PharmaSUG China 2021 - Paper AD-003

Data De-identification of Demographic Detail in Adult and Pediatric Study

Huan Lu, Sanofi

ABSTRACT

Clinical trial data sharing is an increasingly important topic. Making sure that, before sharing, data has been de-identified in a way that protects the identity of subjects and retaining utility for researchers have never been this relevant as today.

One of the challenges is those demographic details collected from clinical trials data. The objective of this paper is to draw a process map to de-identify potential pitfalls of the demographic details, suggest best practices, and offer automation ideas specifically on height, weight, and body mass index (Quetelet index) in adults, pediatric and hybrid studies.

INTRODUCTION

Participant demographic information and physical characteristics, such as age, sex, race, height, weight, etc., are defined as quasi-identifiers, sensitive individual information, or low-frequency event until acceptable risk threshold is achieved, meaning that these details above would be extremely straight-forward to identify someone if they remain raw describing the profile of a participant during a clinical trial.

With the validation of risk assessment, demographic details could be generalized making that it is hard to identify a specific participant. A clinical trial in a specific therapeutic area may allow these details to distribute widely among all participants, such as weight in diabetes study, generalization on demographic details hence become more necessary. Although parts of trial information were published on regulatory authorities' websites long before data sharing, demographic details still need to be reviewed or redacted.

Aggregation and combination are considered good practices when it comes to de-identification. For age, race, etc., aggregation applied reasonably brings sufficient generalization among all participants; for height and weight, combining them into body mass index would be an ideal approach to blur someone's demographic characteristics, which would be the main topic of this paper.

OBJECTIVE

A common case of various clinical trials would be conducted in a population of adults only, where demographic details, such as height and weight, would not vary dramatically nor frequently, especially height. However, in some other specific therapeutic areas, for example, diabetes, weight is the main character collected during every single visit and considered as an important measurement that gives weight a comparatively wide distribution and deviation.

WITH BMI DERIVED

Since height would not change much on an adult, it usually would only be collected once at first visit either by measured or manually filled, this could be combined with any weight at visit to populate body mass index with the very classical equation by Adolphe Quetelet in 1972:

$$Body\ Mass\ Index\ (BMI) = \frac{mass_{(kg)}}{height^2_{(m)}}$$

Equation 1: Body Mass Index (Quetelet Index) Equation

A common use of the BMI is to assess how far a participant’s body weight departs from what is normal or desirable for a person's height, which could also be used to roughly describe someone’s demographic detail without specifically telling someone if tall, short, overweight or underweight.

After combining height and weight into BMI, aggregation on BMI provides a rough demographic fact of a participant without leaking too detailed characteristics and preventing low-frequency events occur.

The WHO refers BMI to nutritional status defining a perfect reference to aggregate BMI:

BMI	Nutritional status
18.5–24.9	Normal weight
25.0–29.9	Pre-obesity
30.0–34.9	Obesity class I
35.0–39.9	Obesity class II
Above 40	Obesity class III

Table 1: The WHO Nutritional Status

By populating the aggregated BMI with nutritional status, solely remaining either height or weight in data would be easy to derive one and the other with the equation, hence simply removing height and weight records from clinical trial data of a study becomes extremely necessary. Only remain aggregated BMI would not undermine the data utility when it comes to data sharing. On one hand, less detailed demographic characteristics could be exposed and shared safely, on the other hand, a participant’s private information, such as height in centimeter and weight in kilogram, were sufficiently protected.

Risk assessment should be considered at this point in case of a certain nutritional status still standing out alone out of all the others that could easily be identified as a low-frequency event, which can be quite concerning or potentially socially or financially harmful to an individual when revealed publicly.

However, on a certain study level, risk assessment may not be always flavored with the aggregation defined by the WHO, altered nutritional status aggregation hence become considerable:

BMI	Nutritional status
Below 29.9	Non-obesity
Above 30.0	Obesity

Table 2: Example of Altered Nutritional Status

Study specified aggregation should be included in supportive documents supporting the authenticity of the data and helping reviewers to better understand the fuzzy demographic details.

WITHOUT BMI DERIVED

Height and weight as quasi-identifiers are extremely sensitive, meaning that they are most intuitive when having someone in mind. When BMI was not collected, aggregation on height and weight hence become an alternative mean.

