Linking Anonymous Data from Student's Questionnaires in Prospective Prevention Trial Using Self-Generated Identification Codes

CHARLES UNIVERSITY
First Faculty of Medicine
Department of Addictology

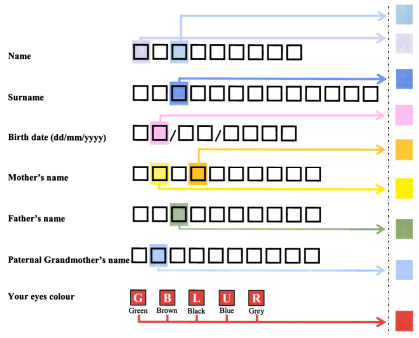
Jaroslav Vacek, Roman Gabrhelík
Department of Addictology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Czech Republic

Introduction

Participation in prevention studies often requires respondents to answer sensitive questions about their risk behavior, such as sexual experiences or substance use. Maintaining anonymity appears necessary not only to obtain valid answers but also to protect the participants and their environment. A self-generated identification code (SGIC) is an anonymous identifier generated from information available to the participant but not to the researcher on the basis of identical instructions for all participants. Linking data on individual level allows to employ statistic methods which outperform those with grouped or aggregated data.

Methods and Sample

We conducted a feasibility study of matching subjects using an anonymous 9-character SGIC adapted from Galanti et al. (2007) and assessed its methodological properties.



Source: Galanti et al. (2007)

Data comes from students who participated in a Czech school-based randomized controlled prevention trial (Gabrhelík et al., 2014). There were five waves of computer-based data collection conducted from September 2013 to December 2015 in 71 schools, with total of 11,492 valid questionnaires collected (initiated with 2,552 sixth-graders, average age 11.94 years, 50.4% female).

Year	2013/14		2014/15		15/16	Total
Wave	**1.**	**2.**	**3.**	**4.**	**5.**	
Collected	2,571	2,316	2,354	2,218	2,133	11,592
No code	19	37	21	19	4	100
Valid N	**2,552**	**2,279**	**2,333**	**2,199**	**2,129**	**11,492**

Those with no code (100 students, less than 1% of all collected questionnaires) were not included in the analysis. Operations with the code included cleaning and converting (e.g. czech digraph CH converted to one character). For linking we used deterministic matching system, pairs were identified on the basis of unique matches of all nine, or a combination of any of eight, seven, six, five, or four characters.

We performed two separated linking procedures for all ten combinations of any two waves. First, we used the SGIC itself (further referred to as to SGIC-only linking). Second, we included the school affiliation variable in the linking procedure in order to reduce the number of possible combinations and thus potentially increase the accuracy of the linking process (further referred to as SGIC+school linking). The efficiency analysis of the anonymous linkage including calculation of precision (positive predictive value), recall (sensitivity, true positive rate) and F-measure (harmonic mean of precision and recall) (Christen & Goiser, 2007) was enabled by use of additional unique numerical anonymous control code (CC) associated with each subject during data collection („gold standard", further referred to as CC linking).

Results

We were able to match 94.7% (18,808) with SGIC-only and 98.4% (19,538) with SGIC+school linking procedure out of all 19,855 possible matches (CC linking). The number of pairs - unique match in a given number of characters shows following table:

Matching characters	SGIC-only			SGIC+school		
	n	%	cumm.	n	%	cumm.
9	10,814	54.5%	54.5%	10,814	54.5%	54.5%
8	4,809	24.2%	78.7%	4,823	24.3%	78.8%
7	2,363	11.9%	90.6%	2,456	12.4%	91.1%
6	652	3.3%	93.9%	873	4.4%	95.5%
5	164	0.8%	94.7%	460	2.3%	97.8%
4	6	0.0%	94.7%	112	0.6%	98.4%
Total	**18,808**	**94.7%**		**19,538**	**98.4%**	

Next table compares frequency of matches in terms of true positives (TP, identical pairs resulting from SGIC and CC linking), false positives (FP, incorrect matches from SGIC linking) and false negatives (FN, incorrect non-matches from SGIC linking). Statistics of matching quality represents precision, recall and F-measure. Precision calculates the proportion of how many of the classified matches (TP + FP) have been correctly classified as true matches (TP). It thus measures how precise a classifier is in classifying true matches. Recall measures the proportion of true matches (TP + FN) that have been classified correctly (TP). It thus measures how many of the actual true matching record pairs have been correctly classified as matches. F-measure as harmonic mean of precision and recall stays high only when both of them have high values.

