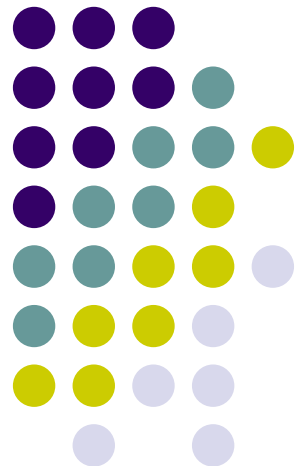
People Related Data Requirements

Walter Schaffer
Commons Working Group
May 19, 2004



An Enumeration Exercise

Analysis—Key Findings:

- Estimated number of personnel involved with these sampled projects:

207,711

An Enumeration Exercise

Analysis—Key Findings (continued):

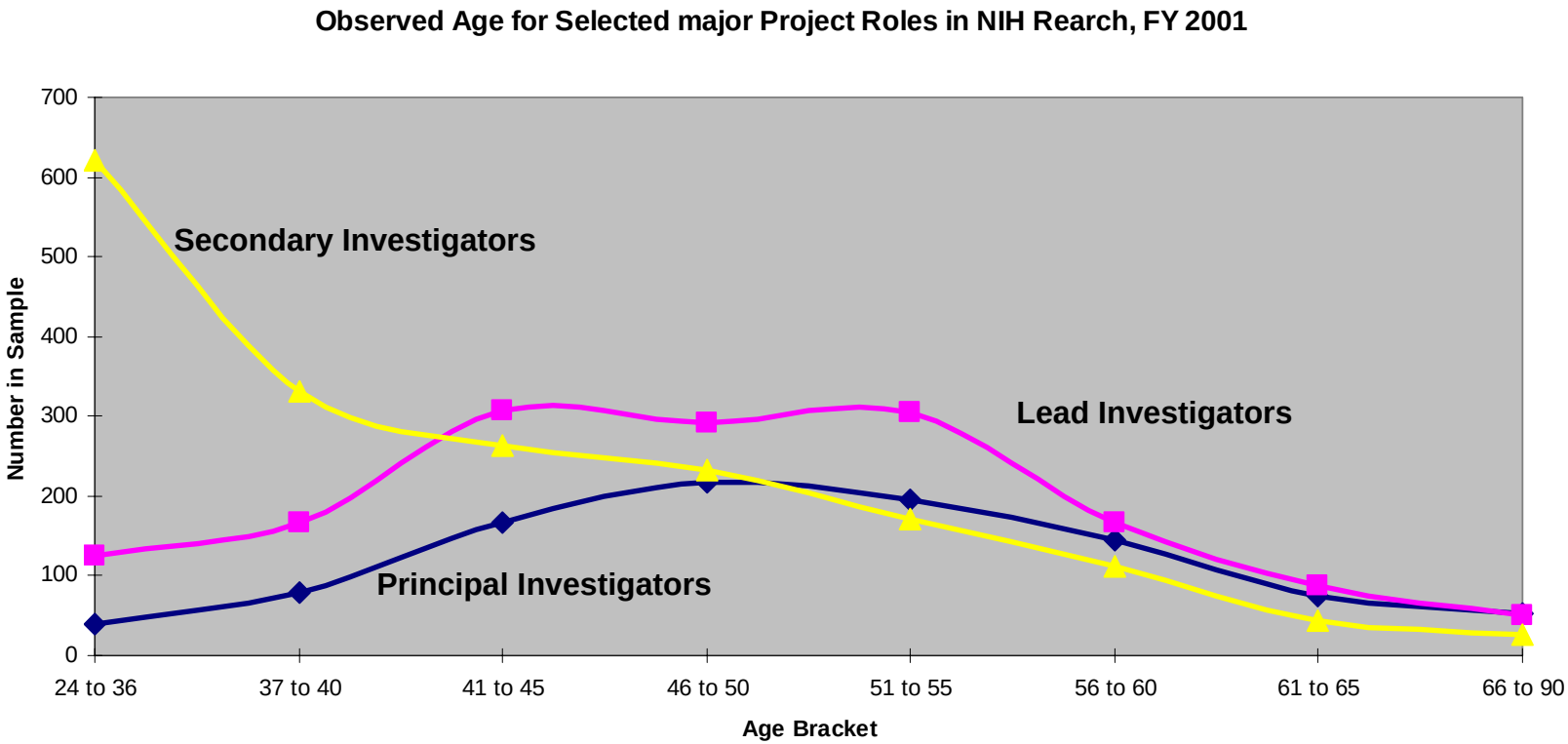
- Estimated 35,967 Principal Investigators
- Estimated 51,678 Lead Investigators (involved in the overall scientific direction of the research)
  - co-principal investigators
  - co-investigators
  - project leaders, etc.
- Estimated 69,360 Secondary Investigators
  - post-doctoral fellows
  - research associates, etc.
- Estimated 23,376 technicians
- Estimated 11,117 students
- Estimated 2,944 administrative staff

