Global School-Based Student Health Survey (GSHS)

Data Processing and Weighting

World Health Organization

Overview

- GSHS cleaning, weighting and analysis* have been standardized to ensure the integrity of the global system and comparability of results.
- WHO has developed a set of R code to automate these tasks.
- These slides present how GSHS data is cleaned and weighted and the input files needed to use the R code.

World Health Organization

* Standard descriptive analysis – this is by no means an exhaustive analysis. Countries are encouraged to perform additional analyses.

Data Cleaning

World Health Organization

Overview

- As part of the standard cleaning process, GSHS data are checked for quality and consistency.
- Checking occurs at both the variable level as well as at the record level.
  - Thus, it is possible for a **variable** to be dropped or a student's **entire response** to be dropped.
- The following slides in this section describe the data edits performed.

Out-of-range edits

- If a student selects a response that does not correspond to one of the possible responses for a question, then the response is set to missing.

- *Example:* If “A” and “B” are the valid response options for a question and a student selects “C”, “D”, “E”, “F”, “G”, or “H,” then the response is set to missing for that student.

World Health Organization

Multi-response edits

- If a student selects more than one response for a question, then the question is set to missing.

- GSHS questions never allow for multiple responses.

World Health Organization

BMI-related edits

- BMI is calculated using the height and weight measures reported in each student's response.
  - If either height or weight are missing, BMI is set to missing.

- Height, weight and BMI are checked to see if they are outside the biologically plausible range*.
  - If height, weight or BMI are implausible, all are set to missing.
  - If age or sex is missing, height, weight and BMI are set to missing since plausible ranges vary by age and sex and it would be impossible to determine plausibility.

World Health Organization

* With a z-score <-5 or >5 according to WHO Growth Standards.

Logical consistency edits

- Logical consistency checks are made for questions in 6 of the core modules.
- These checks ensure that responses are *internally consistent*
  - Example: If a student responds that they did not clean their teeth in the past 30 days but in the following question responds that the toothpaste usually used to brush their teeth in the past 30 days contains fluoride – these responses would not be internally consistent.
- If a check fails, then the responses to **both** questions are set to missing *except* if one of the questions is AGE (age is never set to missing).
- Consistency checks are not exhaustive and there are no consistency checks done for core-expanded or country-specific questions.
- All 46 edits are listed on the following slide.

World Health Organization

Hygiene

1. HY_CLTEETH = A AND HY_FLUORIDE = B,C,D

Injury

2. IN_TIMESINJ = A AND IN_TYPEINJ = B, C, D, E, F, G, H
3. IN_TIMESINJ = A AND IN_CAUSEINJ = B, C, D, E, F, G, H

Tobacco Use

4. TO_TRIEDCIG = B AND TO_AGECIG = B, C, D, E, F, G, H
5. TO_TRIEDCIG = B AND TO_DAYSCIG = B, C, D, E, F, G
6. DE_AGE = A AND TO_AGECIG = E,F,G,H
7. DE_AGE = B AND TO_AGECIG = F,G,H
8. DE_AGE = C AND TO_AGECIG = F,G,H
9. DE_AGE = D AND TO_AGECIG = G,H
10. DE_AGE = E AND TO_AGECIG = G,H
11. DE_AGE = F AND TO_AGECIG = H
12. DE_AGE = G AND TO_AGECIG = H

Alcohol Use

13. AL_AGE = A AND AL_DAYS = B, C, D, E, F, G
14. AL_AGE = A AND AL_DRINKS = B, C, D, E, F, G
15. AL_AGE = A AND AL_INAROW = B, C, D, E, F, G, H
16. AL_AGE = A AND AL_SOURCE = B, C, D, E, F
17. AL_AGE = A AND AL_TROUBLE = B, C, D, E, F
18. AL_AGE =A AND AL_DRUNK = B, C, D, E, F
19. DE_AGE = A AND AL_AGE = E,F,G,H
20. DE_AGE = B AND AL_AGE = F,G,H
21. DE_AGE = C AND AL_AGE = F,G,H
22. DE_AGE = D AND AL_AGE = G,H
23. DE_AGE = E AND AL_AGE = G,H
24. DE_AGE = F AND AL_AGE = H
25. DE_AGE = G AND AL_AGE = H

