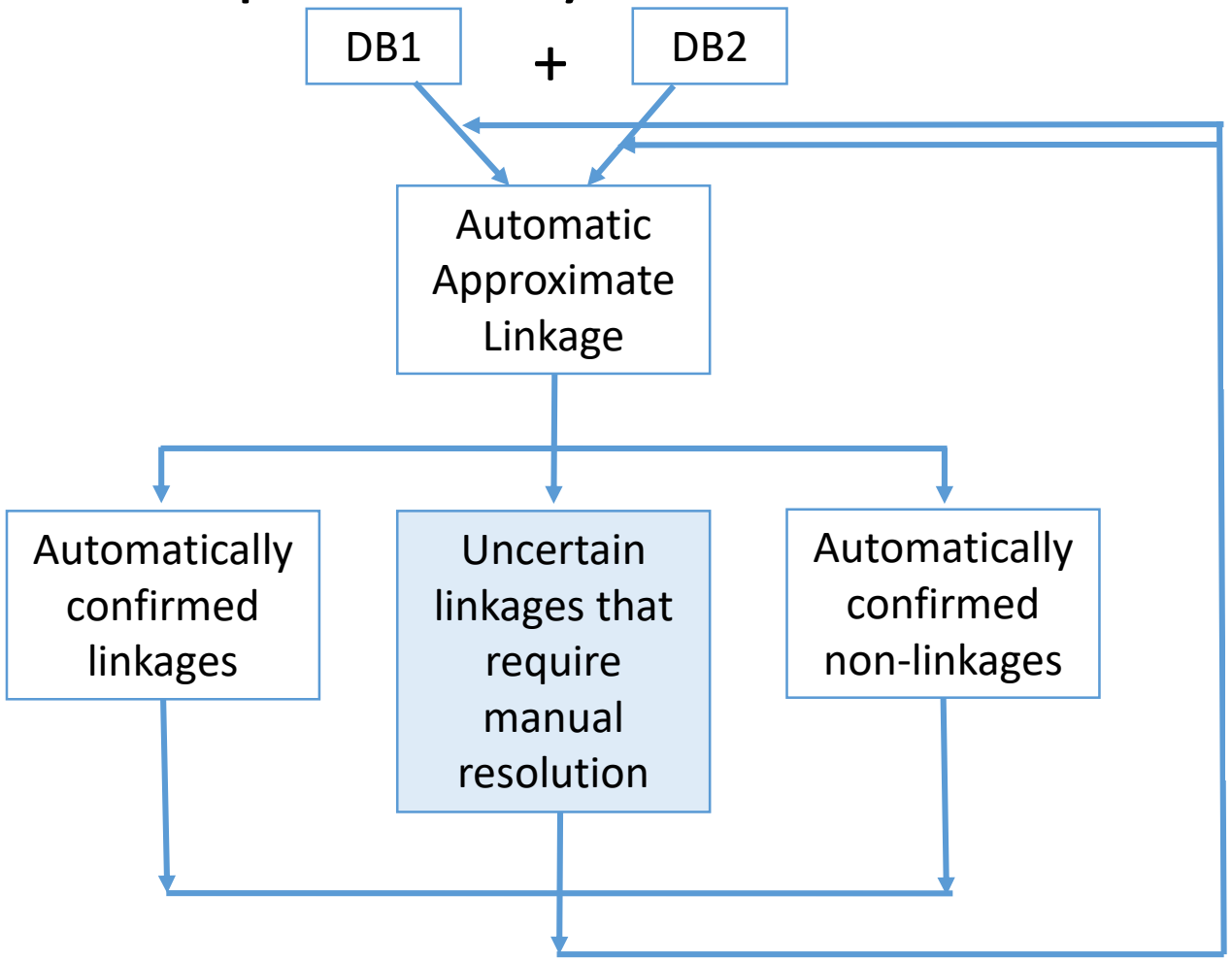
Approximate RL
Human-computer system



- Standardize
- Clean Data
- Build Training Data

- 75%-80% automatics
- 15%-25% manual resolution

Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression

https://pinformatics.org/ppirl/

Phase 1 – Completed Framework on Privacy Preserving Interactive Record Linkage (PPIRL): Privacy & Utility Objective

Phase 2 – Research Needed: Algorithm & Methods Development for Design of SDLink Software and Companion Documents (PCORI proposal)

Approach	**Computational: Agile Software Development (Iterative Spiral Process)**		**Participatory Action Research**
Methods	**Incremental, on-demand, partial disclosure**	**k-anonymity set size & Apriori algorithm**	**Nominal Group Technique & Delphi**
	Aim 1: Effective Info. Disclosure Task1.A: Design User Interface (UI) Task1.B: Algorithm & Implementation Task1.C: Evaluation – What is the trade off between information disclosure and linkage quality ?	**Aim 2: Theoretical Privacy Risk Analysis** Task2.A: Design Budget System Task2.B: Algorithm & Implementation Task2.C: Evaluation – What budget (level of disclosure) is required for high quality linkage ?	**Aim 3: Practical Privacy Risk Analysis** Task3.A: Engage & Education on PPIRL Task3.B: Build Consensus on PPIRL Task3.C: Incorporate into companion documents (*Privacy Statement, IRB Application, DUA*) and SDLink Software
Outcome	**SDLink Software Prototype (Pre-Beta version: Year 2 & Prototype Version 0.5: Year 3)**		**3 SDLink Companion Documents**

Phase 3 – After Project is Completed: Hardening Code – SDLink Software Development & Release (Collaboration with Kitware Inc.)

