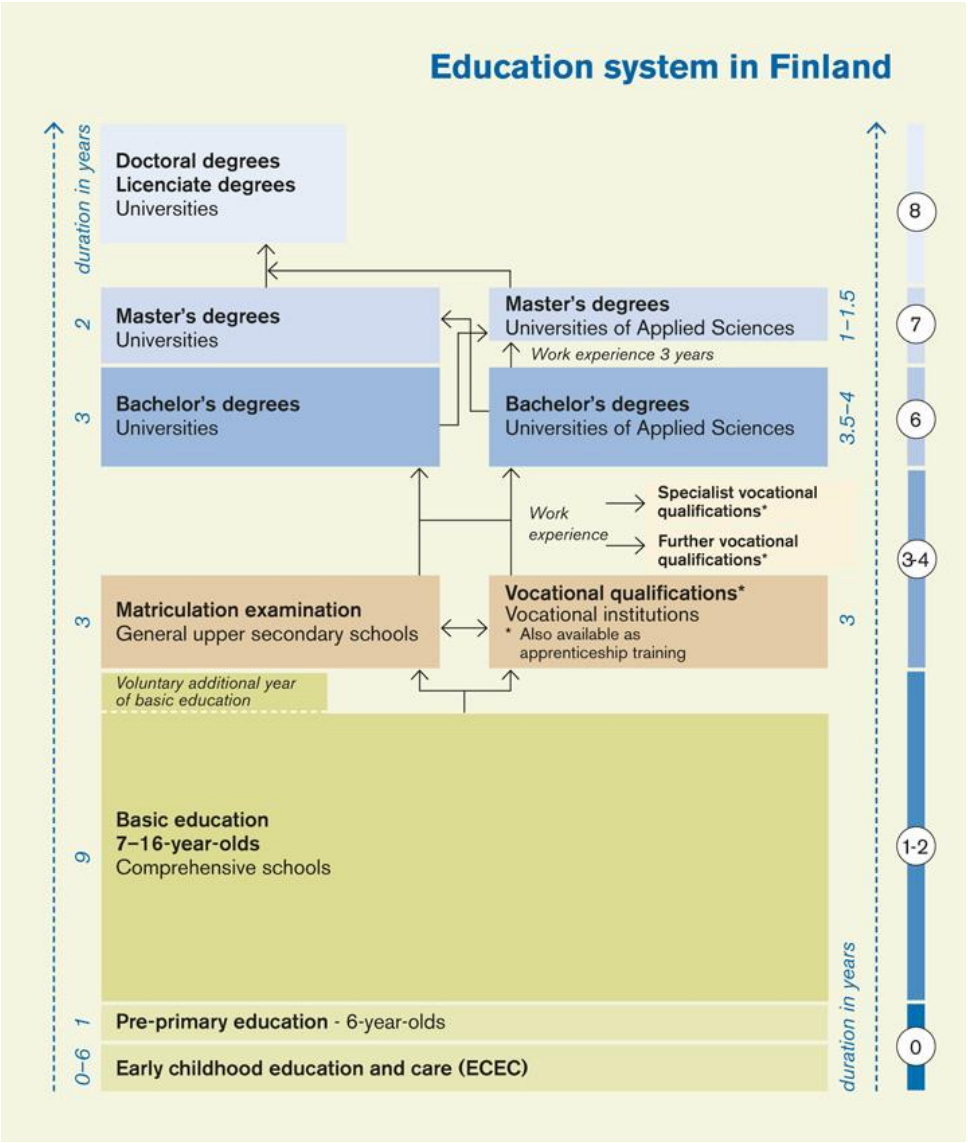
The best of two worlds: Combining longitudinal health and learning to learn surveys with national registry data

Henrik Dobewall, Arja Rimpelä, Lasse Pere, Pirjo Lindfors, Mari-Pauliina Vainikainen, & Sakari Karvonen

Big Data Meets Survey Science (BigSurv18) conference, 27 October 2018

From Data Linkage to Education, 16.00-17.30 (room 40.012)