Drug Use

26. DR_AGE = A AND DR_CANLIFE = B, C, D, E, F
27. DR_AGE = A AND DR_CAN30 = B, C, D, E, F
28. DR_AGE = A AND DR_AMPHLIFE = B, C, D, E, F
29. DE_AGE = A AND DR_AGE = E,F,G,H
30. DE_AGE = B AND DR_AGE = F,G,H
31. DE_AGE = C AND DR_AGE = F,G,H
32. DE_AGE = D AND DR_AGE = G,H
33. DE_AGE = E AND DR_AGE = G,H
34. DE_AGE = F AND DR_AGE = H
35. DE_AGE = G AND DR_AGE = H

Sexual Behaviors

36. DE_AGE = A AND SX_AGE = E,F,G,H
37. DE_AGE = B AND SX_AGE = F,G,H
38. DE_AGE = C AND SX_AGE = F,G,H
39. DE_AGE = D AND SX_AGE = G,H
40. DE_AGE = E AND SX_AGE = G,H
41. DE_AGE = F AND SX_AGE = H
42. DE_AGE = G AND SX_AGE = H
43. SX_EVERSEX = B AND SX_AGE = B, C, D, E, F, G, H
44. SX_EVERSEX = B AND SX_NUMBER = B,C,D,E,F,G
45. SX_EVERSEX = B AND SX_CONDOM = B,C
46. SX_EVERSEX = B AND SX_BC = B,C,D,E,F,G,H

World Health Organization

Variable-level edits

- After all other checks have been implemented, each variable is checked to ensure at least 60% of students have responded.
- If the response rate for a variable is less than 60%, the variable is set to missing for all students.

World Health Organization

Record-level edits

- After all other checks have been implemented, each record is checked to ensure there are at least 20 valid responses
- If a student's response has fewer than 20 valid responses, the response is deleted.
- If a record has 15 or more identical responses in a row, other than "A", the entire record is deleted.

World Health Organization

Weighting

Overview

- Once data have been cleaned, the weighting process can begin.
- Weighting accounts for:
  - the probability of selection of schools and classes
  - non-responding schools, classes and students
  - the distribution of the target population (i.e. students in the targeted grades) by grade and sex
- In addition to analysis weights, PSU and Stratum will also be generated which inform the statistical software about the design of your sample.

Requirements

- All of the following conditions must be met in order to weight GSHS data:
  - the sample was scientifically selected from an up-to-date and complete sampling frame
  - all school-level and class-level forms were accurately completed
  - a high (>60%) overall response rate was obtained

World Health Organization

Weight calculation

- The formula used to calculate analysis weights for most GSHS data sets is

weight = w1 * w2 * f1 * f2 * f3

where:

Base weight
- **w1** = the inverse probability of selecting each school
- **w2** = the inverse probability of selecting each class

Non-response adjustment
- **f1** = a school-level non-response adjustment factor
- **f2** =a student-level non-response adjustment factor (calculated per class)

Post-stratification adjustment
- **f3** = a post-stratification adjustment factor (calculated by sex within each grade)

World Health Organization

PSU and Stratum

- PSU and Stratum describe the complex sample design of the survey
- These numbers are generated as follows:
  - **Schools selected with certainty*** : assign a unique stratum to each school and a unique PSU to each class in each school
  - **All other schools :** sort schools by school weight** and group schools into pairs (if there is an odd number, make one group of three), assign a unique stratum to each pair of schools (or group of three) and a unique PSU to all classes within a given school

World Health Organization

* Large schools with a school weight of 1.0.
** w1 in the weight calculation

PSU and Stratum - example

School Weight	School	Classes	Stratum	PSU
1.0	A	1	1	1
		3	1	2
		6	1	3
1.0	B	2	2	4
		4	2	5
		6	2	6
1.0	C	1	3	7
		3	3	8
		4	3	9
		6	3	10
1.27	E	1	4	11
		2	4	11
		3	4	11
1.38	F	1	4	12
		3	4	12
		5	4	12
1.79	G	2	5	13
		4	5	13
		6	5	13
		8	5	13
1.83	H	1	5	14
		3	5	14
		5	5	14
1.90	I	3	5	15
		6	5	15
		9	5	15

Schools selected with certainty

All other schools (i.e. smaller schools)

World Health Organization