Aim 1: Effective Information Disclosure

- July 2017: User Study 1
  - The study had a total of 104 participants
    - ~20 participants for each of the five modes
    - There were 61 males and 42 females, and one participant did not specify gender.
    - Ages ranged from 18 to 43 years, and the median age was 24 years.
    - About 65% of the participants were from the United States and had English as their native language.
    - About 57% of the participants were either pursuing or already had a graduate degree, and the remaining participants were undergraduate university students.
  - 30 questions

BASE mode

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B

FULL mode

(Icons & colors)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767	1	JUDE	WILLIAM	1	09/09/1906	M	W
	✖			+		⇄		DIFF
	8000003567	1	JUDE	WILLIAM JR	1	09/09/1960	M	B

MODERATE mode

(Close same & partial IDs)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	1	✓	WILLIAM	1	09/09/1906	M	W
	✖			+		⇄		DIFF
	******35**	1	✓	WILLIAM JR	1	09/09/1960	M	B

MINIMUM mode

(Partial names, dates etc)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	1	✓	*******	1	**/**/**06	M	@
	✖			+		⇄		DIFF
	******35**	1	✓	******* JR	1	**/**/**60	M	&

MASKED mode

(Only symbols)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******@@**	1	✓	*******	1	**/**/**@@	✓	@
	✖			+		⇄		DIFF
	******&&**	1	✓	******* &&	1	**/**/**&&	✓	&

ENCRYPTED mode

(Encrypt Data)

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	@@zz@@@@e@@@@@yc@tcflg==	1	HOMwdz8KpFKaTfPE+qr8Xw==	✓	1	/KSKzJ2U5C/fpHmkMqZPqw==	~	%
	DIFF		DIFF			DIFF		DIFF
	&&qw&p&&&&&&m&&&&vlm&&==	1	o1fSci26GzxKx41n1lkRuQ==	✓	1	bJupClSkjj/bmw9DRq07vw==	~	^

Mode 2: Full Mode
Full disclosure with markup

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767	1	JUDE	WILLIAM	1	09/09/1906	M	W
	8000003567	1	JUDE	WILLIAM JR	1	09/09/1960	M	B
2	0000006947	1	BRYANT	MADELINE	1	05/02/1962	F	W
	0000006947	25	MADELINE	BRYANT	•••	05/02/1962	F	W
3	9000018540	•••	SALLY	BYRD	•••	07/04/1960	F	W
	6000008928	∞	JOHN	BYRD	•••	04/07/1960	M	?

Mode 3: Moderate Mode
Moderate disclosure with markup

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	1	✓	WILLIAM	1	09/09/1906	M	W
	✖			✚		⇄		DIFF
	******35**	1	✓	WILLIAM JR	1	09/09/1960	M	B
2	✓	1	&&&&&&	@@@@@@@@	1	✓	F	✓
	✓	2-5	@@@@@@@@	&&&&&&	•••	✓	F	✓
3	@@@@@@@@@@	•••	SALLY	✓	•••	07/04/1960	F	*
	DIFF		DIFF			✖	DIFF	
	&&&&&&&&&&	∞	JOHN	✓	•••	04/07/1960	M	?

Mode 4: Minimum Mode
Minimum disclosure with markup

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	1	✓	*******	1	**/**/**06	M	@
	✗			+		⇄		DIFF
	******35**	1	✓	******* JR	1	**/**/**60	M	&
2	✓	1	&&&&&&	@@@@@@@	1	✓	F	✓
	✓	2-5	@@@@@@@@	&&&&&&	•••	✓	F	✓
3	@@@@@@@@@@	•••	@@@@@	✓	•••	07/04/****	F	*
	DIFF		DIFF			✗	DIFF	
	&&&&&&&&&&	∞	&&&&	✓	•••	04/07/****	M	?

Mode 5: Masked Mode
Masked disclosure with markup

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******@@**	1	✓	*******	1	**/**/**@@	✓	@
	✗			+		⇄		DIFF
	******&&**	1	✓	******* &&	1	**/**/**&&	✓	&
2	✓	1	&&&&&&	@@@@@@@@	1	✓	✓	✓
	✓	2-5	@@@@@@@@	&&&&&&	•••	✓	✓	✓
3	@@@@@@@@@@	•••	@@@@@	✓	•••	@@/&&/****	@	@
	DIFF		DIFF			✗	DIFF	
	&&&&&&&&&&	∞	&&&&	✓	•••	&&/@@/****	&	?