The education system in Finland

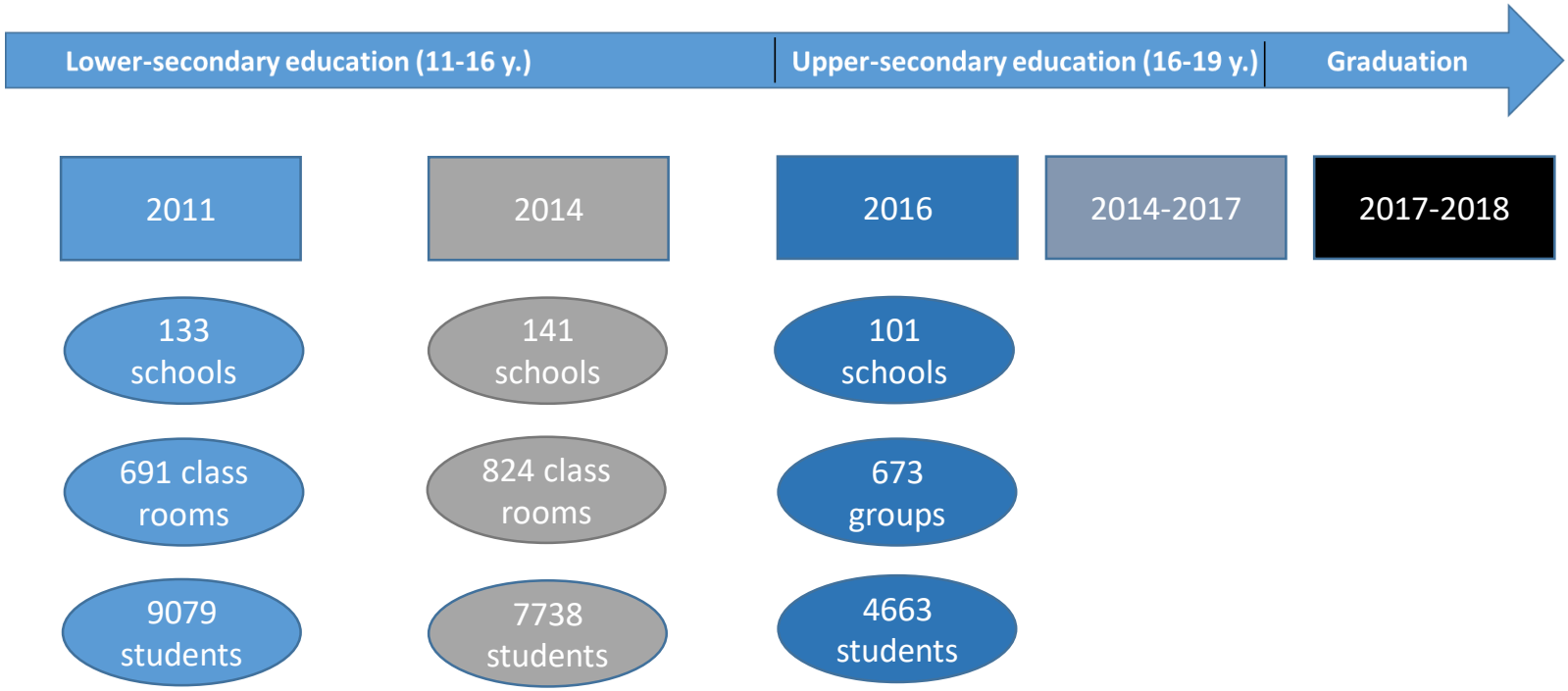


The MetLoFin project

- Our project follows a large cohort of children **from basic to the end of the upper-secondary education** in the 14 municipalities of the Helsinki Metropolitan Region
- These data were collected by the **Center for Educational Assessment** at the University of Helsinki (CEA; led by prof. Risto Hotulainen) and the **Research for Children and Adolescent Health Promotion** research group at the University of Tampere (NEDIS, led by prof. Arja Rimpelä)
- Data on applications for upper-secondary schools were obtained from the **Joint Application Registry** hold by the **Finnish National Agency for Education**

