

Evaluating linkage accuracy in de-identified data

(A case study from my PhD where I used infant nutrition trials linked to school records)

Maximiliane Verfürden

UCL Child Health Informatics Group

January 2020



Was the participant linked to the
right pupil?

Infant participant
in 1993:



First Names
Maximiliane Lara
Surname
Verfürden
Date of birth
07/12/1993
Postcode at recruitment
E8 1BX

Lara Maximiliane
Verfürden
07 / 12 / 1993
E4 0LZ

Pupil A



?

Maxi
Vaughan
12 / 07 / 1993
E8 1BX

Pupil B



Pupil C etc...

WHO IS IT??

What kind of trials am I following up using the administrative education data and why?

Case study: 7 trial cohorts



**protein
and calorie**

iron

**long-chain
polyunsaturated
fatty acids
(LCPUFA)**

**palmitate
position
(sn-2 palmitate)**

nucleotides

- preterm babies
- small for dates

- healthy infants

- term babies
- preterm babies

- term babies

- term babies

2 trials

1 trial

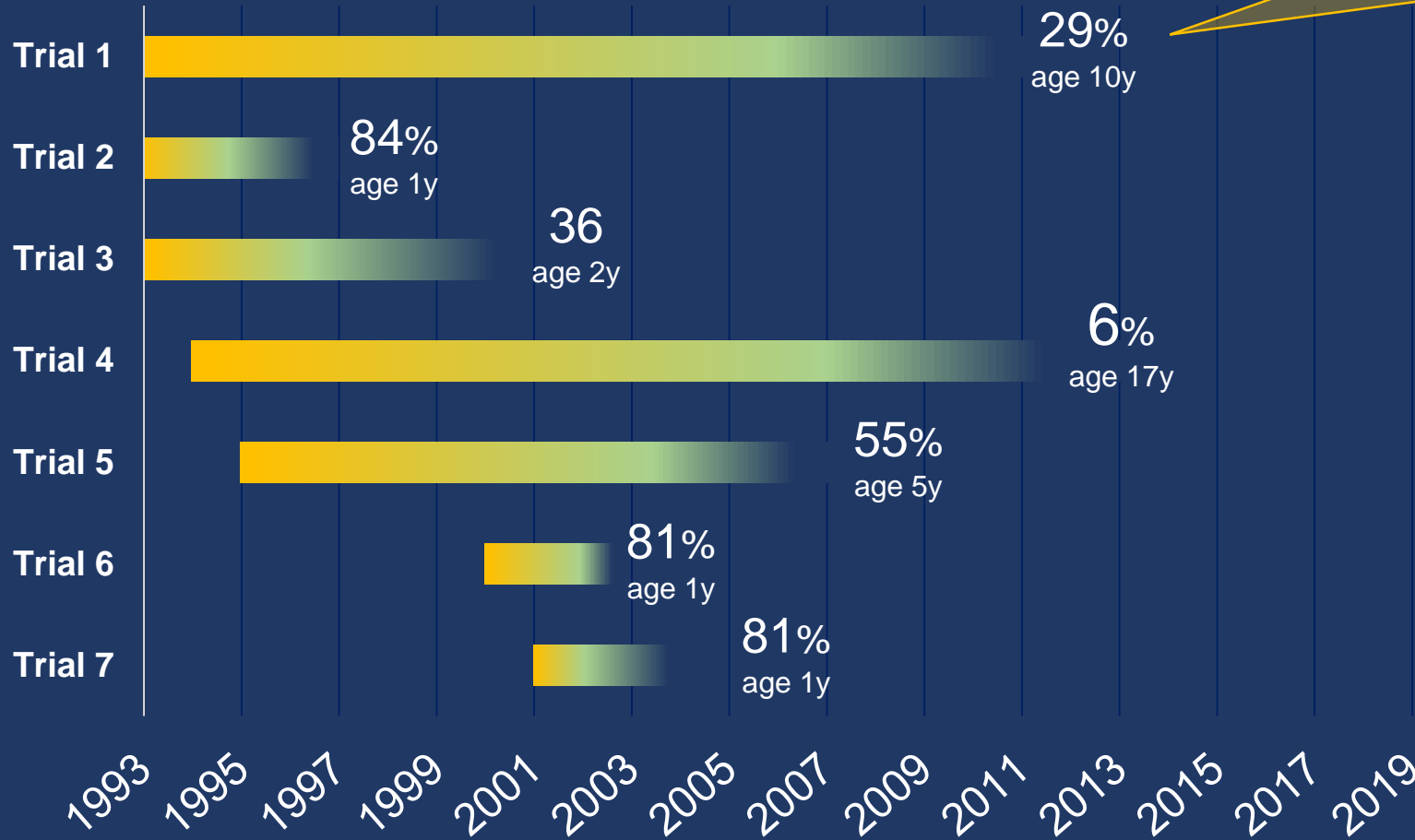
2 trials

1 trial

1 trial

Table 1 Trial populations and interventions

Follow-up in the trials was short and poor

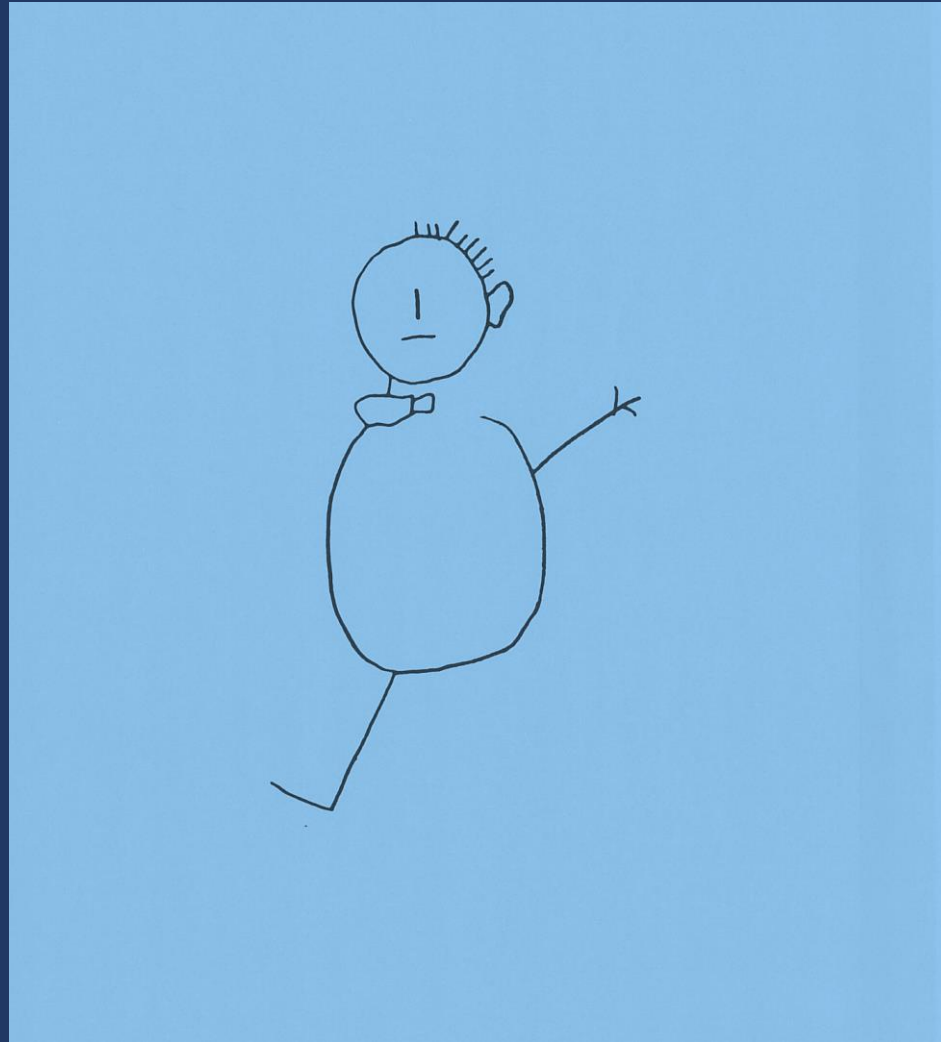


% retention at last follow-up

- Recruitment 1993-2002
- Total 2,788 randomised participants

- youngest ones are 18 years old in 2020

Why even bother?



We have an incomplete picture

To safeguard participant privacy
linkage was done by another institution



Infant participant
in 1993:

First Names
Maximiliane Lara
Surname
Verfürden
Date of birth
07/12/1993
Postcode at recruitment
E8 1BX