Mode 6: Encrypted Mode
Encrypted disclosure with markup

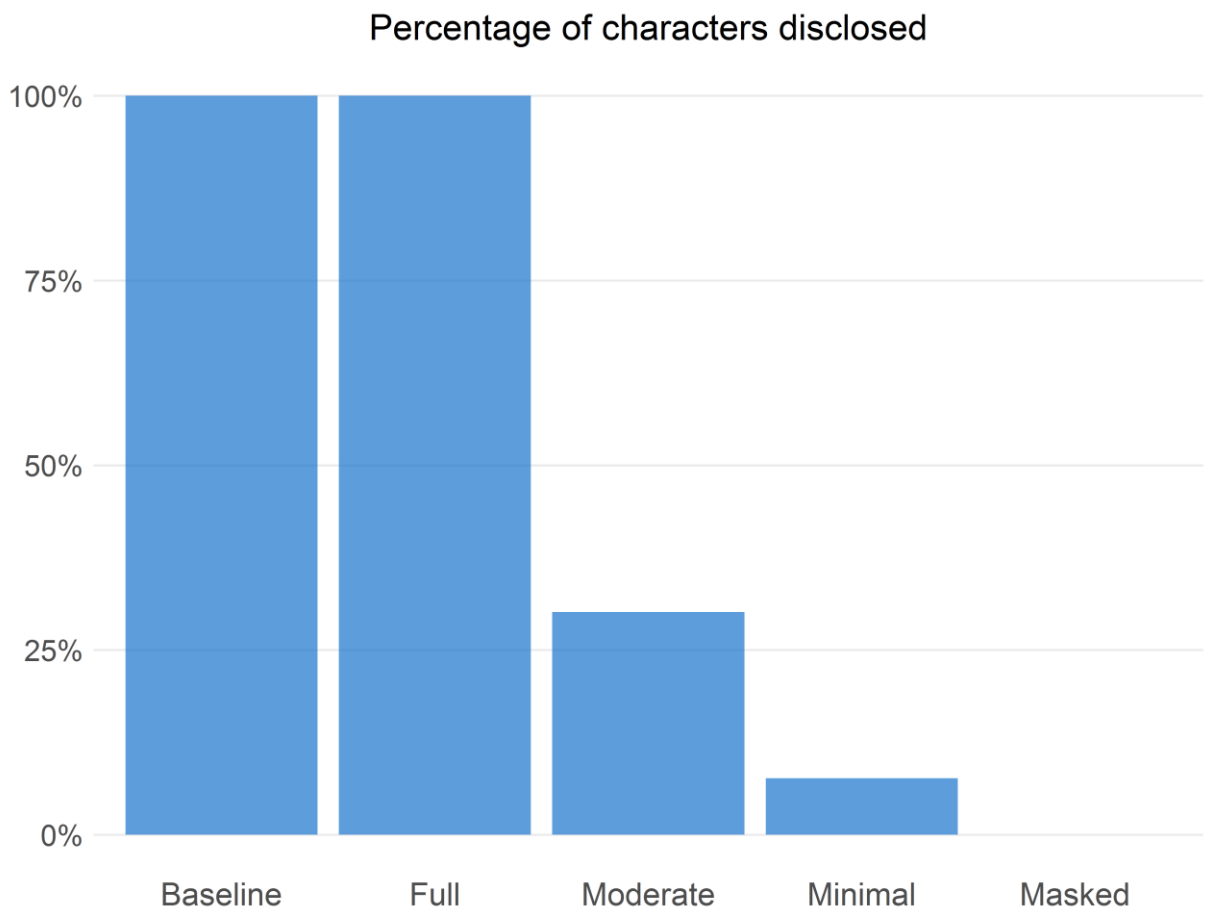
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
31	@@zz@@@@e@@@@@yc@tcflg== DIFF	∞	✓	Pgi+8vEbeh4nP757N9zGdg== DIFF	1	/KSKzJ2U5C/fpHmkMqZPqw== DIFF	~	% DIFF
	&&qw&p&&&&&&m&&&&vlm&&==	1	✓	9SGxq1uytTwBKSC8SpQx8A==	1	bJupClSkjj/bmw9DRq07vw==	~	^
32	✓	2-5	&&&&&&&&	@@@@@@	•••	✓	>	✓
	✓	1	@@@@@@	&&&&&&&&	1	✓	>	✓
33	@@@l@i@g@@@os@@bn@@@@g== DIFF	1	H0Mwdz8KpFKaTfPE+qr8Xw== DIFF	✓	1	JPHm/tFJf/Sa38z+PthPYQ== DIFF	> DIFF	%
	vx&+&&h&v&+&xnmyqmaa&&==	1	o1fSci26GzxKx41n1lkRuQ==	✓	1	AgsX5d/vZ1tRukT6GTxCZw==	~	?

Encrypted Mode vs Full Mode

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
31	@@zz@@@@e@@@@@yc@tcflg==	∞	✓	Pgi+8vEbeh4nP757N9zGdg==	1	/KSKzJ2U5C/fpHmkMqZPqw==	~	%
	DIFF			DIFF		DIFF		DIFF
	&&qw&p&&&&&&m&&&&vlm&&==	1	✓	9SGxq1uytTwBKSC8SpQx8A==	1	bJupClSkjj/bmw9DRq07vw==	~	^
32	✓	2-5	&&&&&&&&	@@@@@@	•••	✓	>	✓
	✓	1	@@@@@@	&&&&&&&&	1	✓	>	✓
33	@@@l@i@g@@@os@@bn@@@@g==	1	H0Mwdz8KpFKaTfPE+qr8Xw==	✓	1	JPHm/tFJf/Sa38z+PthPYQ==	>	%
	DIFF		DIFF			DIFF	DIFF	
	vx&+&&h&v&+&xnmyqmaa&&==	1	o1fSci26GzxKx41n1lkRuQ==	✓	1	AgsX5d/vZ1tRukT6GTxCZw==	~	?

