Feasibility Study of Linking Anonymous Data of Children in Longitudinal School-Based Prevention Research

VACEK, J., GABRHELÍK, R.

Charles University, First Faculty of Medicine, Department of Addictology, Prague, Czech Republic
General University Hospital in Prague, Department of Addictology, Czech Republic

Citation | Vacek, J., & Gabrhelík, R. (2024). Feasibility study of linking anonymous data of children in longitudinal school-based prevention research. *Adiktologie, 24*(2), 99–108. https://doi.org/10.35198/01-2024-002-0006

INTRODUCTION: Longitudinal prevention research in adolescents often involves sensitive data collection, necessitating anonymity. Self-Generated Identification Codes (SGICs) have emerged as a tool for linking anonymous data while preserving privacy, but their effectiveness, especially in large-scale studies, remains underexplored. This study aimed to assess the feasibility of using SGICs for linking anonymous longitudinal data in a school-based substance use prevention study.
METHODS: We utilized a three-armed randomized control trial design, utilizing data from the Czech Unplugged Study. The study involved schoolchildren from 71 schools in the Czech Republic, tracked from 6th to 9th grade. SGICs were used to link anonymous survey data across multiple waves. The sample comprised 2,571 pupils, aged 11–13 years, with data collected over seven waves, resulting in a total of 15,289 questionnaires.

RESULTS: The study demonstrated a high rate of SGIC completion (99.2%) and substantial linkability across survey waves. However, errors in SGICs were observed, with certain characters being more prone to inaccuracies. The results showed that 8.5% of all children's SGICs contained one or more missing or erroneous characters; the proportion of errors gradually decreased over time. **CONCLUSION:** The study's findings provide insights into the practicality and challenges of using SGICs in large-scale, longitudinal studies. SGICs offer a viable solution for linking longitudinal data while maintaining participant anonymity. The study highlights the importance of careful SGIC design and the need for further research into optimizing this methodology for large-scale adolescent studies.

Keywords | **Self-Generated Identification Codes (SGICs) – Longitudinal Prevention Research – Substance Use Prevention – School-Based Prevention – Children – Data Linking – Anonymity – Randomized Control Trial**

Submitted | 31 January 2024 **Accepted |** 30 September 2024

Grant affiliation | The study was created as part of the project Development and pilot verification of a regional educational module for the prevention of mental illness in children and adolescents with an emphasis on risky behaviour (ZD-ZDOVA1-012) that was supported by a grant from Iceland, Liechtenstein and Norway within the EEA Funds; the institutional support programme Cooperatio-Addictology.

Corresponding author | Professor Roman Gabrhelík, PhD, Charles University, First Faculty of Medicine, Department of Addictology, Apolinářská 4, 120 00 Prague 2, Czech Republic **roman.gabrhelik@lf1.cuni.cz**

1 INTRODUCTION

Longitudinal prevention research is essential to investigate the effectiveness of prevention interventions to prevent risky behaviour in children (Catalano et al., 2012).

Involvement in a prevention study generally requires children to disclose sensitive information about their risky behaviour, e.g., underage drinking and smoking, substance use, and sexual activity. One challenge is preserving children's anonymity to prevent sensitive information from being leaked and ensure that children provide honest and unbiased answers (Gfroerer & Kennet, 2014). Moreover, storage of identifying information is restricted by recent privacy regulations, such as the General Data Protection Regulations (GDPR) in the EU. Acquiring permission to store information related to individuals can be a time-consuming and bureaucratic process.

A specific methodological issue that arises in longitudinal designs is the linking of anonymous data collected from individual participants at different time points. To address this issue, self-generated identification codes (SGICs) have gained popularity as a means of linking anonymous data in longitudinal prevention studies, where maintaining accurate records can be particularly challenging (Audette et al., 2020; Carifio & Biron, 1978; Schnell et al., 2010; Yurek et al., 2008). However, their methodological properties have remained only little studied.

Originally proposed by Carifio and Biron (1978), SGICs are anonymous personal identifiers that allow participant data to be linked across multiple survey waves while preserving anonymity. SGIC is usually a string of characters, a combination of letters and digits, generated from elements that cover personally relevant information inherent to the individual (e.g., first name, surname, date of birth, mother's and father's names, name of pets, etc.). One element can cover one or more coding questions of the SGIC. Each character represents an answer to the coding question. The number of elements (coding questions) to generate the SGICs usually ranges from three to eight (Audette et al., 2020; Galanti et al., 2007; Schnell et al., 2010; Yurek et al., 2008).

In general, SGICs are usually used in longitudinal studies to track specific participants across multiple waves of data collection (e.g., Bezençon et al., 2023; Calatrava et al., 2021; Damrosch, 1986; DiIorio et al., 2000; De Medeiros et al.; Galanti et al., 2007; Grube et al., 1989; Kristjansson et al., 2014; Lee et al., 2004; Lippe et al., 2019; Little et al., 2021; McAlister & Gordon, 1986; 2024; Seidel et al.; 2022; Wilson et al., 2010). Other studies used SGICs to pair survey data of participants from different study groups, such as parents and children (Fernandez-Hermida et al., 2013; Jessop, 1982; Kandel, 1973; Vacek et al., 2017).