An Enumeration Exercise

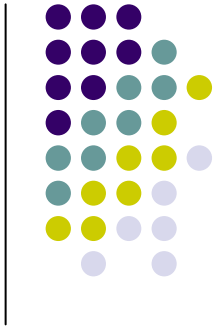
Analysis—Key Findings (continued):

- Estimated Degree breakouts:
  - 96,277 Ph.D.s
  - 39,708 M.D.s
  - 9,720 M.D./Ph.D.s
  - 908 veterinarians
  - 737 dentists
- Estimated 86,629 FTEs
  - 2.4 FTE's per award)
- Estimated total individuals involved in NIH research 212,410 when K awards are included
- Matching SSNs to CGAF showed that some 1,300 of the sampled lead, secondary, and administrative staff had been PIs at some future date

An Enumeration Exercise

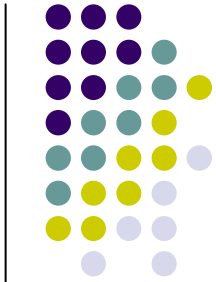


An Enumeration Exercise



- Convenience sample of readily available applications
  - Based upon Type 5 (Form 2590) applications for FY 2002 from all ICs
  - Stratified across award types
  - *Not* randomized
  - Assumes no inherent bias
- Selected mechanisms
  - comprised 35,967 (or 88.4%) of all awards
  - Comprised $14.1 billion (or 94.6%) of disbursed dollars
- Sampled 1,039 Awards
  - 7,000 individuals
  - 3½ weeks

An Enumeration Exercise – A Fictional Form 2590



PERSONNEL REPORT

GRANT NUMBER

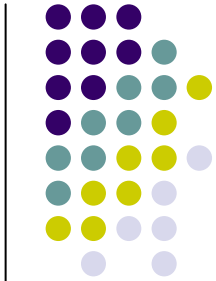
Place this form at the end of the signed original copy of the application. Do not duplicate.

5P50 CA90949-02

All Key Personnel for the Current Budget Period

Name	Degree(s)	SSN	Role on Project (e.g. *Pf*, Res. Assoc)	Date of Birth *(MM/DDIYY)*	Annual % Effort
O'Neill, Shaquille	M.D., Ph.D.	000-00-0000	**Principal Investigator**	8/8/1955	31%
Summitt, Pat	Ph.D.	000-00-0000	Co-Leader	10/15/1959	45%
Ming, Yao	--------	000-00-0000	Programmer	3/26/1952	20%
Glass, Phillip	--------	000-00-0000	Research Assistant	11/24/1962	100%
Rimbaud, Alfonse	Ph.D.	000-00-0000	Co-Investigator	4/12/1943	5%
Jarry, Alfred	Ph.D.	000-00-0000	Co-Investigator	5/2/1965	10%
Keester, Mike	M.D.	000-00-0000	Co-Investigator	1/17/1958	15%
Olekola, Festus Olufunsha	M.D.	000-00-0000	Co-Investigator	3/17/1965	25%
Bonecuffer, Virgil T.	M.D., Ph.D.	000-00-0000	Co-Investigator	1/5/1956	10%
Nohboddi, Donatella	Ph.D.	000-00-0000	Postdoctoral Fellow	9/9/1969	40%
Louis-Dewey, Hugh E.	RN.	000-00-0000	Research Nurse	5/9/1949	100%
Breton, Adriana	M.D.	000-00-0000	Core Leader	11/28/1968	15%
O'Ferra, Marge N.	MD.	000-00-0000	Project Leader	3/9/1959	10%

People Data Needs

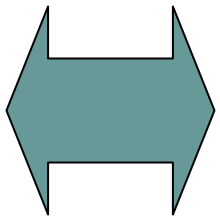


- **No longer enough to record information on PI alone**
  - Need composition of entire team
  - Source of information about the workforce
  - Use of actuarial approaches to estimate replacement needs and the size of the training programs
  - Involvement of clinicians and others with special backgrounds
  - Training outcome studies
  - Essential for proper stewardship
- **Information present in paper applications**
  - Need to ensure that comparable data is available in electronic applications
  - Need professional profiles on all professional-level contributors
  - How this information is requested and reported?
  - Who controls registration and data entry?
  - How are existing professional profiles linked to transaction records?

