Health Care Policy and Substance Abuse Treatment Access	OLFSON, MARK	COLUMBIA UNIVERSITY HEALTH SCIENCES	NY	221461
NIA	5R13AG066368-02	-	Social Neuroscience of Grief: 2020 Vision and Social Neuroscience of Grief: Early Adversity and Later Life Reversibility	O'CONNOR, MARY-FRANCES	UNIVERSITY OF ARIZONA	AZ	9165

- The project listing displays information such as funding IC, organization name, and funding amount
- Exportable
- *Women's Health* project level data available starting FY2020

NOTE: FY2023 data will be released along with the President's budget

Public Project Text and Data

Project Details

- Description
- Details
- Sub-Projects
- Publications
- Patents
- Outcomes
- Clinical Studies
- News and More
- History
- Similar Projects

Combining Testosterone Therapy and Exercise to Improve Function Post Hip Fracture

Project Number	Contact PI/Project Leader	Awardee Organization
5R01AG051647-04	BINDER, ELLEN FOther PIs	WASHINGTON UNIVERSITY

Description

Abstract Text

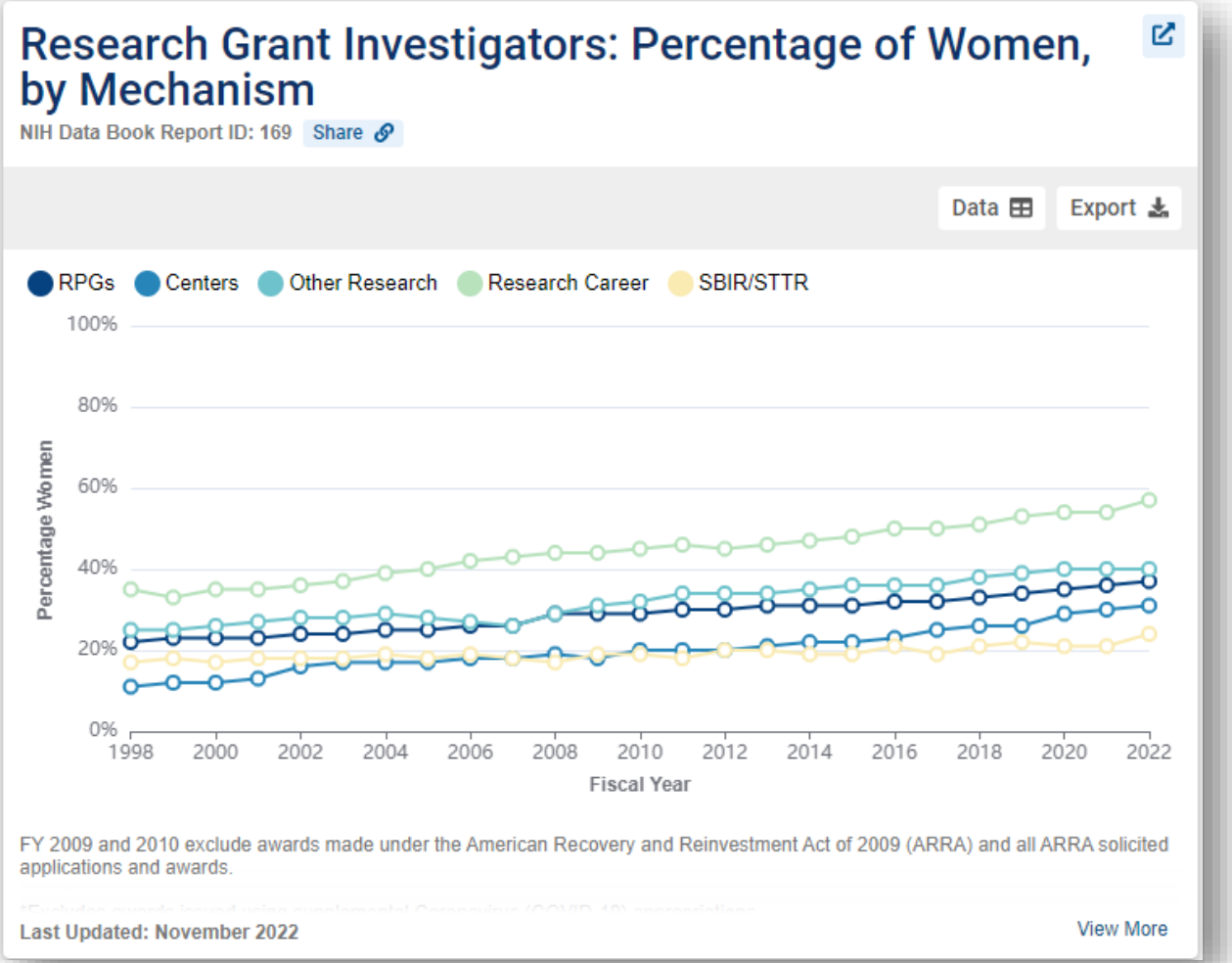
Hip fractures are common among older women and can have a devastating impact on their ability to remain independent. A clinically important functional decline and failure to recover following a hip fracture has been documented as much as a year after the fracture, even among individuals who were functioning at high levels before the event. Age-associated androgen deficiency in women contributes to deficits in muscle mass, strength and power that are common in this patient population before the fracture, and are exacerbated afterward. A pilot study of testosterone (T) supplementation in elderly female hip fracture patients has demonstrated the feasibility of T treatment in this population, and showed gains in lean body mass (LBM) and muscle strength with active drug, compared to placebo. The benefits of exercise in restoring muscle strength and physical function after a hip fracture have been documented. However, it remains unclear whether T treatment can augment the effects of exercise on mobility and patient-reported function, or whether any observed benefits are sustained beyond the period of active

- Clicking a project will show the project's Title, Abstract, and Public Health Relevance text and other RCDC categorizations.
- Identify publications, clinical studies, patents, etc. associated with the project

NIH Data Book

- The NIH Data Book summarizes answers to commonly asked questions about the NIH budget and extramural programs – including data on awards by gender.

- Data are updated annually



RCDC Inclusion Statistics Report

- Available only for automated RCDC reports
- Participants can be counted in more than 1 category
- Inclusion data does **not** map to RCDC budget data as the data are processed differently
  - Inclusion data processed by administering IC
  - Budget data processed by funding IC

NIH RCDC Inclusion Statistics Report

NIH RCDC Inclusion Statistics Report

Summary | Detail | RISR FAQs | Contact Us

Filter RCDC Categories | Total, NIH | 2021 | Sex/Gender | Exclude Single Sex/Gender Studies | Clear Filters | Export

RCDC Category	Total Participants	Female Participants	% Female Participants	Median % Female Participants	Male Participants	% Male Participants	Median % Male Participants	Participants of Unknown or Unreported Sex/Gender
ALS	1,687	804	48%	51%	857	51%	49%	26
Acquired Cognitive Impairment	511,726	316,108	62%	58%	192,304	38%	41%	3,314
Acute Respiratory Distress Syndrome	100,213	44,439	44%	44%	47,007	47%	55%	8,767
Adolescent Sexual Activity	49,288	26,486	54%	50%	21,949	45%	46%	853
Agent Orange & Dioxin	4,966	2,541	51%	51%	2,425	49%	49%	<12
Aging	1,646,954	1,008,463	61%	55%	593,858	36%	43%	44,633

Thank You!

Questions?