SGICs have several advantages as well as limitations. One advantage is that the use of SGICs does not require researchers to collect personally identifiable information of participants, e.g., names and date of birth. It has been reported that individuals are more willing to participate in the study and provide more accurate answers when they are not required to disclose any traceable information (Audette et al., 2020; Johnson, 2014). Ultimately, SGICs should contribute to improving the response rate and the response validity in research studies (Audette et al., 2020; DiIorio et al., 2000; Kristjansson et al., 2014). Furthermore, the use of SGICs allows researchers to comply with legal and ethical issues associated with adolescent research (e.g., obtaining informed consent and approval from Institutional Review Boards) (Audette et al., 2020; Carifio & Biron, 1978; Kristjansson et al., 2014, Schnell et al., 2010). One potential disadvantage of SGICs is the low linkage performance, specifically the proportion of unmatched cases due to errors in the codes (false-negative matches) and incorrectly matched cases due to low differentiation (false-positive matches) (Schnell et al., 2010).

Several previous longitudinal prevention studies reported overall match rates of successfully paired questionnaires. For example, Grube et al. (1989) used seven-element SGIC in a study on adolescent substance use. They were able to pair 71% of the questionnaires using all seven elements and another 17% when they allowed one character to be erroneous. DiIorio et al. (2000) used eight-element SGIC in a 3-year longitudinal prevention study on HIV in college students. They successfully linked 61.2% of the questionnaires using all eight characters and an additional 11.7% using five to seven characters. Galanti et al. used nine-character SGIC based on seven elements in a multi-centric longitudinal prevention study of substance use to assess the feasibility of linking anonymous data of schoolchildren. Using all nine characters, the proportion of successfully linked SGICs was 61% and increased to 91.9% when using a combination of any six characters. Kristjansson et al. (2014) used five-character SGIC in a longitudinal prevention study with adolescents. They reported an overall match rate of 61%.

Most recently, Calatrava et al. (2022) attempted to employ a fully anonymous coding system that does not rely on personal data. Instead, this system was based on various non-personal elements such as the participant's largest pet, most memorable vacation, most hated food, preferred sport, favorite color and number. Unfortunately, this approach yielded a relatively low success rate in linking data, with only a 30% overall match rate, comprising a 29% recall and an 80% precision rate (Calatrava et al., 2022).

Although other innovative methods for ensuring respondent anonymity in research are emerging, such as the computer-assisted HIDE, CANDIDATE, or BRIDGE methods (Sandnes, 2021b, 2021a, 2021c), these involve a process in which the respondent enters a name which is not saved but directly transformed into a human-readable code through an automatic and instant process of phonetic transformation, hashing, and truncating. Currently, there are no results about matching rate available from practical studies conducted on real subjects using these techniques.

Previous studies reported varying success in linking participant data. According to the literature, to lower error and omission rates, SGICs should satisfy the following characteristics: (1) stability, or the capacity of elements (coding questions) to produce the same answer each time an individual is asked to answer this question, to ensure that the code remains constant even over extended time periods; (2) variability, or the capacity of response options to differ among participants, to ensure that each participant generates a unique code; and (3) proximity, or personal relevance (salience) of the elements to the participant, to ensure that information is easy to remember and difficult to forgot (Audette et al., 2020; DiIorio et al., 2000; Yurek et al., 2008).

So far, little research has systematically studied the methodological properties and efficiency of SGICs in large-scale longitudinal experimental prevention studies involving schoolchildren. Therefore, the aim of this article is to describe the feasibility of linking anonymous longitudinal data using SGICs in schoolchildren participating in a randomized controlled trial focused on substance use prevention. Specifically, this study aimed to (1) estimate the methodological properties of the code characters and elements by analyzing errors and omissions in the codes and identifying the components most prone to errors in SGICs; and (2) evaluate the efficiency of linking anonymous longitudinal data of children.

2 METHODS

2.1 Design

This is a feasibility record-linkage study of linking anonymized longitudinal data of schoolchildren using SGICs.

We utilized data from the Czech Unplugged Study, a school-based three-armed randomized control trial to prevent and reduce adolescent substance use in the Czech Republic (Gabrhelík et al., 2014). The methodology was previously described in Vacek et al. (2017), evaluating the feasibility of between-group linking of children-parents data using SGICs.

2.2 Sample

All schoolchildren in sixth grades (11–13 years) from 71 randomly selected schools in 4 regions in the Czech Republic were enrolled in the 2013–2014 school year to take part in the study. These 71 schools represented 133 classes with 3,017 pupils, of which 446 were not present in the first wave of data collection. The participation of 2,571 pupils with average age of 11.94 years represents 85 % of sample set of research. A more detailed description of the procedure, justification of the selection of regions and the size of the set can be found in the publication of Gabrhelík et al. (2014). Pupils were monitored from the 6th to the 9th grade, i.e. in the period from the pretest (before the start of the intervention in September 2013) to the last retest (36 months after the intervention in June 2017) for a period of 44 months. The data collection always took place at the beginning and end of the school year, in early autumn and late spring, except for the 9th grade, in which only the final testing took place (7th wave of data collection). A total of 1,728 pupils from 61 schools took part in the latest testing in 9th grade, which represents more than two-thirds of the pupils who took part in the first wave. During seven waves, pupils filled out a total of 15,289 questionnaires. *Table 1* shows the number of participating respondents, retention in the study across waves, and the rate of successfully completed SGICs. The decrease in the count of participants was stable over time and initial sample was reduced to about two-thirds (67.2%) compared to the first wave. Overall, less than one percent of pupils did not fill in the anonymous identification code, which is roughly true in all waves. There were no statistically significant differences in sociodemographic characteristics (gender, type of school, region, experimental group) between participants who completed the SGIC and those who did not.