Aim of presentation

- Showing the gains of **combining** students' answers to health and learning to learn **surveys** with national **registry** data
- by use of their ***personal identification code*** (Finnish: henkilötunnus)



Longitudinal health and learning to learn surveys

2014-2017

435+ institutions

192+ fields

14034 (99.5%) students

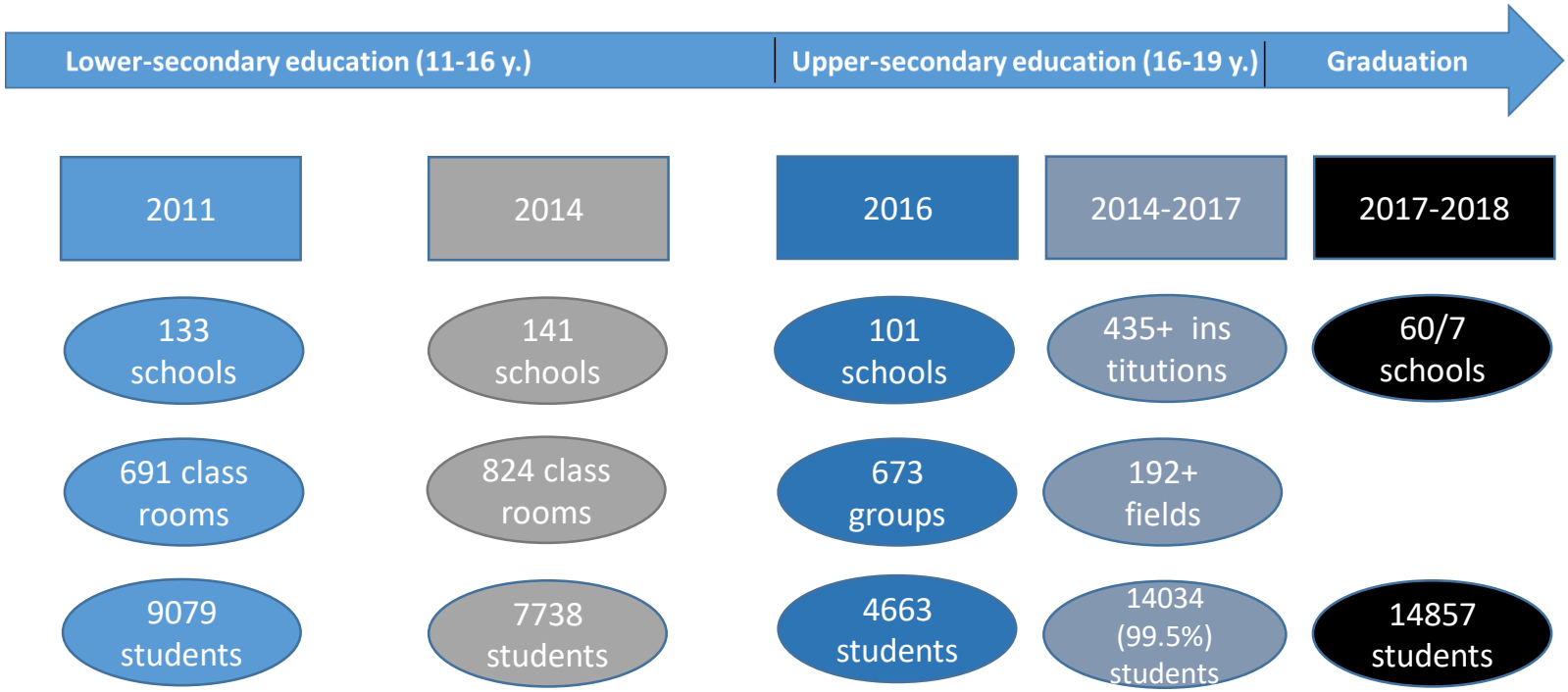
National registry data

2017-2018

60/7 schools

14857 students

Directly from schools

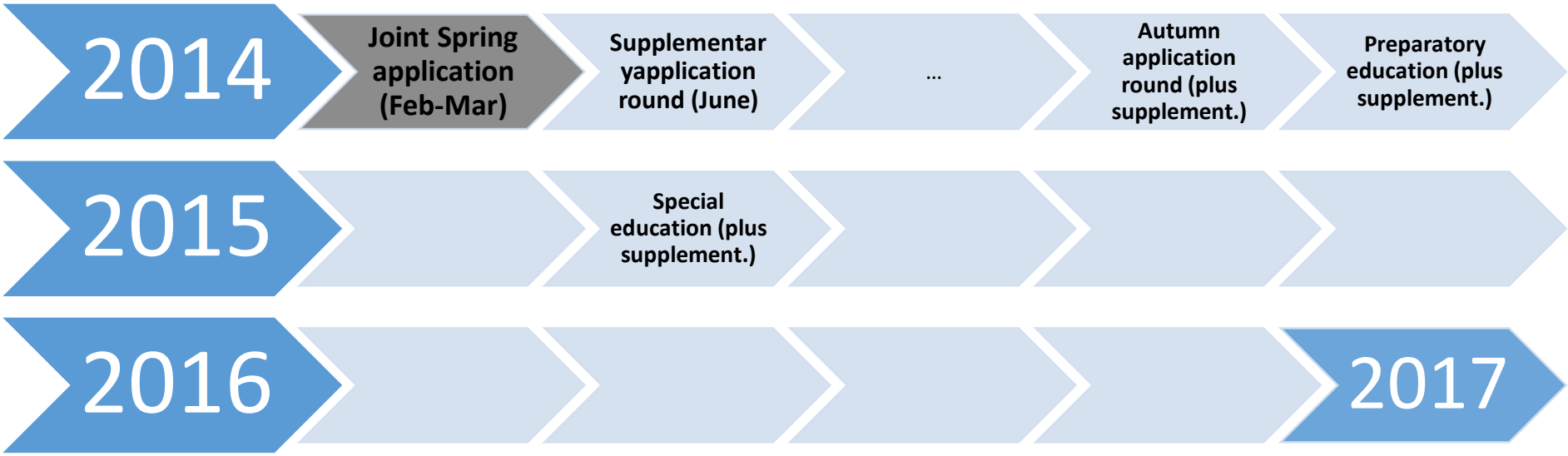


Longitudinal health and learning to learn surveys

National registry data

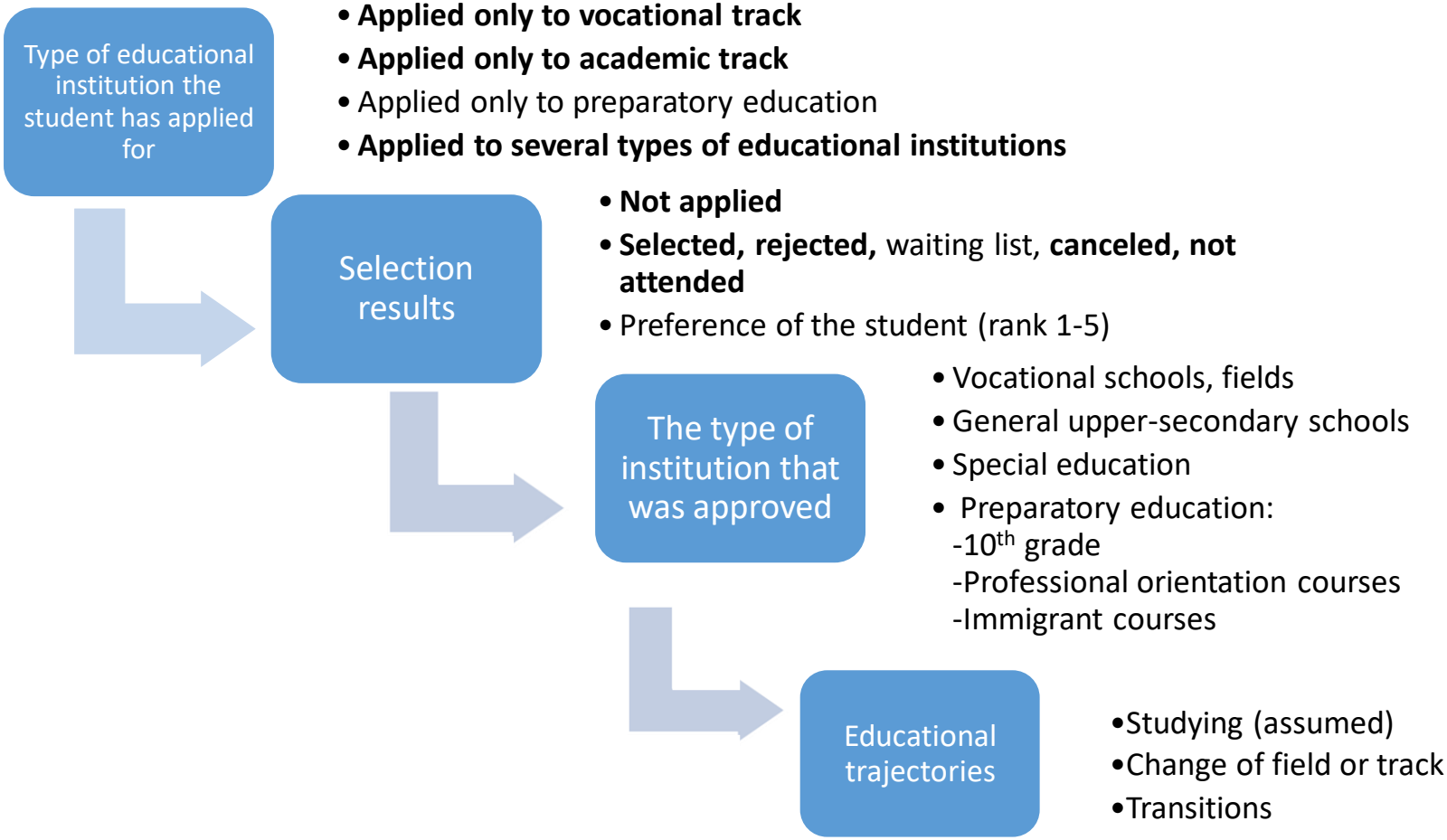
Directly from schools

Timeframe of the Joint Application Registry

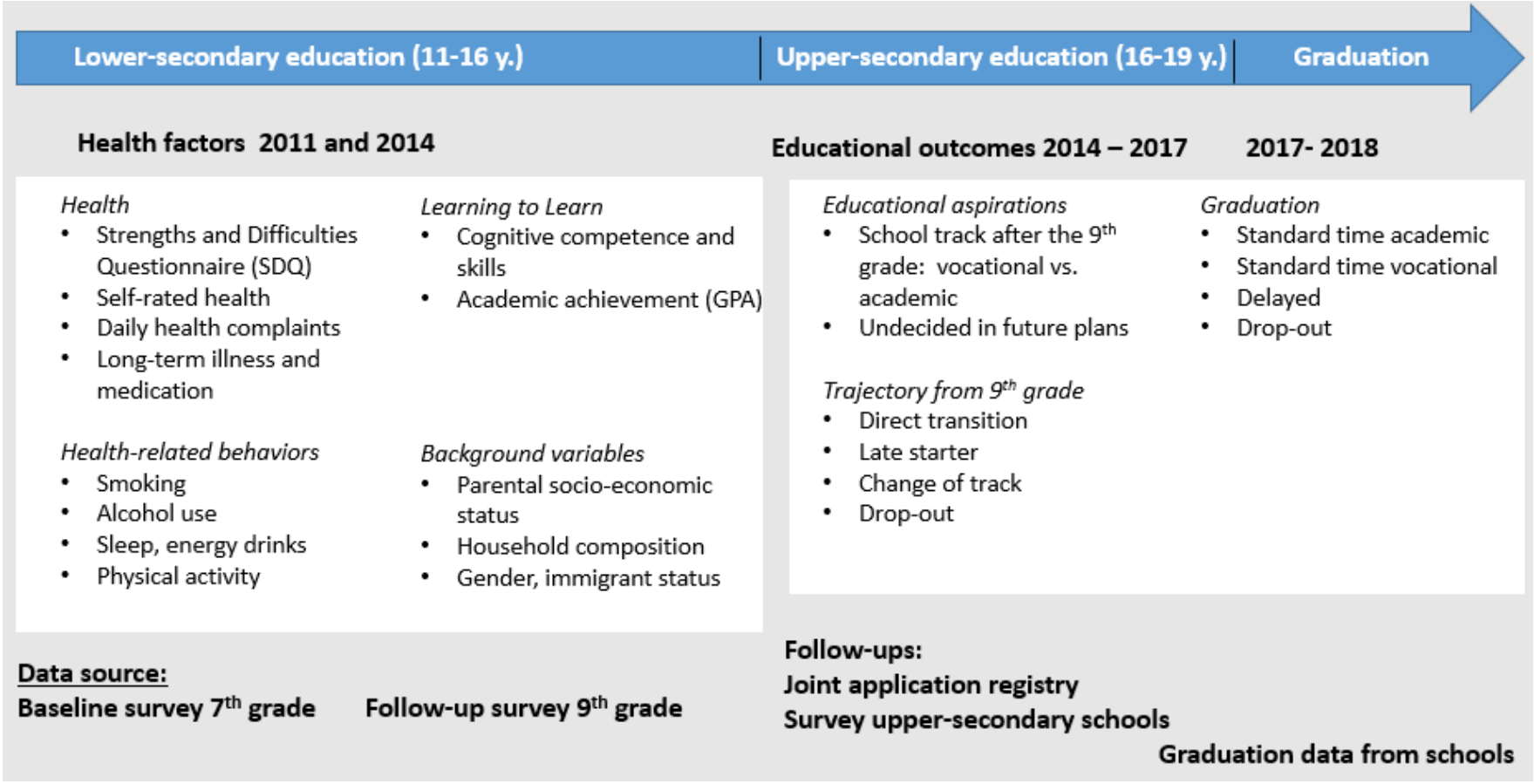


- National registry includes data of **20 application rounds**
- In Spring and Autumn the student **can apply per round to a maximum of 5 (3) study places**
- Study place preference is ranked in the order in which s/he wishes it to be selected
- Plus supplementary application rounds

Joint Application Registry data structure