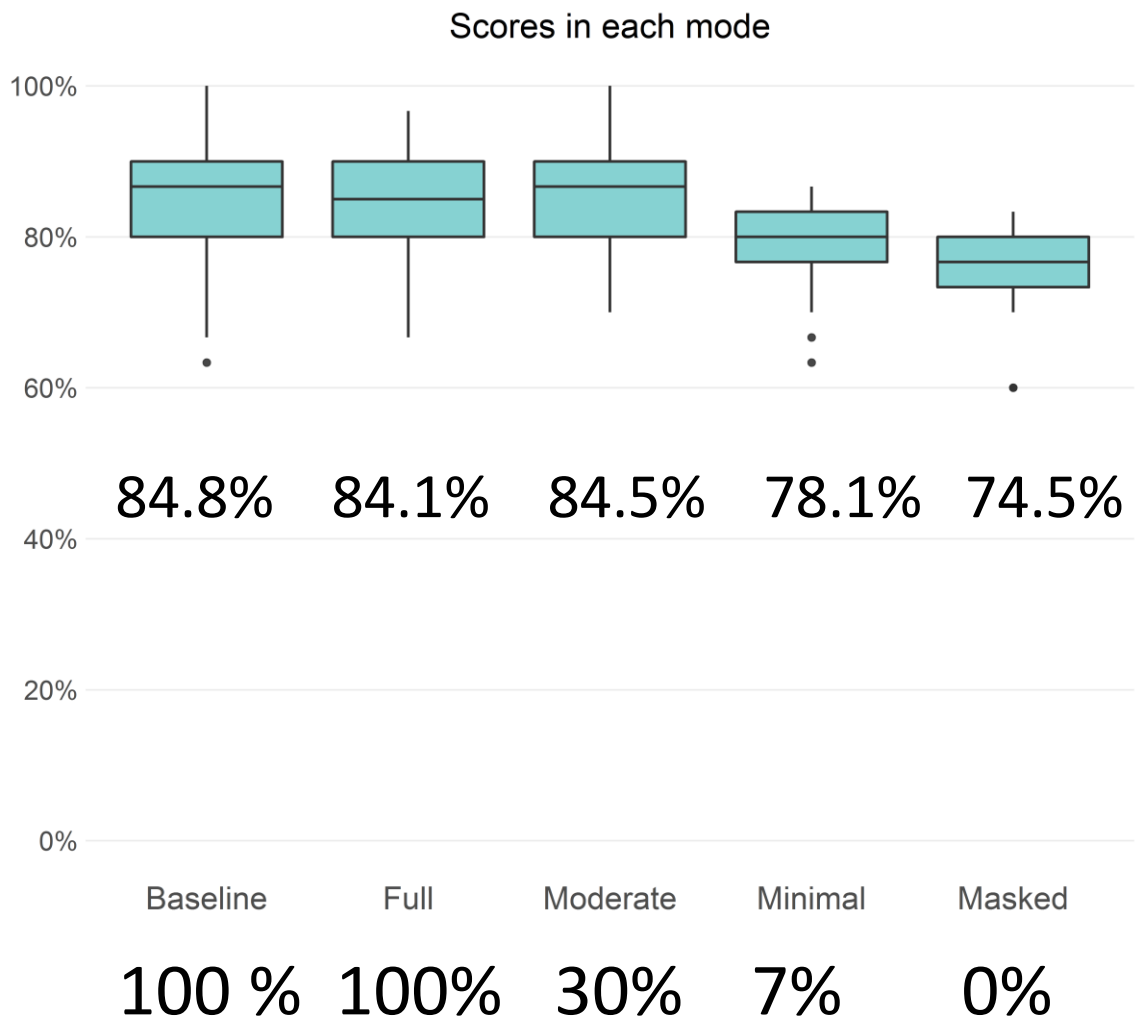
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767	1	JUDE	WILLIAM	1	09/09/1906	M	W
	✗			+		⇄		DIFF
	8000003567	1	JUDE	WILLIAM JR	1	09/09/1960	M	B
2	0000006947	1	BRYANT	MADELINE	1	05/02/1962	F	W
	0000006947	2-5	MADELINE	BRYANT	•••	05/02/1962	F	W
3	9000018540	•••	SALLY	BYRD	•••	07/04/1960	F	W
	DIFF		DIFF			✗	DIFF	
	6000008928	∞	JOHN	BYRD	•••	04/07/1960	M	?