2.3 Data collection process

We used computer-based online questionnaires to collect the data from the children in each school. To preserve anonymity and at the same time be able to match data from each participant in all waves, we used two types of codes. Besides the SGIC, information on which questionnaires were completed by the same respondent was ensured through automatically generated unique identification numerical codes (UINC). In each wave of data collection, each participating child received an UINC, a 10- to 13-digit verification variable automatically generated by the system, which was used both to enter the online questionnaire and to match questionnaires in all waves. A discreet envelope for each child and a class record sheet was brought to school by the research assistant. The teacher was given a class record sheet for distributing discrete envelopes. The sheet contained the serial number of the envelope, which was printed on the visible side of the envelope (UINC was printed on the inside and was not public visible). From the moment it was handed over to the teacher, the interviewer no longer met the sheet and was not allowed to look at it. The teacher wrote the children's names into the list on the sheet to the sequential numbers. The child received a corresponding envelope with the same serial number as in the class record sheet. Children opened a discrete

Table 1 | Sample, retention rate, SGIC completion rate

Count of questionnaires	Wave of data collection							Total
	1	2	3	4	5	6	7	
Total collected n	2 571	2 316	2 354	2 218	2 133	1 969	1 728	15 289
Retention rate[i] %	100.0	90.1	91.6	86.3	83.0	76.6	67.2	
SGIC completed n	2 553	2 279	2 340	2 206	2 130	1 941	1 721	15 170
SGIC completion rate[ii] %	99.3	98.4	99.4	99.5	99.9	98.6	99.6	99.2

Notes: i) % of wave 1, ii) % of all with completed SGIC envelope and have rewritten a multi-digit numerical code into the computer to enter the questionnaire. Trained administrator provided children the pen-and-paper SGIC sheets; children were instructed to self-generate their SGICs based on instruction written on the sheets and to transcribe the characters into online questionnaires. The class record sheet with the children's names and envelope serial numbers was kept by the teacher and had been placed in a safe at school for use in the next test wave. In subsequent waves, the envelopes were distributed according to the serial numbers listed in the class record sheet - the sheet was exclusively manipulated by the teacher. After completing the questionnaire, the children tore open the discreet envelope and the SGIC sheet. Torn envelopes and forms were placed in a bin and shuffled so that no one else could acquire their personal information used to generate the code or match serial numbers of envelopes with UINC.

2.4 Self-Generated Identification Code

We used Galanti's adaptation of SGIC (Galanti et al., 2007) to link children's anonymous longitudinal data. Galanti's SGIC meets the basic requirements for anonymity while sufficient stability, variability, and proximity could be expected.

Generating a nine-character code consisting of a combination of letters and digits requires answering a predetermined set of coding questions based on the child's personal information. These include the third and first letters of a child first name, the third letter of their surname, the second digit of the number of the day of their date of birth, fourth and second letters of mother's first name, the third letter of father's first name, the second letter of paternal grandmother's name, and the first letter of child eye colour.

With regard to the content and graphical adaptation, we used SGIC sheets identical to those of Galanti et al. (2007). We used the same derivation rules for typesetting (the use of upper-case letters, distinguishing between the number "0" and the capital letter "O") and handling missing digits. Original instructions to fill out the SGIC were supplemented regarding the Czech language-specific features, i.e., using only characters included in the Czech alphabet, restriction of diacritics use, and use of the "Ch" as a one-letter character.

2.5 Linking procedure

To link children's anonymous data across survey waves, Schumacher's deterministic matching system was used (Schumacher, 2007). Pairs were identified based on unique matches at the level of all nine characters or a combination of any eight, seven, six, five, or four characters. Initially, we sorted out only those pairs linked in all nine characters (fully matched). We then applied Schnell's off-one rule (Schnell et al., 2010) to link the data—the remaining codes were declared as "matched" if they differed at most in one character. We repeated the off-one rule until the four-character level.

Microsoft Office Excel® and the Fine-grained record integration and linkage tool (FRIL; Jurczyk et al., 2008) were used for the linking procedures. We generated shortened variations of SGICs representing all possible combinations at the level of eight to four characters for each participant. Then, we compared full-length and shortened codes (if they had at least four identical characters in the same positions). Only unique pairs at each level of code length (nine to four) were considered a match.

Two separate linking procedures were executed: (1) SGIC-only linking, where we used SGICs itself to link the data and (2) SGIC+school linking, where we included school affiliation in the linking procedure to increase its accuracy via less number of compared codes.

2.6 Statistical analysis

Overall error rate

A simple descriptive analysis of missing or erroneous characters (omitted or mistyped) was performed to calculate the overall error rate of SGICs.

Overall error rate in pairs

The proportion of matched pairs with discrepant characters at each position of the code and the overall proportion of discrepant characters in pairs were calculated to determine the components that contribute most to errors in the SGICs.

Overall error rate and analysis of discrepant characters in pairs allowed us to estimate the methodological properties of characters and elements of SGIC in terms of variability, stability, and proximate relevance to the participant.