Lara Maximiliane
Verfürden
07 / 12 / 1993
E4 0LZ

Pupil A

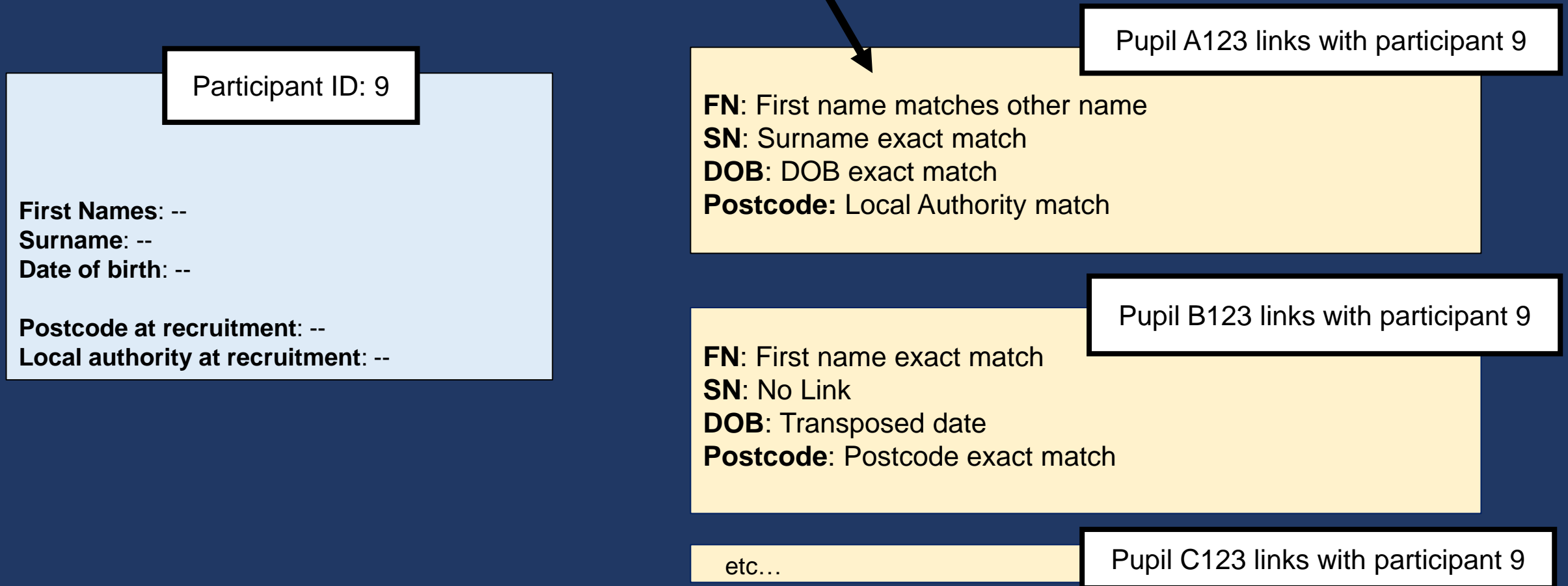
?

Maxi
Vaughan
12 / 07 / 1993
E8 1BX

Pupil B

Pupil C etc...

But because the linked data we receive is de-identified, we requested information on *how* they matched for the best matching three pupils:



Flags available for each participant-pupil pair after linkage

Identifier	Description
First name	First name and other first name both exact match
	First name matches other name in both directions
	First name exact match
	First name matches other name
	Other name exact match
	First name truncated at any hyphen matches
	First name matches via common name alternatives lookup
	Pattern match function - % of 2 letter combinations from longer of two names that don't appear in shorter is 30% or less
	Pattern match function - % of 2 letter combinations from longer of two names that don't appear in shorter is 60% or less - AND first character of first name matches
	First name / surname match in both directions
No Link	
Surname	Surname exact match (including alternative surnames)
	Surname truncated at any hyphen matches
	Pattern match function - % of 2 letter combinations from longer of two names that don't appear in shorter is 30% or less
	Pattern match function - % of 2 letter combinations from longer of two names that don't appear in shorter is 60% or less - AND first character of surname matches
	First name / surname match in both directions
No Link	
Date of Birth	DOB exact match
	Day on source matches month on match, and vice versa; year matches (i.e. transposed date)
	Day and month match (i.e. wrong year)
	Day and year match (i.e. wrong month)
	Month and year match (i.e. wrong day)
	Either source or match DOB is 1st January; year matches
	Either source or match DOB is 1st September; year matches
No Link	
Location	Postcode exact match
	Local Authority match
	Neighbouring / nearby Local Authority match
No checks	

This information is available for each participant-pupil pair:

Trial Participant ID	Pupil ID	Match level for			
		First name	Surname	DOB	Location
9	A123	First name matches other name	Surname exact match	DOB exact match	Local Authority match
9	B123	First name exact match	No Link	Transposed date	Postcode exact match
9	C123	No link	Surname truncated at any hyphen matches	Month and year match (i.e. wrong day)	Nearby Local Authority match
17	Q456
etc

This information can be transformed into match weights:

Trial Participant ID	Pupil ID	Match weight
9	A123	24.6
9	B123	9
9	C123	4
17	Q456	...
etc