Percentage of Characters Disclosed



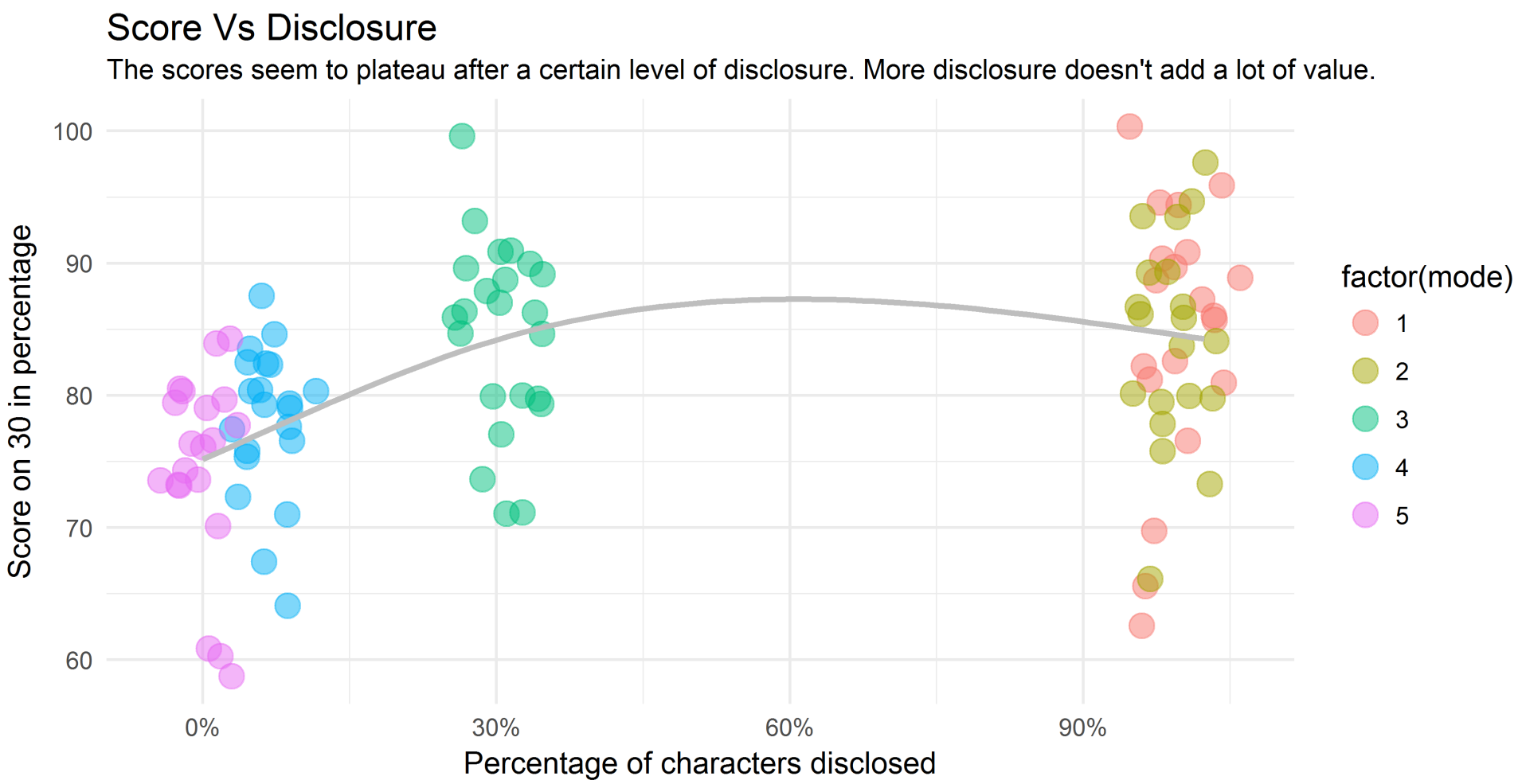
100 % 100% 30% 7% 0%

Accuracy Score by Disclosure Mode

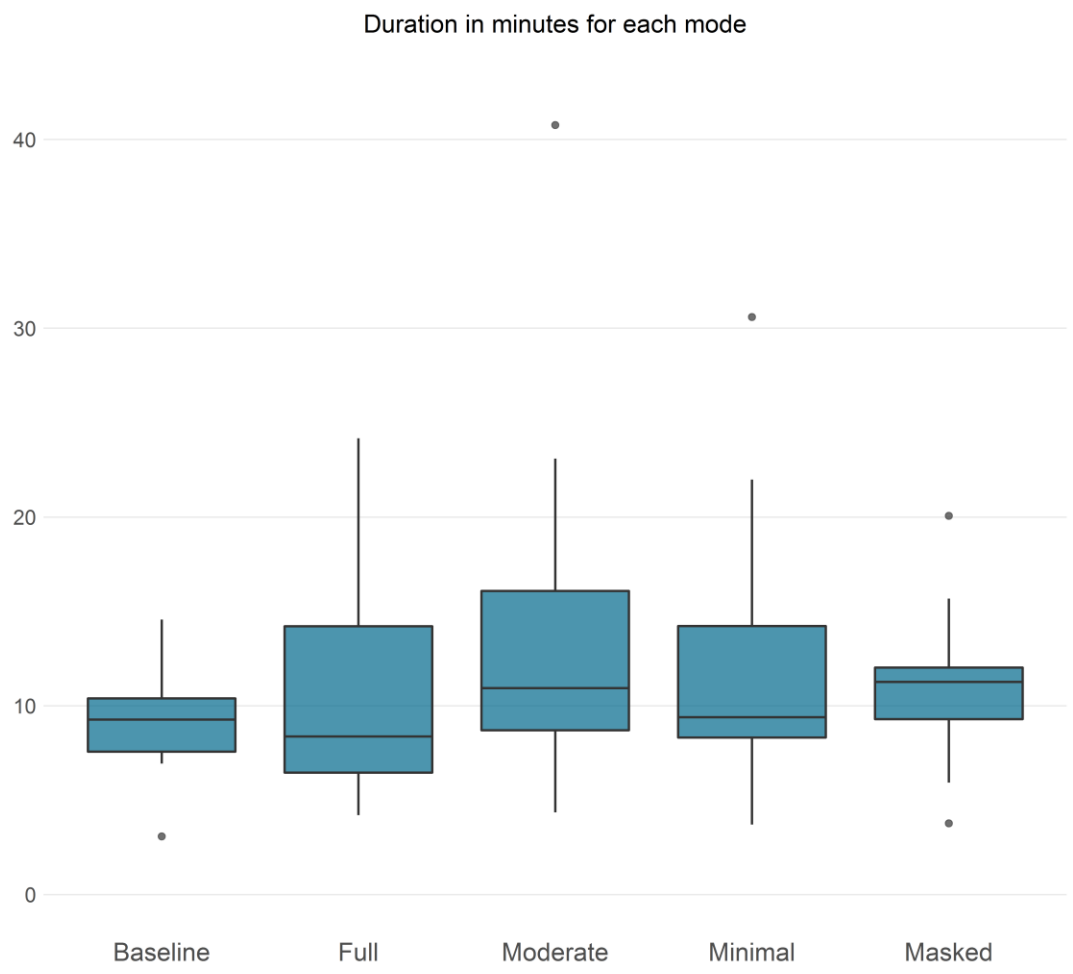


- We can get comparable results to full mode with only 30% disclosure with appropriate masks (moderate mode)
- As we mask more values for privacy, quality of results start to suffer ($p<0.001$)
- However, even legally de-identified data with proper masks can be linked properly for most situations
  - 0% disclosure still had 75% accuracy
  - Incremental disclosure can significantly improve privacy protection with negligible impact on quality of linkage

Score vs Disclosure

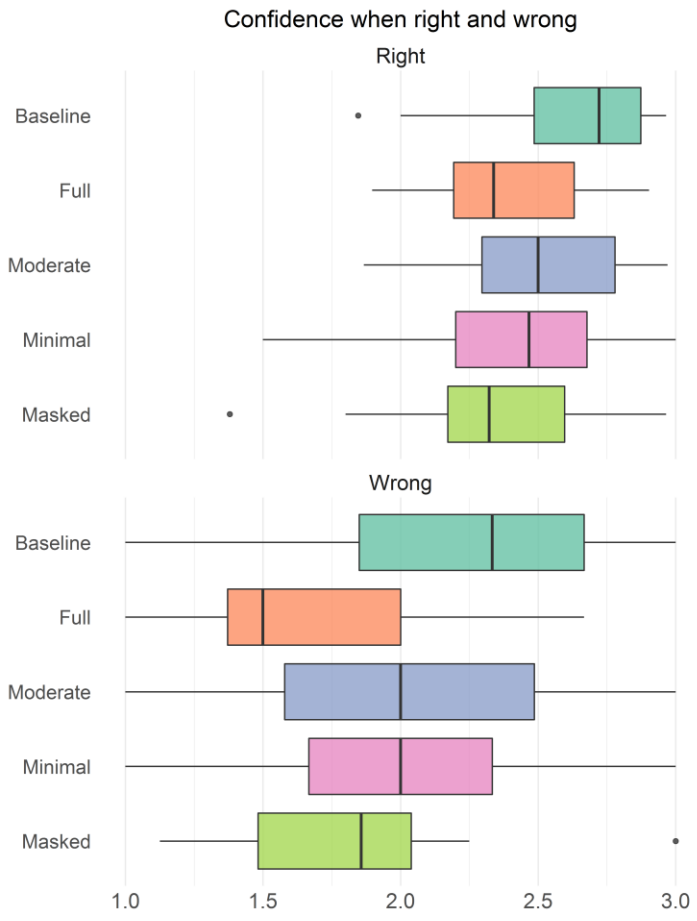


Time by Disclosure Mode



- Comparable across all modes
- More information (supplemental mark up and frequency icons) has more variability among participants
  - Probably due to differences in participants speed of processing information

