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
Lower-secondary education (11-16 y.) | Upper-secondary education (16-19 y.) | Graduation

- 2014-2017
- 435+ institutions
- 192+ fields
- 14034 (99.5%) students

National registry data
Directly from schools

- Lower-secondary education (11-16 y.)
- Upper-secondary education (16-19 y.)
- Graduation

- 2017-2018
- 60/7 schools
- 14857 students
Lower-secondary education (11-16 y.) | Upper-secondary education (16-19 y.) | Graduation

2011 |
---|
133 schools
691 class rooms
9079 students

2014 |
---|
141 schools
824 class rooms
7738 students

2016 |
---|
101 schools
673 groups
4663 students

2014-2017 |
---|
435+ institutions
192+ fields
14034 (99.5%) students

2017-2018 |
---|
60/7 schools
14857 students

Longitudinal health and learning to learn surveys
National registry data
Directly from schools

H. Dobewall et al. The best of two worlds
Timeframe of the Joint Application Registry

2014
- Joint Spring application (Feb-Mar)
- Supplementary application round (June)
- Autumn application round (plus supplement.)
- Preparatory education (plus supplement.)

2015
- Special education (plus supplement.)

2016

2017

- 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 Registry data structure

Type of educational institution the student has applied for
- Applied only to vocational track
- Applied only to academic track
- Applied only to preparatory education
- Applied to several types of educational institutions

Selection results
- Not applied
- Selected, rejected, waiting list, canceled, not attended
- Preference of the student (rank 1-5)

The type of institution that was approved
- Vocational schools, fields
- General upper-secondary schools
- Special education
- Preparatory education:
  - 10th grade
  - Professional orientation courses
  - Immigrant courses

Educational trajectories
- Studying (assumed)
- Change of field or track
- Transitions
## Individual-level data

<table>
<thead>
<tr>
<th>Lower-secondary education (11-16 y.)</th>
<th>Upper-secondary education (16-19 y.)</th>
<th>Graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Health factors 2011 and 2014</strong></td>
<td><strong>Educational outcomes 2014 – 2017</strong></td>
<td><strong>2017-2018</strong></td>
</tr>
<tr>
<td>Health</td>
<td>Educational aspirations</td>
<td></td>
</tr>
<tr>
<td>• Strengths and Difficulties</td>
<td>• School track after the 9th grade:</td>
<td></td>
</tr>
<tr>
<td>Questionnaire (SDQ)</td>
<td>vocational vs. academic</td>
<td></td>
</tr>
<tr>
<td>• Self-rated health</td>
<td>• Undecided in future plans</td>
<td></td>
</tr>
<tr>
<td>• Daily health complaints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Long-term illness and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>medication</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Health-related behaviors</strong></td>
<td><strong>Trajectory from 9th grade</strong></td>
<td></td>
</tr>
<tr>
<td>• Smoking</td>
<td>• Direct transition</td>
<td></td>
</tr>
<tr>
<td>• Alcohol use</td>
<td>• Late starter</td>
<td></td>
</tr>
<tr>
<td>• Sleep, energy drinks</td>
<td>• Change of track</td>
<td></td>
</tr>
<tr>
<td>• Physical activity</td>
<td>• Drop-out</td>
<td></td>
</tr>
<tr>
<td><strong>Learning to Learn</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cognitive competence and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Academic achievement (GPA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Background variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Parental socio-economic status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Household composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Gender, immigrant status</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data source:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline survey 7th grade</td>
<td>Follow-up survey 9th grade</td>
<td></td>
</tr>
<tr>
<td><strong>Follow-ups:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joint application registry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey upper-secondary schools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduation data from schools</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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.

Educational trajectories (half-a-year sensitive)

<table>
<thead>
<tr>
<th>Event</th>
<th>Group</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student never applied via the joint application system</td>
<td>Not in upper-secondary education (n=683)</td>
<td>71</td>
<td>0.5</td>
</tr>
<tr>
<td>Student was never accepted for a study place / cancelled it</td>
<td></td>
<td>540</td>
<td>3.8</td>
</tr>
<tr>
<td>Participating in preparatory classes after having previously studied in an upper-secondary school</td>
<td></td>
<td>20</td>
<td>0.1</td>
</tr>
<tr>
<td>Student did not start to study again after participating in preparatory education</td>
<td></td>
<td>52</td>
<td>0.4</td>
</tr>
<tr>
<td>Student in vocational track since 2014</td>
<td>Vocational track (n=5124)</td>
<td>4004</td>
<td>28.4</td>
</tr>
<tr>
<td>Change within vocational track</td>
<td></td>
<td>361</td>
<td>2.6</td>
</tr>
<tr>
<td>Student started later than 2014 to study in vocational track</td>
<td></td>
<td>258</td>
<td>1.8</td>
</tr>
<tr>
<td>Successful transition from preparatory education into vocational track</td>
<td></td>
<td>208</td>
<td>1.5</td>
</tr>
<tr>
<td>Changing from academic to vocational track</td>
<td></td>
<td>293</td>
<td>2.1</td>
</tr>
<tr>
<td>Student in academic track since 2014</td>
<td>Academic track (n=8002)</td>
<td>7769</td>
<td>55.1</td>
</tr>
<tr>
<td>Change within academic track</td>
<td></td>
<td>5</td>
<td>0.0</td>
</tr>
<tr>
<td>Student started later than 2014 to study in academic track</td>
<td></td>
<td>73</td>
<td>0.5</td>
</tr>
<tr>
<td>Successful transition from preparatory into academic track</td>
<td></td>
<td>126</td>
<td>0.9</td>
</tr>
<tr>
<td>Changing from vocational to academic track</td>
<td>Excluded (n=296)</td>
<td>29</td>
<td>0.2</td>
</tr>
<tr>
<td>Student did not get the place via joint application system</td>
<td></td>
<td>83</td>
<td>0.6</td>
</tr>
<tr>
<td>Student had applied for special education at some point</td>
<td></td>
<td>213</td>
<td>1.5</td>
</tr>
</tbody>
</table>

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