Overall match rate

The overall match rate was calculated to determine the efficiency of SGIC-only and SGIC + school linking procedures. First, we matched the children's data using UINC. These codes remained identical throughout all survey waves. We then calculated the overall match rate as the ratio of successfully paired SGICs divided by all pairs matched using unique computer-generated codes.

Quality measures

Precision, Recall, F-measure were calculated based on proportions of true positives (TP), false positives (FP), and false negatives (FN) matches in SGIC-only and SGIC + school linking to determine the quality of anonymous linking.

Precision (positive predictive value) refers to the precision of the classifier in classifying true matches. Precision was calculated as the proportion of classified matches (TP + FP) from correctly classified true matches (TP) according to the formula (Christensen & Goiser, 2007):

$$Precision = \frac{TP}{(TP + FP)}$$

Recall (sensitivity, true positive rate) is the proportion of true matches (TP + FN) correctly classified as matches (TP) calculated according to the formula (Christensen & Goiser, 2007):

$$Recall = \frac{TP}{(TP + FN)}$$

F-measure (f-score) refers to the overall accuracy of anonymous linking and was calculated as a harmonic mean of precision (P) and recall (R) according to the formula (Christensen & Goiser, 2007):

$$Recall = \frac{Prec \times Rec}{Prec + Rec} = \frac{2TP}{2TP + FP + FN}$$

Accuracy - overall correctness, classification accuracy - indicates the proportion of correctly classified codes out of all possible ones, it also takes true negatives into account.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

2.7 Ethics

This study was approved by the Institutional Review Board of the General University Hospital in Prague.

3 RESULTS

3.1 Overall error rate

Out of the total of 15,170 SGICs, 1,289 (8.5%) included at least one missing or erroneous character (*Table 2*). Of the total 136,530 SGIC's characters (15,170 codes times 9 SGIC's characters in each code), 1,413 (1.03%) were missing or erroneous. With each wave of data collection, the proportion of SGICs with missing or erroneous characters decreased. In the first wave, we revealed 14.06% of SGICs with missing or wrong characters, while in the last seventh wave it was 4.59%.

Overall, the exceptionally high percentage of children made a mistake in the eighth character of SGIC – the second letter of paternal grandmother's first name (7.82% of all SGICs, 84.01% of all errors). Otherwise, in other components, the proportion of missing or erroneous characters was low (< 1%). We found no errors in the first and second components of SGIC the letters of children first names. A minimum count of errors were also found in the third, fourth and sixth components.

3.2 Overall error rate in pairs

Overall, the highest error rates (14% and 12.6%) showed the eighth character (the second letter of the paternal grandmother's first name) and the ninth character (an abbreviation of the child's eye colour) of the SGIC (*Table 3*).

High error rates (12.4% and 11.3%) were also found for the first and second characters, which coded the third and first letter of the child's first name, respectively. In total, 3,738 pairs (10% of all possible ones) had discrepancies in both the first and second characters. In 94% of them, we found that the discrepancies were caused by an interchange in the order of the letters.

We observed a similar confusion for the fifth and sixth characters that coded the fourth and second letters of the mother's name. Of the 1,641 pairs (4.4% of all possible), the interchange of the order of the characters caused an error in 70.3%.

Table 2 | Frequency and percentage of missing or erroneous characters in children SGIC

Character of SGIC	Survey wave							Total errors		
	1	2	3	4	5	6	7	In all waves	% of all codes	% of all errors
1 - Third letter of child's first name	-	-	-	-	-	-	-	-	-	-
2 - First letter of child's first name	-	-	-	-	-	-	-	-	-	-
3 - Third letter of child's surname	-	1	-	-	-	-	-	1	0.01	0.07
4 - Second digit of day in child's birthdate	3	1	-	-	1	-	-	5	0.03	0.35
5 - Fourth letter of mother's first name	5	5	14	11	8	5	1	49	0.32	3.47
6 - Second letter of mother's first name	-	4	-	-	-	-	-	4	0.03	0.28
7 - Third letter of father's first name	24	21	18	16	10	10	11	110	0.73	7.78
8 - Second letter of paternal grandmother's first name	345	237	177	126	127	101	74	1 187	7.82	84.01
9 - Abbreviation (one letter) of colour of child's eyes	5	8	8	11	8	13	4	57	0.38	4.03
Total errors	382	277	217	164	154	129	90	1 413		100
Total codes with errors	359	248	193	149	144	117	79	1 289	8.50	
% of codes with errors	14.06	10.88	8.25	6.75	6.76	6.03	4.59	8.50		
N (total codes)	2 553	2 279	2 340	2 206	2 130	1 941	1 721	15 170	100	

Table 3 | The proportion of pairs with discrepant characters at each code position

	Sum of matching characters in a real pair*								
Character of SGIC	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**Total (%)**
1 - Third letter of child's first name	5.3	46.5	61.2	79.3	83.2	90.5	94.1	100	12.4
2 - First letter of child's first name	3.0	43.4	58.9	76.8	83.2	87.3	100	100	11.3
3 - Third letter of child's surname	6.9	6.6	9.4	8.2	17.5	23.8	41.2	0	3.2
4 - Second digit of day in child's birthdate	4.6	5.5	9.4	7.0	16.2	15.9	41.2	100	2.5
5 - Fourth letter of mother's first name	8.5	17.5	29.8	68.1	85.8	93.7	94.1	100	8.0
6 - Second letter of mother's first name	5.5	15.4	26.6	68.0	80.6	90.5	88.2	100	6.9
7 - Third letter of father's first name	8.4	12.9	19.8	24.4	31.7	58.7	82.4	100	5.4
8 - Second letter of paternal grandmother's first name	28.9	29.4	45.2	37.7	57.0	88.9	94.1	100	14.0
9 - Abbreviation (one letter) of color of child's eyes	29.0	22.8	39.5	30.5	45.0	50.8	64.7	100	12.6

* The column with 9 matching characters is omitted from the table, as this indicates a complete match and a zero error rate for each character.