Confidence Level by Correctness of Decision

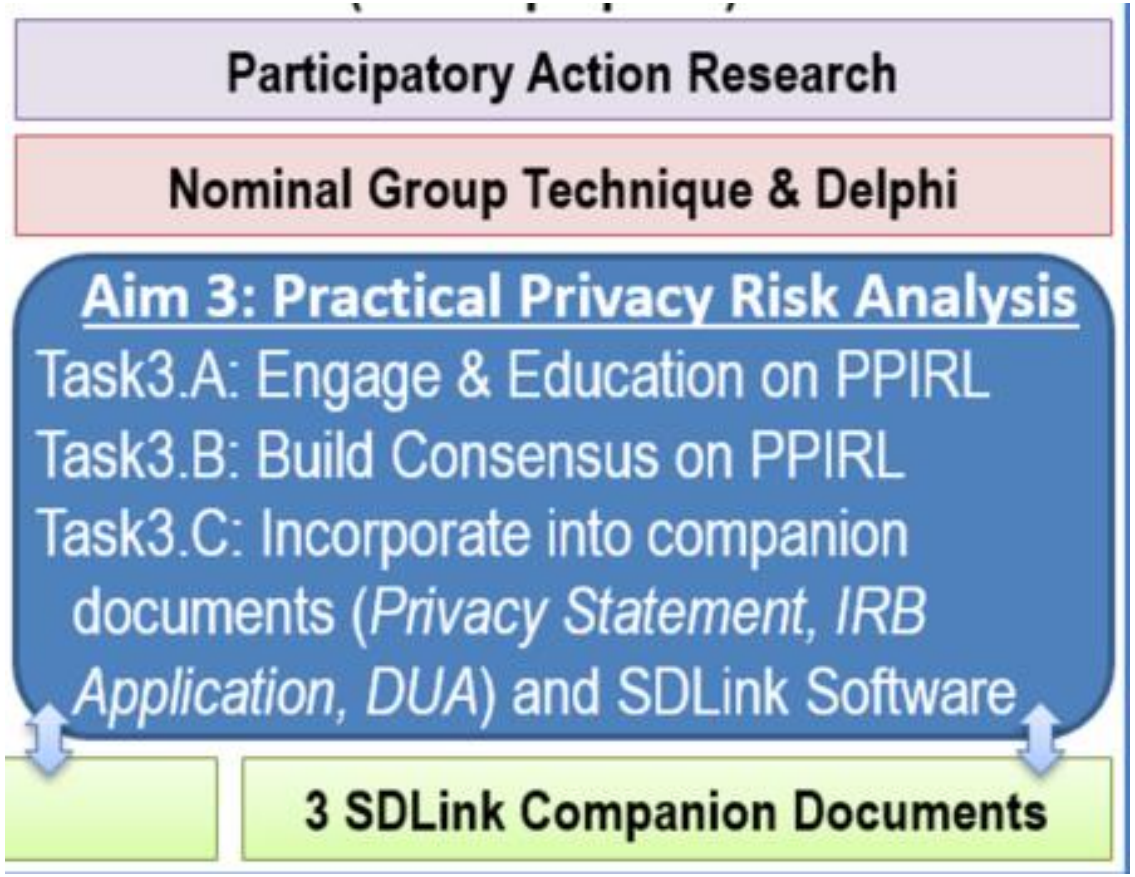


- Higher confidence when answers are correct (top) compared to when answers were wrong (bottom)
- Full mode is least confident when wrong answer
  - More information introduces more uncertainty in wrong decisions, but not sufficient to change the answer

PPIRL

- Aim 2: Theoretical Privacy Risk Analysis
  - User Study 2: Spring 2018
- Fall 2018: Beta release
  - Summative evaluation: UAB & UT Houston