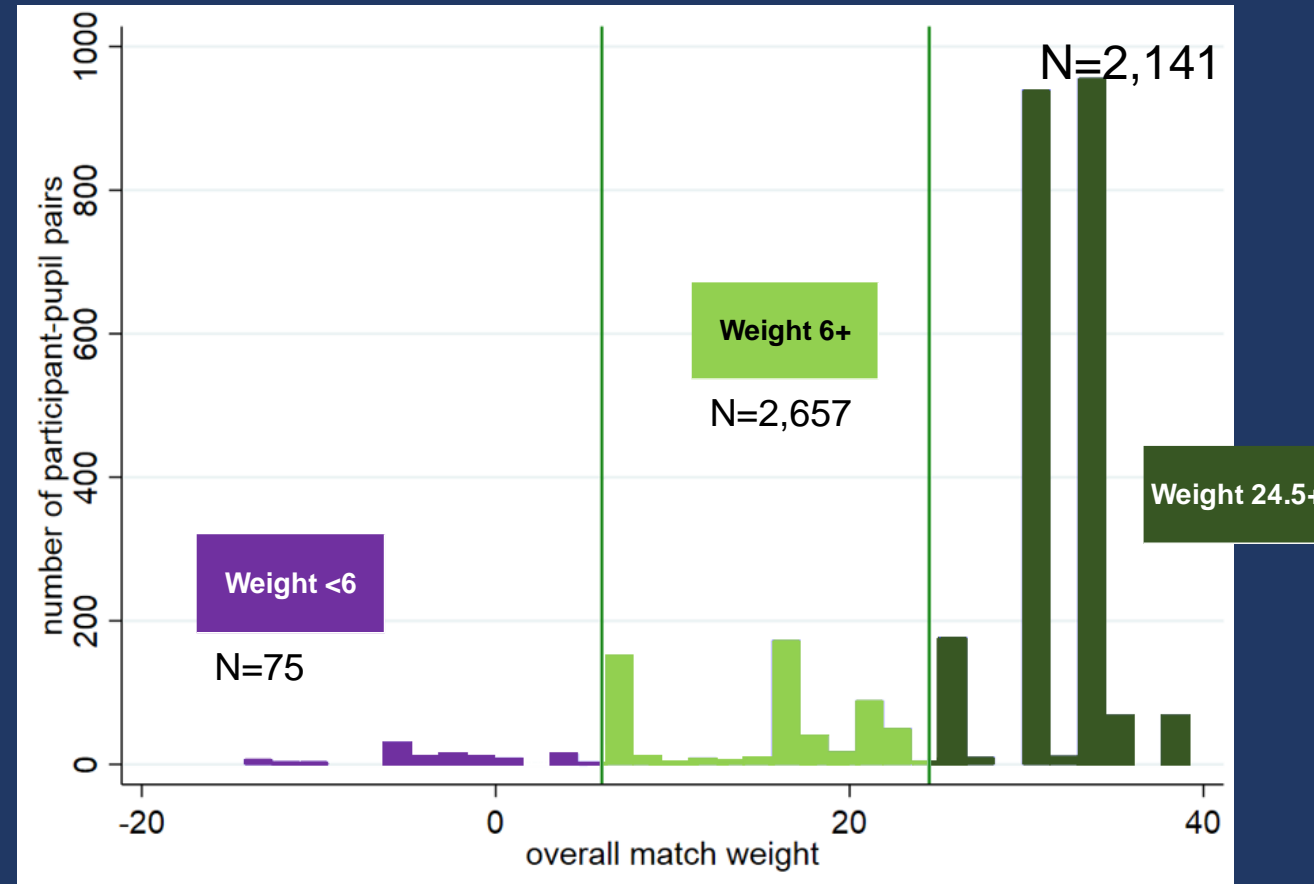
...higher match weights correspond to better fitting identifying information

Further reading: Ivan P. Fellegi & Alan B. Sunter (1969) A Theory for Record Linkage, Journal of the American Statistical Association, DOI: [10.1080/01621459.1969.10501049](https://doi.org/10.1080/01621459.1969.10501049)

Ranking participant-pupil pairs by how well they match

72.2% matched on all identifiers

Trial Participant ID	Pupil ID	Match weight
9	A123	24.6
9	B123	9
9	C123	4
17	Q456	...