Individual-level data



Challenges of analyzing, interpreting, and reporting national registry data

- Based on the application data we do not see whether a student had **dropped out** of / **graduated** from upper secondary education → Obtaining graduation data
- No information on **double degrees** available
- **Missing data pattern** in registry data (e.g., some students study without confirming the place)
- About 300 students were **found elsewhere** in upper-secondary schools than suggested by joint application registry → Might be explained by the **youth guarantee** that took effect in 2013

Procedure and ethical considerations

- The study protocols were approved by the Ethical Committee of the National Institute of Health and Welfare
- A **parental consent** was obtained in two municipalities where it was required by the educational authorities, while in the other municipalities information letters were sent to the parents of the student
- The questionnaires were **filled as part of the normal schoolwork**; participation was voluntarily, and the students were instructed that they can decline to answer any question or withdraw from the survey at any time

Data management

- For each student an ID and password were generated to **pseudonymize** the questionnaires/data collection
- Data linkage with the national registry was done at CEA by a data manager who does not analyze the data himself

Consequences of the new European general data protection regulation

Basic principles one should not violate:

- 1) No sensitive information is given out without consent. When it comes to large registries the consent may be overruled due to practical reasons (i.e. too many cases to be contacted individually) AND the new information is not linked with other data that requires consent (in our case questionnaire material)
- 2) Giving out new information is not harmful for the subjects (e.g., criminal records)
- 3) What is new is the more strict regulation on indirect identification

Aim of future linkages is to understand students transition into working life:

- Finnish Linked Employee-Employer Database (FLEED)
- The Compulsory School Registry
- In effect – due to the new regulation – we will have to obtain funding to be able to contact the students and to obtain their consent; without such a permission we cannot conduct the linkage ethically

Empirical example: Health across adolescence and students' educational aspirations

- **Educational aspirations pattern the future of the students to a great extent**
- We test the **selection hypothesis** (West, 1991) which suggests that health can affect educational attainment
- Previous work did not find associations with self-rated health (Madarasova Geckova et al., 2010)
- We tested a broad range of health factors (**SDQ, self-rated health, daily health complaints, long term illness**) and were able to control for students' grade point average and socio-demographic background

Results of the multi-level multinominal logistic regression with base outcome applying to academic track [among 11-12 / 15-16 years olds]

Intra-class correlations school-level = 15.1% / 17.2%.