Aim 3: Practical Privacy Risk Analysis

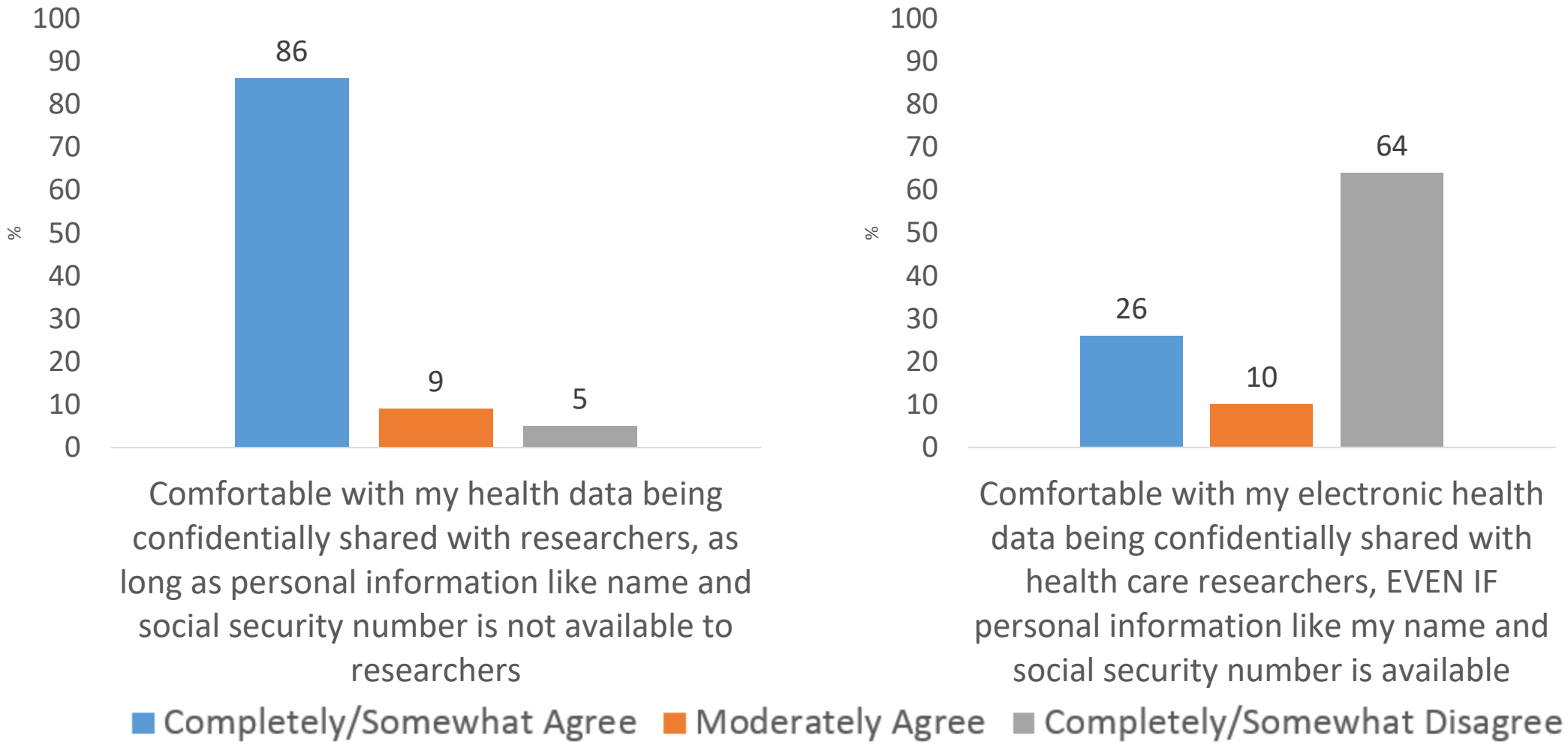


- Template IRB application & DUA
  - Nov 2017: ELSI NGT Session
  - Nov 2018: ELSI Delphi
- Privacy statement
  - Feb 2018: Patient NGT Session
  - Feb 2019: Patient Delphi
  - Feb 2020: Summative Evaluation (Survey)
- User Committee Meeting
  - Every six months: Feb & Aug

Improving Methods for Linking Secondary Data Sources for CER/PCOR

- A PCORI Record Linkage Project at Duke University
  - Sean O'Brien & Emily O'Brien
  - July 2014-
- O'Brien E.C., Rodriguez A.M., Kum H.-C., Schanberg L., O'Brien S.M., Setoguchi S. Patient perspectives on the linkage of health data for clinical research: insights from a survey in the United States. Oral presentation (#O17-3) at the 2017 World Congress of Epidemiology; Saitama, Japan. August 20, 2017.

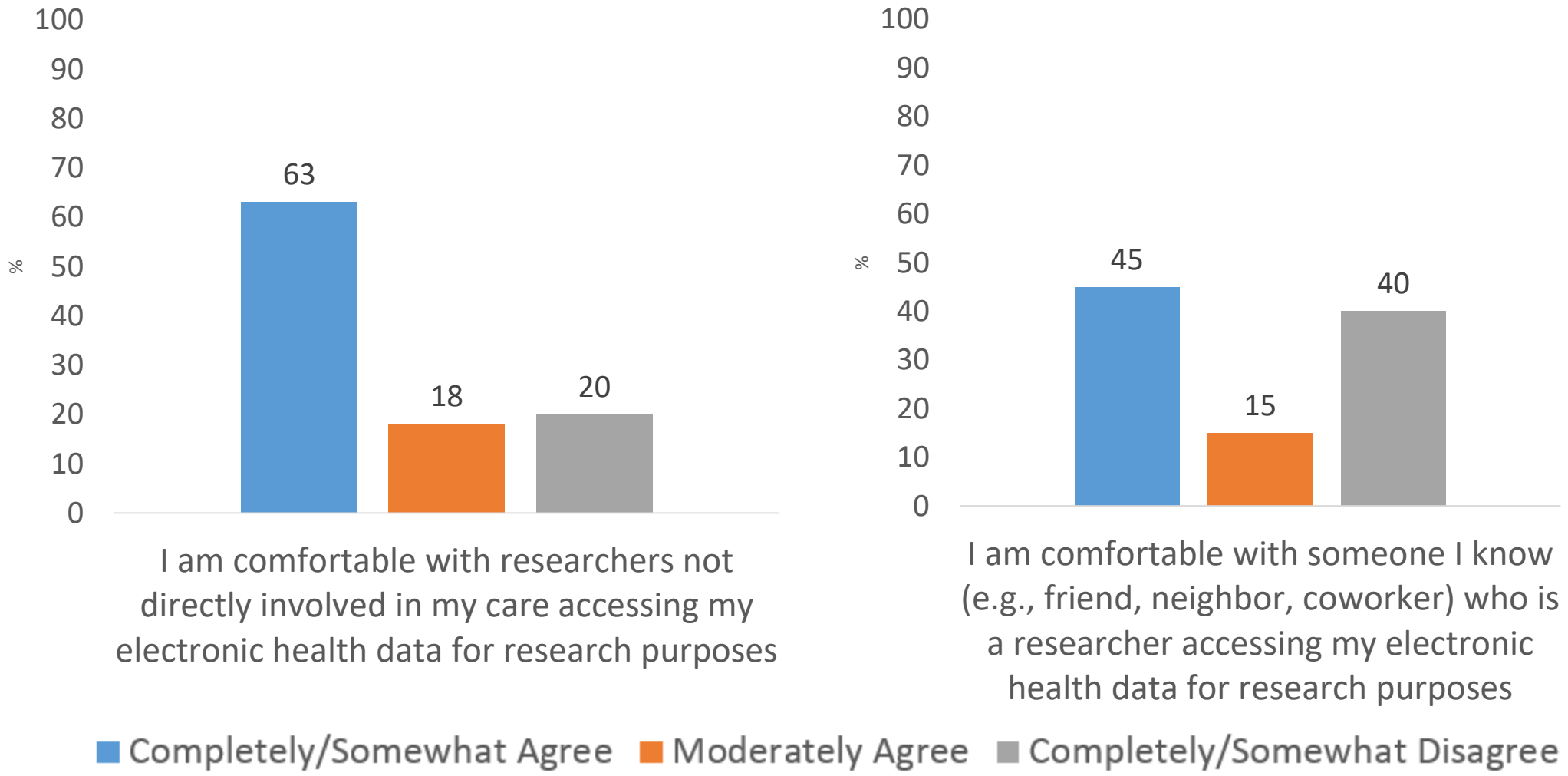
Figure 1. Data Sharing Comfort (n=3516): Sharing PII confidentially