	SGIC-only	SGIC+school
True positives (TP)	18,808	19,538
False positive (FP)	300	112
False negative (FN)	1,047	317
Precision (P) = TP/(TP+FP)	0.984	0.994
Recall (R) = TP/(TP+FN)	0.947	0.984
F-measure = 2*((P*R)/(P+R))	0.965	0.989

Both linking procedures (SGIC-only, SGIC+school) shows overall very high values of quality statistics.

Following table shows percent of errors in each character of the code in identified pairs by SGIC+school linking (by count of characters used for linking and in total, nine-character match = zero error, i.e. omitted in table). This conformity analysis reveals frequent mutual swap of the characters 1-2 and 5-6 (reverse order of the letters in the given names of a child and mother). The most problematic characters (red color in the table) comes from paternal grandmother's first name and eyes color.

Code characters	Count of matching characters					Total
	8	7	6	5	4	
Third letter of child's first name	5.12%	46.95%	60.82%	75.22%	81.25%	12.12%
First letter of child's first name	2.63%	43.57%	58.88%	72.61%	79.46%	10.92%
Third letter of child's surname	6.88%	7.08%	8.71%	9.57%	18.75%	3.31%
Second digit of day in child's birth date	4.54%	5.78%	8.82%	6.74%	16.96%	2.50%
Fourth letter of mother's first name	8.73%	17.35%	30.24%	68.26%	87.50%	7.80%
Second letter of mother's first name	5.18%	14.66%	28.18%	67.83%	77.68%	6.42%
Third letter of father's first name	8.09%	12.87%	20.39%	27.17%	34.82%	5.36%
Second letter of paternal grandma's first name	30.06%	29.32%	47.42%	40.65%	58.04%	14.52%
Abbreviation (one letter) of child's eyes color	28.76%	22.43%	36.54%	31.96%	45.54%	12.57%

Discussion and Conclusion

Our study indicated favorable outcomes - high number of unique matches and guarantee of anonymity to the participants at the same time. We have proved that SGIC linking with code introduced by Galanti et al. (2007) is both possible and effective, even more with an objective grouping variable. Use of other variables available from the questionnaires (e.g., gender, siblings - number, age, sex) may increase number of questionnaires matched. Based on results of further analyses we recommend to use at least 5 characters of the SGIC for linking.

On the basis of our results and in compliance with the reviewed literature we recommend:

1. maintaining a larger number of the elements on which the SGIC is based;
2. considering the employment of more objective variables (e.g., sex);
3. keeping the procedure for generating the code as uncomplicated as possible (i.e., no change in the order of the letters in the elements);
4. using control mechanisms, e.g., a PC or the Internet, to check whether the codes were completed adequately.

An adjusted form for generating the SGIC based on Galanti et al. (2007) is available as a supplementary material from authors on request.

References

- Christen, P., & Goiser, K. (2007). Quality and Complexity Measures for Data Linkage and Deduplication. In F. J. Guillet & H. J. Hamilton (Eds.), Quality Measures in Data Mining (pp. 127-151). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gabrhelík, R., Orosová, O., Miovský, M., Voňková, H., Berinšterová, M., & Minařík, J. (2014). Studying the effectiveness of school-based universal prevention interventions in the Czech Republic and Slovakia. *Adiktologie, 14(4)*, 403-408.
- Galanti, M. R., Siliquini, R., Cuomo, L., Melero, J. C., Panella, M., & Faggiano, F. (2007). Testing anonymous link procedures for follow-up of adolescents in a school-based trial: The EU-DAP pilot study. *Preventive Medicine, 44(2)*, 174-177.

Contact:
jaroslav.vacek@lf1.cuni.cz

Grant support: This study was supported by the Czech Science Foundation, Grant No. 13-23290S, and Charles University in Prague (PRVOUK-P03/LF1/9). No conflict of interest.