3.3 Total match rate/efficiency of linking

We were able to link 37,313 using unique computer-generated codes each child received prior to data collection (*Table 4*).

Using any combination of four or more characters, we were able to link 36,308 pairs (97.6% of all real pairs) in SGIC-only linking and 37,205 (99.7% of all real pairs) in SGIC + school linking (Table 3).

Using all nine characters, we were able to link 21,001 pairs (56.4%) in SGIC-only linking and 21,017 (56.3%) in SGIC + school linking. Using any combination of eight characters, the number of successfully matched pairs has increased by almost a quarter for both linking procedures; we could link an additional 8,756 pairs (23.5%) in SGIC-only linking and 8,777 (23.5%) in SGIC + school linking. For both SGIC-only and SGIC + school, the number of successfully linked pairs continued to increase with decreasing numbers of characters used for linking.

3.4 Quality measures

We compared the number of real pairs to SGIC-only pairs and found 35,467 identical and 2,610 different pairs. Of all 36,308 SGIC-only pairs, 2.3% were linked incorrectly (false-positive cases), and 4.7% were not identified (false-negative cases) (*Table 5*).

Table 4 | The proportion of successfully matched pairs by the count of characters used for linking

Count of characters used for linking	**SGIC-only linking**			**SGIC + school linking**		
	N	**%**	**Cumulative %**	**N**	**%**	**Cumulative %**
Matched on 9 characters	21 001	56.4	56.4	21 017	56.3	56.3
Matched on 8 characters	8 756	23.5	80.0	8 777	23.5	79.8
Matched on 7 characters	4 473	12.0	92.0	4 528	12.1	92.0
Matched on 6 characters	1 524	4.1	96.1	1 577	4.2	96.2
Matched on 5 characters	530	1.4	97.5	893	2.4	98.6
Matched on 4 characters	24	0.1	97.6	413	1.1	99.7
Totally matched	36 308	97.6		37 205	99.7	

Table 5 | The classification metrics evaluating the performance of the SGIC-only and SGIC + school linking procedures compared to real pairs by the count of the characters used for linking

Count of characters used for linking	**SGIC-only linking**						**SGIC + school linking**					
	True positive		**False positive**		**False negative**		**True positive**		**False positive**		**False negative**	
Matched on 9 characters	20 997	20 997	4	4	32	32	21 017	21 017			12	12
Matched on 8 characters	8 742	29 739	14	18	47	79	8 777	29 794			16	28
Matched on 7 characters	4 359	34 098	114	132	217	296	4 522	34 316	6	6	54	82
Matched on 6 characters	1 115	35 213	409	541	473	769	1 558	35 874	19	25	50	132
Matched on 5 characters	250	35 463	280	821	637	1 406	816	36 690	77	102	98	230
Matched on 4 characters	4	35 467	20	841	363	1 769	206	36 896	207	309	183	413
Totally matched	35 467		841		1 769		36 896		309		413	

Table 6 | The quality metrics evaluating the performance of the SGIC-only and SGIC + school linking procedures compared to real pairs by the count of the characters used for linking

Count of characters used for linking	SGIC-only linking						SGIC + school linking					
	Precision		Recall		F-measure		Precision		Recall		F-measure	
Matched on 9 characters	1.000	1.000	0.998	0.998	0.999	1.000	1.000	1.000	0.999	0.999	1.000	1.000
Matched on 8 characters	0.998	0.999	0.995	0.997	0.997	1.000	1.000	1.000	0.998	0.999	0.999	1.000
Matched on 7 characters	0.975	0.996	0.953	0.991	0.963	0.999	0.999	1.000	0.988	0.998	0.993	0.999
Matched on 6 characters	0.732	0.985	0.702	0.979	0.717	0.988	0.988	0.999	0.969	0.996	0.978	0.998
Matched on 5 characters	0.472	0.977	0.282	0.962	0.353	0.914	0.914	0.997	0.893	0.994	0.903	0.995
Matched on 4 characters	0.167	0.977	0.011	0.952	0.020	0.499	0.499	0.992	0.530	0.989	0.514	0.990
Totally matched												

Precision = TP/(TP + FP), Recall (sensitivity) = TP/(TP + FN), F-measure = 2 × ((P × R)/(P + R))

Similarly, we compared the number of real pairs with SGIC + school pairs and found 36,896 identical and 722 different pairs. Of all 37,205 SGIC + school pairs, 0.8% were false-positive and 1.1% false-negative.

At the level of 4 or more characters, we found 841 false-positive and 1,769 false-negative cases for SGIC-only linking and 309 false-positive and 413 false-negative cases for SGIC + school linking. The false positivity at all nine characters in the case of only SGIC links suggests that one of the children gradually appeared in two different schools.