Patients are concerned when PII is shared

Solution: disclosure control can help

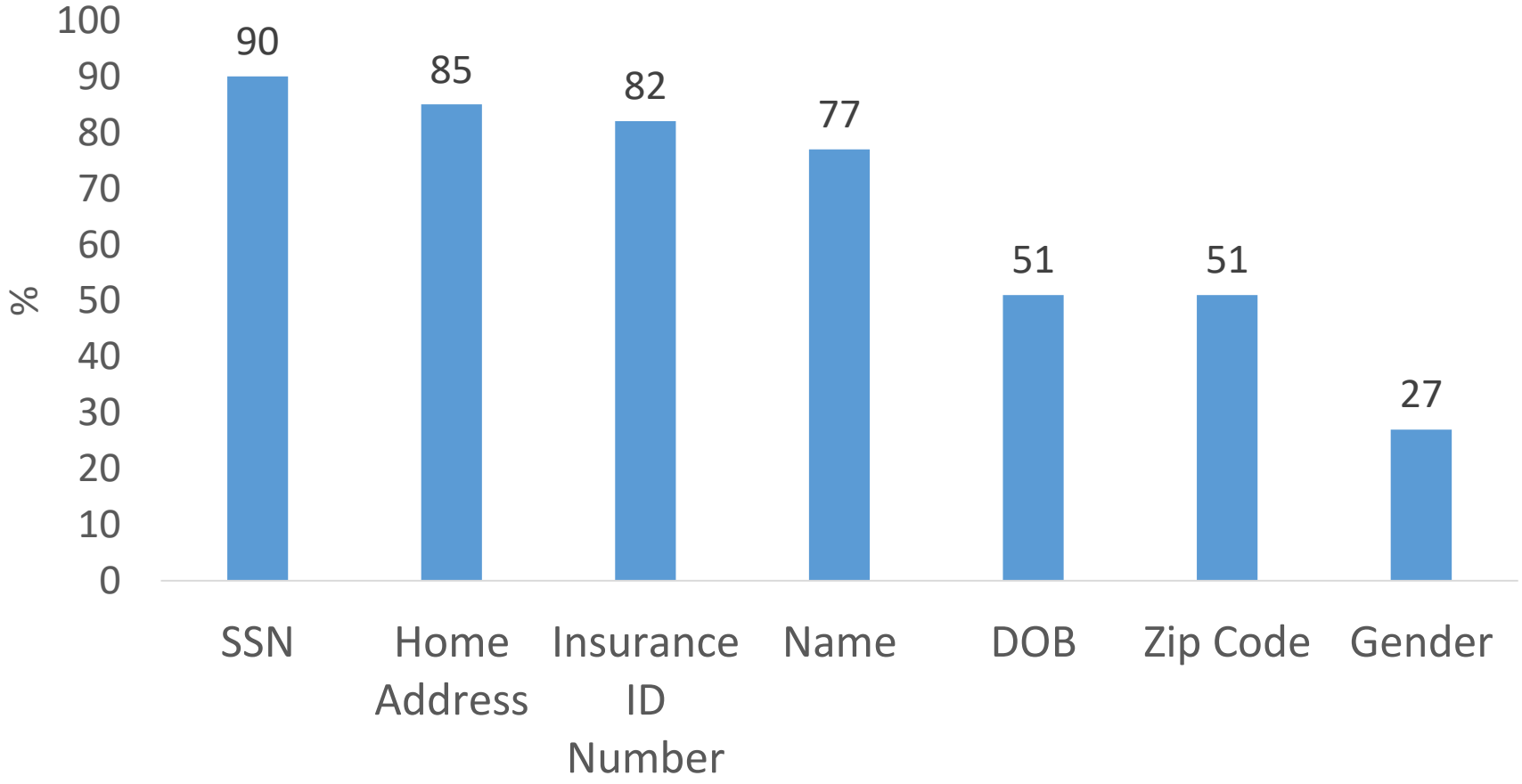
Figure 2. Data Sharing Comfort (n=3516): Local Privacy



More patients are concerned when someone who can recognize them (e.g., someone who knows them) accesses EHR for research

Solution: disclosure control can reduce people who know you, recognizing you

Figure 3. % reporting they would be “extremely” or “much more comfortable” with removal of the following identifiers (n=3516):



What attributes are they most concerned about ?

Solution: Focus on masking IDs, names, and addresses

Goal: Build consensus on template IRB application & DUA when using PPIRL framework

- The questions we plan to ask at the NGT session are:

1. What do you perceive as the benefits when using the PPIRL framework for record linkage?
    - Potentially, allows for linking data that would otherwise not be possible.
    - Encourages use of only needed information, minimizing risk
    - Minimizes risk of re-identification, reduces risk of breach of confidentiality
2. What do you perceive as the risks when using the PPIRL framework for record linkage?
    - Mislinking risks
    - size and quality of data matters; a bad database makes linking difficult when data is masked
    - Disproportionate data sampling, which leads do an increase in bias
3. What other information would you like or need to know when reviewing the IRB application for research?
4. When using the PPIRL framework, what information is needed in the DUA?
    - How to communicate risk to lawyers, so that the risk is stated in the DUA accurately
    - Since the DUA is fixed, how can the software adjust to the DUA
    - Expert determination