Applying to vocational track was predicted by

- **Emotional and behavioral problems** (SDQ "Average risk" ref.) regarding "Slightly raised to high risk" **1.40 (1.08-1.82)** / 1.24 (0.95-1.60) and **"**Very high risk" 1.43 (0.96-2.13) / **1.41 (1.03-1.92)**
- And **self-rated health** ("Very or quite good health" ref.) regarding "Average, quite or very poor health" 1.25 (0.97-1.63) / **1.71 (1.33-2.20)**

Being undecided between tracks was predicted by

- **SDQ** ("Average risk" ref.) regarding "Slightly raised to high risk" **1.27 (1.02-1.57)** / **1.30 (1.05-1.60)** and **"**Very high risk" **1.48 (1.06-2.07)** / **1.53 (1.18-2.00)**
- And **self-rated health** ("Very or quite good health" ref.) regarding "Average, quite or very poor health" **1.48 (1.20-1.82)** / **1.27 (1.03-1.57)**

Conclusions

- Using national registry data reduces measurement error and missing data → improved results
- Yet, registry data are not perfect because they were collected for a different purpose than research and tend to include systematic errors
- Combining them with multilevel-longitudinal survey allow researchers to address unique and important research questions
- Future linkages are at least problematic due to the new European general data protection regulation

Thank you for your interest!

Questions: henrik.dobewall@uta.fi

A word on double standards:

There was nothing *murky* about the data collection, which was conducted following normal – and ethically approved – procedures

- https://yle.fi/uutiset/osasto/news/public_health_researcher_apologises_for_murky_data_collection_methods/10387629

UUTISET > NEWS

News 5.9.2018 12:11 | updated 5.9.2018 13:03

Public health researcher apologises for murky data collection methods

Participants in the 600,000-person survey were surprised and concerned to learn that THL had requested their credit profiles for its research.

Recommend 8 people recommend this. Be the first of your friends.

UUTISET > NEWS

News 9.4.2018 17:30 | updated 9.4.2018 17:30

Report: New EU data directive won't prevent Finnish agencies from selling personal data

Last year, Finland's Population Register Centre sold residents' personal data to the tune of 10.5 million euros, according to Lännen Media.

- https://yle.fi/uutiset/osasto/news/report_new_eu_data_directive_wont_prevent_finnish_agencies_from_selling_personal_data/10150955

Educational trajectories (half-a-year sensitive)

	Group	Frequency	Percent
Student **never applied via the joint application system**	**Not in upper-secondary education (n=683)**	71	0,5
Student was **never accepted** for a study place / **cancelled it**		540	3,8
Participating in **preparatory classes** after having previously studied in an upper-secondary school		20	0,1
Student **did not start to study again** after participating in preparatory education		52	0,4
Student in **vocational track since 2014**	**Vocational track (n=5124)**	**4004**	28,4
Change within vocational **track**		361	2,6
Student started later than 2014 to study in vocational track		258	1,8
Successful transition from preparatory education into vocational track		208	1,5
Changing from academic to vocational track		293	2,1
Student in **academic track since 2014**	**Academic track (n=8002)**	**7769**	55,1
Change within academic track		5	0,0
Student started later than 2014 to study in academic track		73	0,5
Successful transition from preparatory into academic track		126	0,9
Changing from vocational to academic track		29	0,2
Student did **not** get the place **via joint application system**	**Excluded (n=296)**	83	0,6
Student had applied for **special education** at some point		213	1,5

- Are there any Big Data approaches to tackle these trajectories, which we could adopt to improve our research?