If all nine characters are used, the precision equals 1.000 in both SGIC-only and SGIC + school linking, which means no false positivity was observed (*Table 6*). In the SGIC + school linking, the quality measures of precision, recall, and f-measure are all very high up to a level of five characters.

In the SGIC-only linking, we could observe a substantial decline in quality below the seven characters used for linking. Using only four characters resulted in very low values of quality measures (precision = 0.167, recall = 0.011, f-measure = 0.020) in SGIC-only linking.

Overall accuracy was 0.938 for SGIC-only linking and 0.983 for SGIC + school linking. The specificity was 0.826 for SGIC-only linking and 0.935 for SGIC + school linking.

4 DISCUSSION

In this study, we assessed the feasibility of linking anonymous longitudinal data using SGICs in schoolchildren participating in a three-armed randomized controlled prevention trial on substance use. Our findings suggest that a substantial proportion of survey data can be successfully linked using nine-character SGIC. However, some minor amendments to the SGIC sheet would be beneficial to achieve even higher efficiency.

In total, we were able to link 56.4% of children's SGICs using all nine characters and an additional 23.5% by applying the one-off rule (using any eight characters). By employing the school variable into the linking, we could link 56.3% of SGICs using all nine characters and an additional 23.5% using the one-off rule. Similar to our study, Galanti et al. (2007) reported a total match rate of 61% for pairing with all nine characters. Other longitudinal prevention studies using SGICs with different numbers of characters and elements reported a total match rate of between 60–71% (DiIorio et al., 2000; Grube, Morgan & Kearney, 1989; Kristjansson et al., 2014).

The efficiency of anonymous linking with SGIC was generally high. In SGIC-only linking, we were able to identify 2.3% false positive matches and 4.7% false-negative matches. Very high values of precision, recall, and F-measure (> 0.9) were observed up to the level of seven characters. However, between the sixth- and five-character-levels, the F-measure decreased substantially from 0.717 to 0.353. The results have considerably improved when employing the school variable in linking. In SGIC + school linking, we found 0.8% false-positive matches and 1.1% false-negative matches. The precision, recall, and F-measure did not fall below 0.9 up to the level of five characters used for pairing. A substantial decline in F-measure is noticeable until between the fifth- and four-character-levels (from 0.903 to 0.513). Therefore, to maintain sufficient pairing quality, the results indicate that linking could be performed up to a seven-character level in SGIC-only linking and five-character level when using a grouping variable (e.g., school affiliation).

The methodological properties of SGIC were examined by analysing errors and omissions in codes. The results showed that 8.5% of all children's SGICs contained one or more missing or erroneous characters. The proportion of SGICs with errors gradually decreased with each subsequent wave of data collection from 14.06% in first wave to 4.59% in the last seven wave. In line with Yurek, Vasey and Havens (2008), the highest error rates had those elements with less proximate relevance to the participant. Most errors (7.82%) were observed in eighth character derived from the child's parental grandmother's first name. It is possible that such a high error rate is because some children may not know or no longer have grandmothers or are confused about their real names. Similarly, this could explain a higher proportion of errors (0.73%) found in the child's father's first name. Previous methodological studies using SGICs based on similar elements reported the primary error rates ranged around 1.5% for characters derived from the date of birth, 8.9%

for characters derived from the participant's name, 16.7% for characters derived from father's name, and 18.8% for characters derived from the mother's name (Audette, Hammond & Rochester, 2020). In comparison to these studies, our results showed substantially lower error rates for those elements.

Overall error rate in pairs (discrepant characters) in pairs revealed that most mismatches (14%) occurred with the eighth character, which coded the paternal grandmother's first name. It was also difficult for many children to decide on their eye colour, as 12.6% of SGIC pairs had discrepant characters in this ninth position. Moreover, surprisingly high rates of discrepant characters were found for the first and second characters (12.4% and 11.3%, respectively), which coded the third and first letter of the child's first name. Further analysis showed that the discrepancy in these two positions was most likely caused by an interchange in the order of the letters. Therefore, to lower the error rates, we suggest amending the SGIC's sheet in such a way that no two characters derived by the same elements are placed next to each other.

Studies on differences between matched and unmatched subjects suggest that analysing data only from matched subjects who filled out both the survey questionnaire and the SGIC may be burdened with a systematic sampling error. Teichman, Rahab & Barnea (1993) found that unmatched students are significantly more likely to have lifetime and current experience with alcohol use than their matched counterparts and that they differ in socio-demographic background. Grube et al. (1989) and Kristjansson et al. (2014) provided evidence that unmatched adolescents are more likely to use tobacco, alcohol, and illicit substances than their matched participants. These findings potentially undermine the validity of longitudinal prevention research studies where unmatched subjects comprise a substantial part of the study population. The results of such studies should therefore be interpreted with some caution.

According to Agley et al. (2021) SGIC often include personal details (like birth month or middle initial), which, despite technically being anonymous, can lead to apprehensions among young participants and their parents. Such concerns might affect their willingness to participate, the quality of their responses, or even cause displeasure. A deeper exploration of the technical challenges faced in the SGIC process, especially when dealing with large datasets. The suggestion to explore the feasibility of computer-based SGICs indicates a move towards integrating more technology in data collection and linking processes. This could pave the way for more automated, efficient, and error-resistant methods, leveraging advancements in digital tools and software.