Training and Career Development Outcome Studies

Trainee

- **Level of Training**
- **Field of Training**
- **Personal Factors**
  - Age/Sex/Ethnicity
- **Selection Factors**
  - Test Scores
- **Program Factors**
  - Duration
  - Completion
- **Institutional Factors**



Career Outcome Factors

- **Research Involvement**
- Further Training
- Employment
  - Level
  - Setting
- Publications
- Citations
- Patents
- Grant Applications/Awards
- Advisory or Review Panels
- **Clinical Involvement**
- Specialty Certifications
- Clinical Activities

Overview of Early Career Outcomes and Group Comparisons in the Biomedical Sciences

Outcome	Observed Results			More Progress on Outcome for NRSA Trainees and Fellows vs.	
	NRSA Trainees and Fellows	Ph.D.s from NIH Training Institutions	Ph.D.s from Non-NIH Training Institutions	NIH Training Institutions	Non-NIH Training Institutions
Completed Ph.D. in less time (years)[a]	6.5	6.9	7.0	+	+
Pursued postdoctoral training[b]	77.9	59.9	47.6	+	+
Working in a research career position in 1995	81.3	74.0	69.8	+	+
Had academic, tenure-line position	39.3	29.1	32.0	+	+
Employed by top-ranked academic institution	36.6	23.2	15.7	+	+
Applied for one or more NIH/NSF grants	46.9	35.0	29.6	+	+
Awarded an grant (of those who applied)	66.8	55.0	47.2	+	+
Number of post-Ph.D. journal publications[c]	12.8	9.7	8.9	+	+
Average citations to published articles[c]	28.5	24.7	18.9	+	+

Note. Unless indicated otherwise, included are those individuals who received their Ph.D. in the biomedical sciences between 1981 and 1988. A "+" indicates the observed difference (unadjusted) was significant and in the direction where NRSA trainees and fellows outperformed their comparison group counterparts favorable ways. Enclosing the "+" by a box indicates that NRSA predoctoral support was found to be statistically significant in helping to explain the observe difference, after adjusting for the influence of other variables; shading of the box indicates that its role was marginally significant ($p < 0.06$).

[a]1981-92 Ph.D.s
[b]1981-90 Ph.D.s
[c]1981-82 Ph.D.s

Pion, G. The Early Career Progress of NRSA Predoctoral Trainees and Fellows, 2001

Source of Evaluation Information

Trainee

• Level of Training	-----	2271 – TFF
• Field of Training	-----	2271 – TFF
• Age/Sex/Ethnicity	-----	2271 – TFF
• Test Scores	-----	ETS
• Duration of Training	-----	2271 – TFF
• Completion of Training	-----	2271 – TFF
• Degree Earned	-----	2271 – TFF