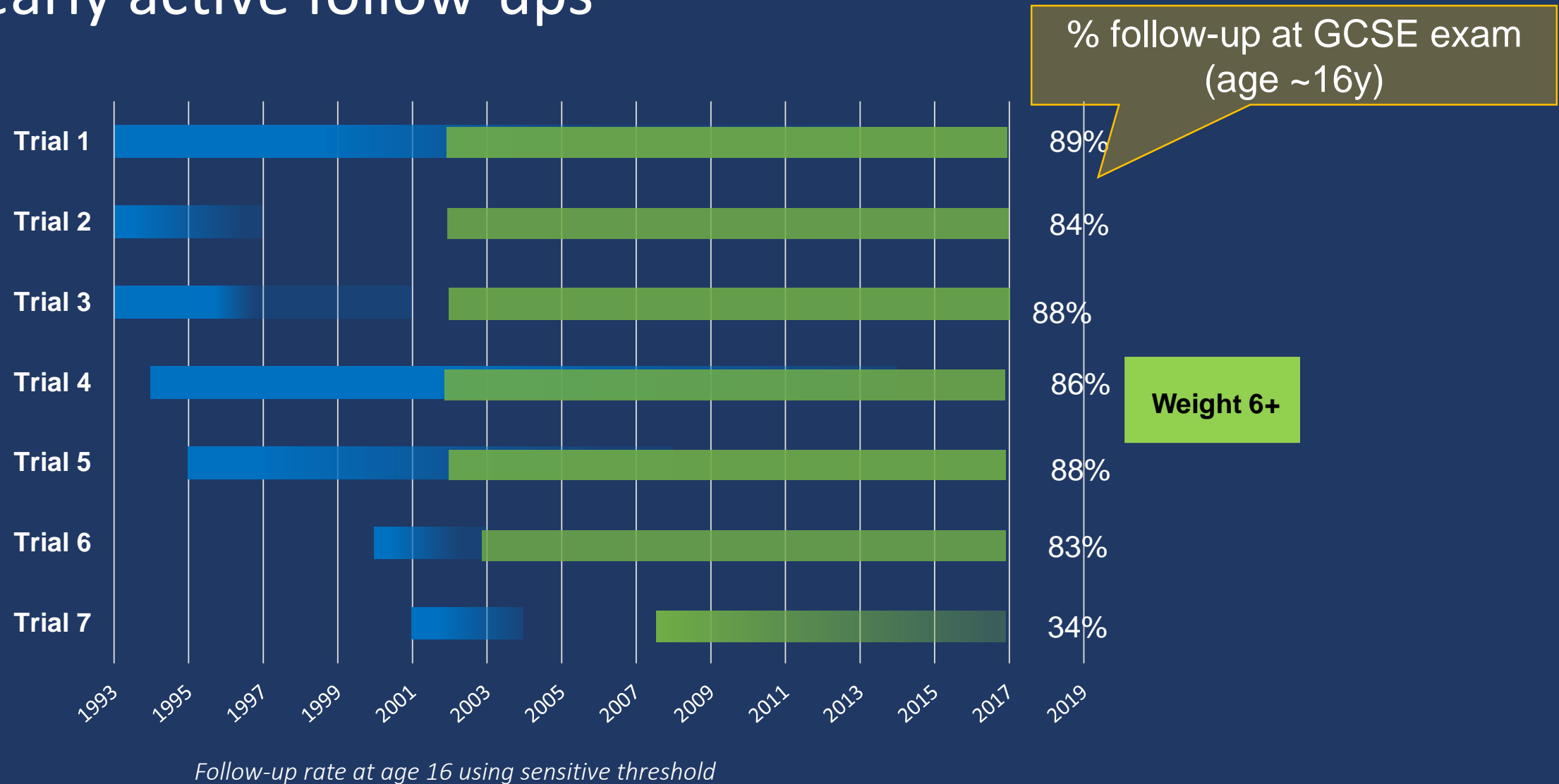


Participant-pupil pairs ranked by match weight

If multiple pupil candidates: deciding the best match using a mixture of review and weights