The study's findings could aid data linking methodologies in longitudinal research, particularly in the fields of adolescent health and education. By demonstrating the effective use of SGICs, it opens avenues for maintaining participant anonymity while ensuring accurate data tracking over extended periods. This approach could be especially beneficial in sensitive research areas where participant confidentiality is paramount, potentially increasing participation rates and the validity of responses.

Discussing the long-term reliability of SGICs, especially in multi-year studies, is crucial. The study provides evidence of the robustness of SGICs over a significant duration, which is essential for longitudinal research involving adolescents. However, there are concerns about the potential for errors over time, especially if participants' recall of the SGIC components diminishes. Future research could explore methods to enhance the stability of SGICs over longer periods and in diverse populations.

This study has strengths. First, the study successfully demonstrates the application of self-generated identification codes (SGICs) for linking anonymous longitudinal data of schoolchildren in a randomized controlled trial. This approach is crucial for maintaining participant anonymity, especially in sensitive research involving minors. Second, the study contributes to the field by exploring the methodological properties and efficiency of SGICs, addressing a gap in research regarding the use of SGICs in large-scale, longitudinal experimental prevention studies.

This study also has limitations. First, the research is based on a specific socio-demographic setting (Czech schoolchildren), which may limit the generalizability of the findings to other contexts or cultures. Second, overall, the use of SGIC and school linking acknowledges the limitations that might arise when using SGICs alone and leverages additional, non-sensitive information (like school affiliation) to bolster the integrity of the data linking process in a research study.

5 CONCLUSIONS

This study provided further evidence that SGICs is a suitable tool for linking anonymous longitudinal data of children participating in school-based prevention studies. Overall, the study represents a significant step forward in the field of longitudinal prevention research, offering valuable insights into both the potential and challenges of using SGICs for linking anonymous data in sensitive research contexts. It would be beneficial if future studies could verify our findings in different socio-demographic settings (e.g., children of different ages, from different cultural environments etc.). Future studies could also assess the feasibility of using computer-based SGICs for linking anonymous data because it may be one of the ways to eliminate errors from transcribing paper-based SGICs into electronic form.

Authors' contributions: RG and JV designed the study and wrote the protocol. JV performed the initial drafting of the manuscript. JV conducted the statistical analyses. JV managed the literature searches and summaries of related work and participated in the data interpretation and manuscript preparation. RG and JV finalized the manuscript. All the authors contributed to and have approved the final manuscript.

Declaration of interest: The authors declare that they have no conflict of interest.

REFERENCES

Agley, J., Tidd, D., Jun, M., Eldridge, L., Xiao, Y., Sussman, S., Jayawardene, W., Agley, D., Gassman, R., & Dickinson, S. L. (2021). Developing and validating a novel anonymous method for matching longitudinal school-based data. *Educational and Psychological Measurement, 81*(1), 90–109. https://doi.org/10.1177/0013164420938457

Audette, L. M., Hammond, M. S., & Rochester, N. K. (2020). Methodological issues with coding participants in anonymous psychological longitudinal studies. *Educational and Psychological Measurement, 80*(1), 163–185. https://doi.org/10.1177/0013164419843576

Bezençon, V., De Santo, A., Holzer, A., & Lanz, B. (2023). Escape Addict: A digital escape room for the prevention of addictions and risky behaviors in schools. *Computers & Education, 200*, Article 104798. https://doi.org/10.1016/j.compedu.2023.104798

Calatrava, M., de Irala, J., Osorio, A., Benítez, E., & Lopez-del Burgo, C. (2022). Matched and fully private? A new self-generated identification code for school-based cohort studies to increase perceived anonymity. *Educational and Psychological Measurement, 82*(3), 465–481. https://doi.org/10.1177/00131644211035436

Carifio, J. (1994). Sensitive data and students' tendencies to give socially desirable responses. *Journal of Alcohol and Drug Education, 39*(2), 74–84. https://www.jstor.org/stable/45092024

Carifio, J., & Biron, R. (1978). Collecting sensitive data anonymously: The CDRGP technique. *Journal of Alcohol and Drug Education, 23*(2), 47–66. https://www.jstor.org/stable/45091343

Catalano, R. F., Fagan, A. A., Gavin, L. E., Greenberg, M. T., Irwin, C. E., Ross, D. A., & Shek, D. T. (2012). Worldwide application of prevention science in adolescent health. *The Lancet, 379*(9826), 1653–1664. https://doi.org/10.1016/S0140-6736(12)60238-4

Damrosch, S. P. (1986). Ensuring anonymity by use of subject-generated identification codes. *Research in Nursing & Health, 9*(1), 61–63. https://doi.org/10.1002/nur.4770090110

De Medeiros, P. F. P., Valente, J. Y., Rezende, L. F. M., & Sanchez, Z. M. (2024). Patterns of alcohol access among Brazilian adolescents: A latent class analysis. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-024-01389-8

DiIorio, C., Soet, J. E., VanMarter, D., Woodring, T. M., & Dudley, W. N. (2000). Focus on research methods – An evaluation of a self-generated identification code. *Research in Nursing & Health, 23*(2), 167–174. https://doi.org/10.1002/(sici)1098-240x(200004)23:2<167::aid-nur9>3.0.co;2-k

Fernandez-Hermida, J. R., Calafat, A., Becona, E., Secades-Villa, R., Juan, M., & Sumnall, H. (2013). Cross-national study on factors that influence parents' knowledge about their children's alcohol use. *Journal of Drug Education, 43*(2), 155–172. https://doi.org/10.2190/DE.43.2.d

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), 402–408.