Career Outcome Factors

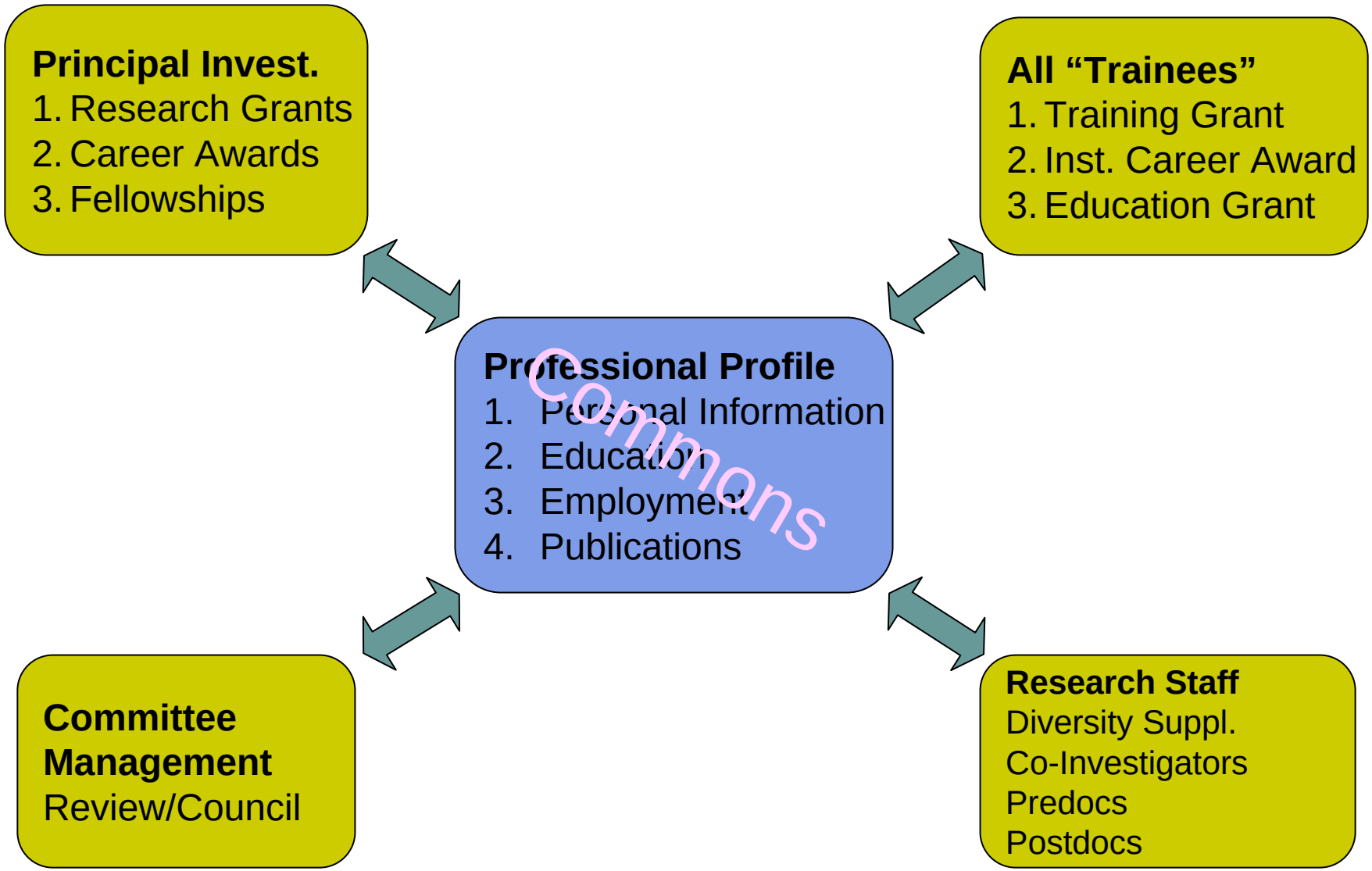
• Comparison Groups	-----	DRF
• Further Training/Education	-----	SDR & TA Module
• Employment Setting/Rank	-----	SDR, FRS, IRP
• Publications	-----	NLM & ISI
• Citations	-----	ISI
• Grant Applications/Awards	-----	398/416 – CGAF
• Staff on Research Grants	-----	2590
• Patents	-----	iEdison
• Advisory or Review Panels	-----	Committee Management
• Clinical Certification	-----	FRS
• Clinical Activities	-----	FRS

FRS = AAMC Faculty Roster System
NLM = National Library of Medicine
ETS = Educational Testing Service
TFF = Trainee Fellow File
IRP = Intramural Research Program
SDR = NSF/NIH Survey of Doctoral Recipients
ISI = Institute of Scientific Information
DRF = Doctorate Record File – Surv. Earned Doc.
CGAF = Consolidate Grant Applicant File

Professional Profile (PPF)

1. Personal Information
   - Name
   - SSN
   - Sex
   - Ethnicity
   - Race
   - Date of Birth
   - Address
   - Phone Number
   - FAX Number
   - E-Mail
   - Citizenship
   - Disability
2. Education
   - Degree Awarded/Expected
   - Year Awarded/Expected
   - Institution
   - Major
   - Minor
3. Employment
   - Employer
   - Address
   - Work Phone
   - Work FAX
   - Work E-Mail
   - Start Date
   - End Date
   - Rank
   - Position
4. Publications
   - NLM Accession Number
   - Citation Text

PPF Data Entry Model



Co-Investigators

- Need to give appropriate credit to Co-Investigators and others who make substantive contributions to the design of a research project
  - Catalyze team science
  - Reduce barriers to the formation of interdisciplinary teams
  - Recognition
  - Coordinate with:
    - Roadmap
    - BECON
    - Research Business Models - OSTP
- Implementation
  - More than one lead on a project?
  - Equal footing?
  - Single lead with identified and recognized co-leads?
  - Separate awards?
  - Financial allocation within an individual project?
  - Reporting:
    - CRISP
    - NGA

People Data Needs

- No longer enough to record information on PI alone
- Need to retain all the people-related information present in paper applications
  - How should this data be collected in a totally electronic environment?
  - How do we achieve an overall reduction in respondent burden and still obtain the information we need?