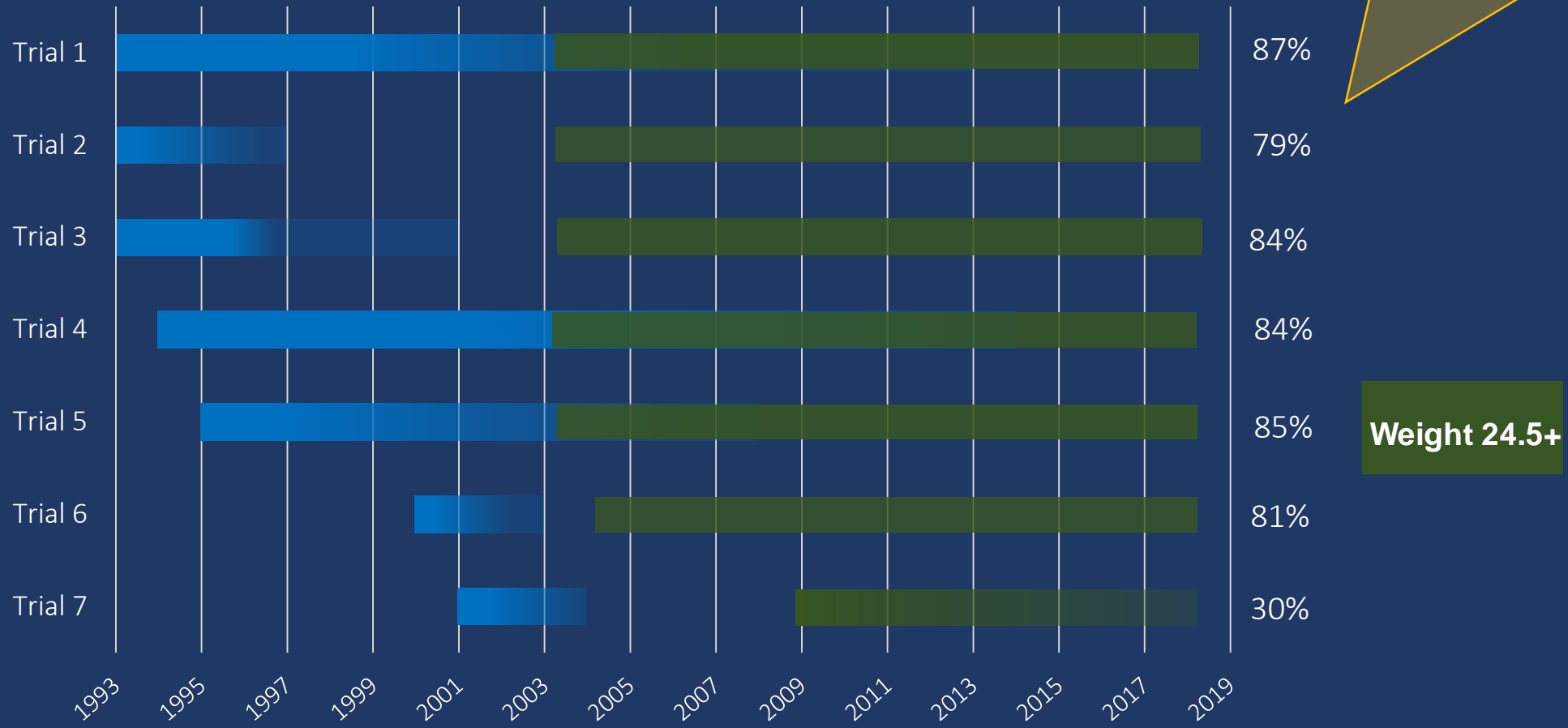
- age at first link was >21
- information on death and time of death available
- if highest match-weight for a participant was $10\% >$ second best match automatically kept the best match. If difference lower, I manually reviewed

Using pairs with weights >5 produces higher follow-up rates than early active follow-ups



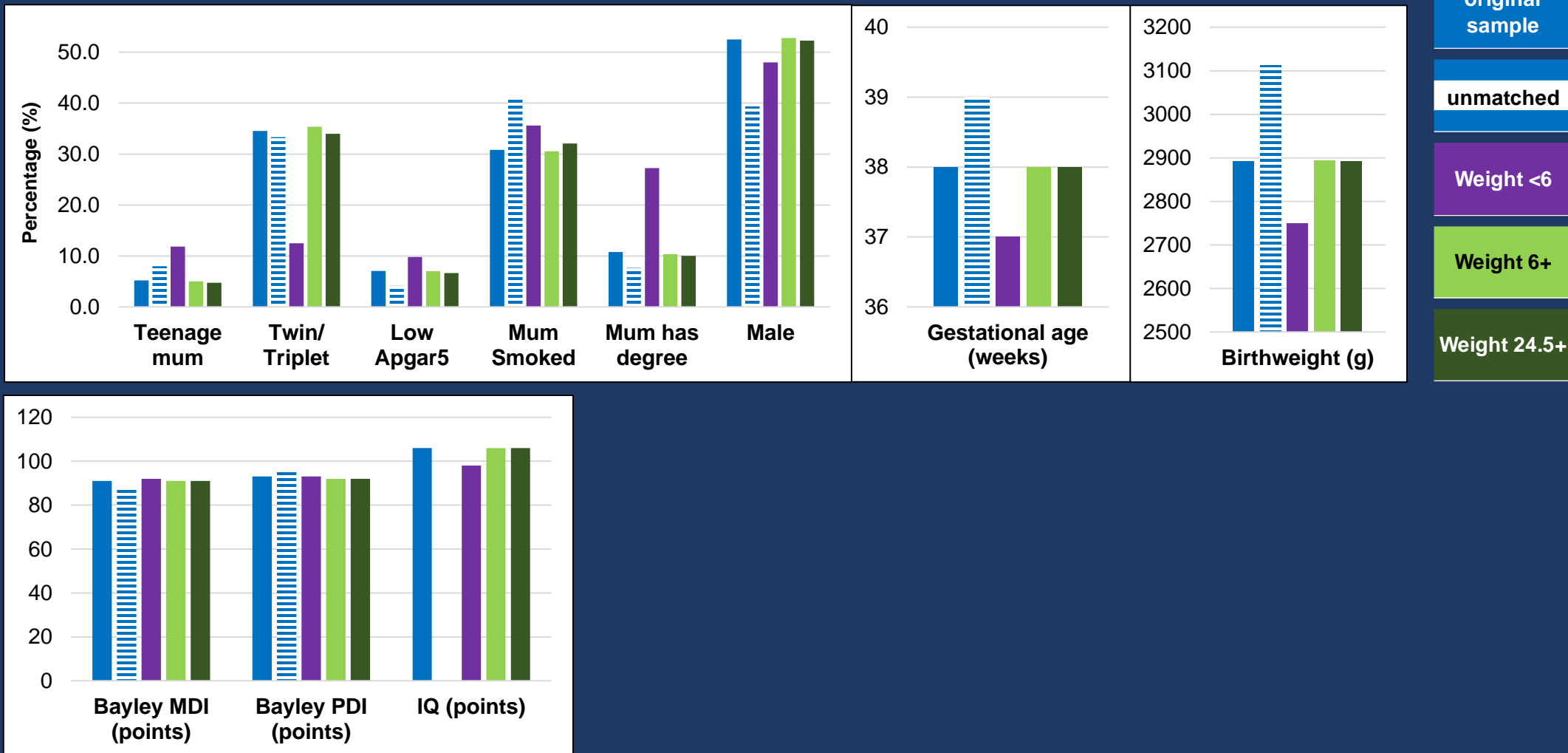
Using only the highest weighted pairs still produces higher follow-up rates than original follow-ups:

% follow-up at GCSE exam
(age ~16y)

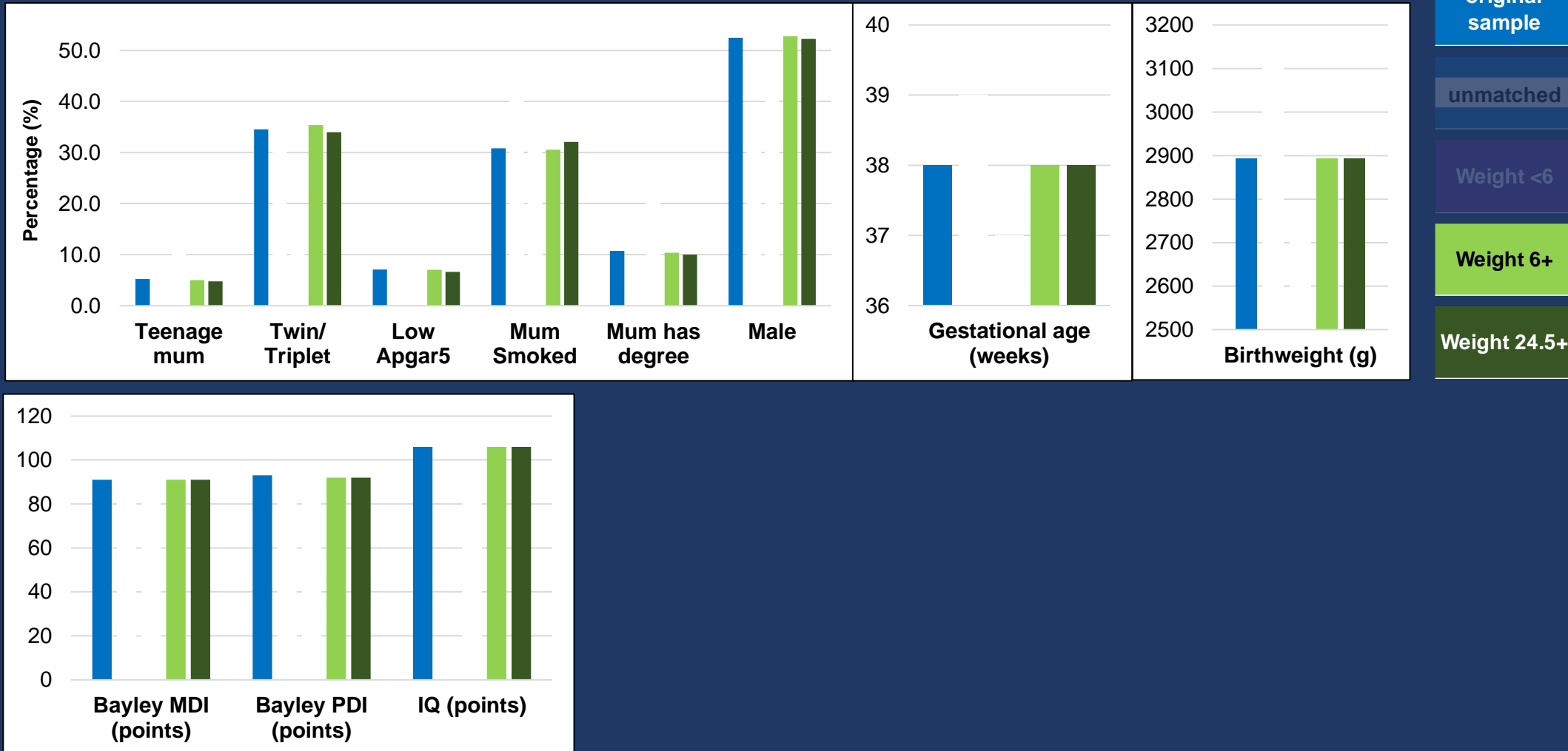


Follow-up rate at age 16 using specific threshold

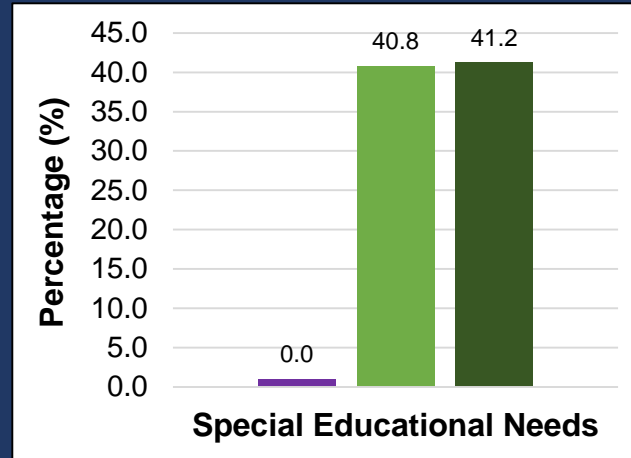
Pairs with high match weights remain representative of original randomised sample



Pairs with high match weights remain representative of original randomised sample



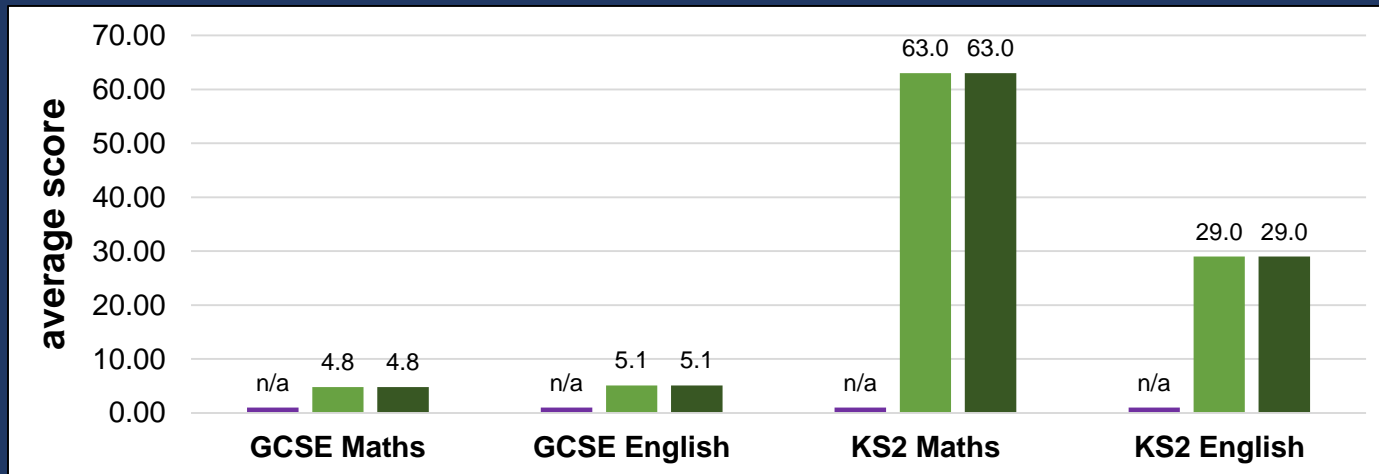
Pairs with high match weights have similar exam scores at age 11 and 16 (low weights have a lot of missing variables)



Weight <6

Weight 6+

Weight 24.5+



Discussion

- Data linkage produces higher retention rates (with fewer resources needed)
- Be aware – its possible to link to wrong pupil records
- ‘Black box’ can be addressed with descriptive linkage flags
- Match weights can help to choose between participant-pupil pairs



Further reading on cost comparisons:
Llewellyn-Bennett et al.
Post-trial follow-up methodology in large randomised controlled trials: a systematic review (2018)
<https://doi.org/10.1186/s13063-018-2653-0>

Thanks to my supervisory panel!



Prof Ruth Gilbert



Prof Mary Fewtrell



Prof John Jerrim



Dr Katie Harron