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. https://doi.org/10.1016/j.ypmed.2006.07.019

Gfroerer, J., & Kennet, J. (2014). Collecting survey data on sensitive topics: Substance use. In T. P. Johnson (Ed.), *Health Survey Methods* (pp. 447–472). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118594629.ch17

Grube, J. W., Morgan, M., & Kearney, K. A. (1989). Using self-generated identification codes to match questionnaires in panel studies of adolescent substance use. *Addictive Behaviors, 14*(2), 159–171. https://doi.org/10.1016/0306-4603(89)90044-0

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). Springer Berlin Heidelberg.

Jessop, D. J. (1982). Topic variation in levels of agreement between parents and adolescents. *Public Opinion Quarterly, 46*(4), 538–559. https://doi.org/10.1086/268751

Jurczyk, P., Lu, J. J., Xiong, L., Cragan, J. D., & Correa, A. (2008). Fine-grained record integration and linkage tool. *Birth Defects Res A Clin Mol Teratol, 82*(11), 822–829. https://doi.org/10.1002/bdra.20521

Kandel, D. (1973). Adolescent marihuana use – Role of parents and peers. *Science, 181*(4104), 1067–1070. https://doi.org/10.1126/science.181.4104.1067

Kristjansson, A. L., Sigfusdottir, I. D., Sigfusson, J., & Allegrante, J. P. (2014). Self-generated identification codes in longitudinal prevention research with adolescents: A pilot study of matched and unmatched subjects. *Prevention Science, 15*(2), 205–212. https://doi.org/10.1007/s11121-013-0372-z

Lee, B. C., Westaby, J. D., &Berg, R. L. (2004). Impact of a national rural youth health and safety initiative: Results from a randomized controlled trial. *American Journal of Public Health, 94*, 1743–1749. https://doi.org/10.2105/Ajph.94.10.1743

Lippe, M., Johnson, B., & Carter, P. (2019). Protecting student anonymity in research using a subject-generated identification code. *Journal of Professional Nursing, 35*(2), 120–123. https://doi.org/10.1016/j.profnurs.2018.09.006

Little, M. A., Bonilla, G., McMurry, T., Pebley, K., Klesges, R. C., & Talcott, G. W. (2021). The feasibility of using self-generated identification codes in longitudinal research with military personnel. *Evaluation & the Health Professions*. https://doi.org/10.1177/01632787211031625

McAlister, A., & Gordon, N. P. (1986). Attrition bias in a cohort study of substance-abuse onset and prevention. *Evaluation Review, 10*(6), 853–859. https://doi.org/10.1177/0193841x8601000609

Sandnes, F. E. (2021a). BRIDGE: Administering small anonymous longitudinal HCI studies with snowball-type sampling. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (pp. 287–297). Springer International Publishing. https://doi.org/10.1007/978-3-030-85610-6_17

Sandnes, F. E. (2021b). CANDIDATE: A tool for generating anonymous participant-linking IDs in multi-session studies. *Plos One, 16*(12), Article e0260569. https://doi.org/10.1371/journal.pone.0260569

Sandnes, F. E. (2021c). HIDE: *Short IDs for robust and anonymous linking of users across multiple sessions in small HCI experiments*. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–6. https://doi.org/10.1145/3411763.3451794

Seidel, A.-K., Morgenstern, M., Galimov, A., Pedersen, A., Isensee, B., Goecke, M., & Hanewinkel, R. (2022). Use of electronic cigarettes as a predictor of cannabis experimentation: A longitudinal study among German youth. *Nicotine & Tobacco Research, 24*(3), 366–371. https://doi.org/10.1093/ntr/ntab166

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of self-generated identification codes. *Evaluation Review, 34*(5), 391–418. https://doi.org/10.1177/0193841X10387576

Schumacher, S. (2007, January 18). *Probabilistic versus deterministic data matching: Making an accurate decision*. Information Management Special Reports.

Teichman, M., Rahav, G., & Barnea, Z. (1993). Alcohol consumption of matched and unmatched adolescents in a longitudinal study. *Journal of Child & Adolescent Substance Abuse, 2*(3–4), 67–86. https://doi.org/10.1300/J272v02n03_09

Vacek, J., Vonkova, H., & Gabrhelik, R. (2017). A successful strategy for linking anonymous data from students' and parents' questionnaires using self-generated identification codes. *Prevention Science, 18*(4), 450–458. https://doi.org/10.1007/s11121-017-0772-6

Wilson, A. L. G., Hoge, C. W., McGurk, D., Thomas, J. L., Clark, J. C., & Castro, C. A. (2010). Application of a new method for linking anonymous survey data in a population of soldiers returning from Iraq. *Annals of Epidemiology, 20*(12), 931–938. https://doi.org/10.1016/j.annepidem.2010.08.008

Yurek, L. A., Vasey, J., & Havens, D. S. (2008). The use of self-generated identification codes in longitudinal research. *Evaluation Review, 32*(5), 435–452. https://doi.org/10.1177/0193841X08316676