One of the approaches is to use fixed intervals replacing height and weight if the value falls into that, it is easy to implement and adjust the intervals if they need to be wilder or narrower, however, using fixed intervals regardless of the population distribution cannot guarantee that there would always be no low-frequency event as aggregated height and weight, it is determined by this method's nature. For example, using 10 cm as a step to aggregate height:

Height (cm)	Aggregated Height(cm)
172	(170, 180]
178	(170, 180]
178	(170, 180]
162	(160, 170]
174	(170, 180]
178	(170, 180]
178	(170, 180]
......	......

→ Low-frequency Event

Table 3. Example of Height Aggregation Using Fixed Intervals with Low-frequency Event

One of the other approaches is more recommended, which is to use percentile. From wild to narrow, using median, quantile or percentile could guarantee that size of each interval be comparatively similar because the thresholds were calculated based on the population distribution. In this way, the difference could be significant between the intervals and insignificant within the interval. Even though low-frequency events cannot be prevented, with careful review and moderate further adjustment, it is still a practically good way to aggregate height and weight when BMI is absent. For example, using 174 cm as a median to aggregate height:

Height (cm)	Aggregated Height(cm)
172	(160, 174]
178	(174, 188]
178	(174, 188]
162	(160, 174]
174	(160, 174]
178	(174, 188]
178	(174, 188]
......	......

Table 4. Example of Height Aggregation Using Quantile

Converting height and weight into BMI could be another way to deal with the problem since BMI was not collected during a clinical trial. Once BMI was derived and height and weight were removed, the problem would become the same as above.

CHALLENGE

In terms of referring BMI to nutritional status, there are some other conditions required if the referral works appropriately, one of the major conditions is age. The fact that the WHO assumes the nutritional status is only applied to adults over 20 years old then BMI would fall into one of the categories above, which becomes one of the challenges when a study involves children or adolescents. With further thoughts, BMI is also recommended for use in children and adolescents.

IN PEDIATRIC STUDIES

In children, BMI is calculated for adults first, then compare with z-scores or percentiles. During childhood and adolescence, the ratio between weight and height varies with sex and age, hence the thresholds determining the nutritional status of those aged 5–19 years shall be gender-and-age-specific.

According to the WHO child growth standards, 2007 BMI-for-age reference for children aged 5-19 years provides two distributional categories to de-identify BMI for age 5-19 years: z-scores and percentiles. The z-scores method used to construct the 2007 WHO references relied on GAMLSS with the Box-Cox power exponential distribution (Rigby and Stasinopoulos, 2004) and selected models simplified to the LMS model (Cole and Green, 1992) since none of the references required adjustment for kurtosis, which hence makes the z-scores method more desirable. For example, the thresholds of the 2007 BMI-for-age reference for children aged 5-19 years, overweight is defined as a BMI-for-age value over +1 SD and obesity as a BMI-for-age value over +2 SD.

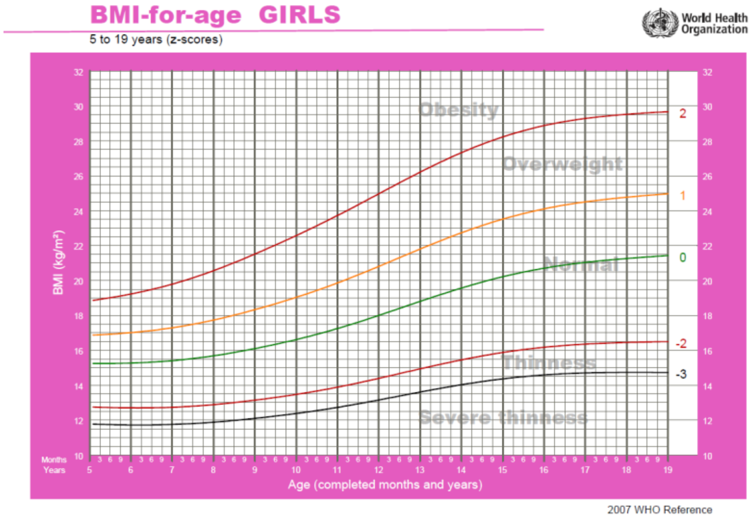


Figure 1. BMI-for-age 5 to 19 years Girls vs. Nutritional Status (z-scores chart)

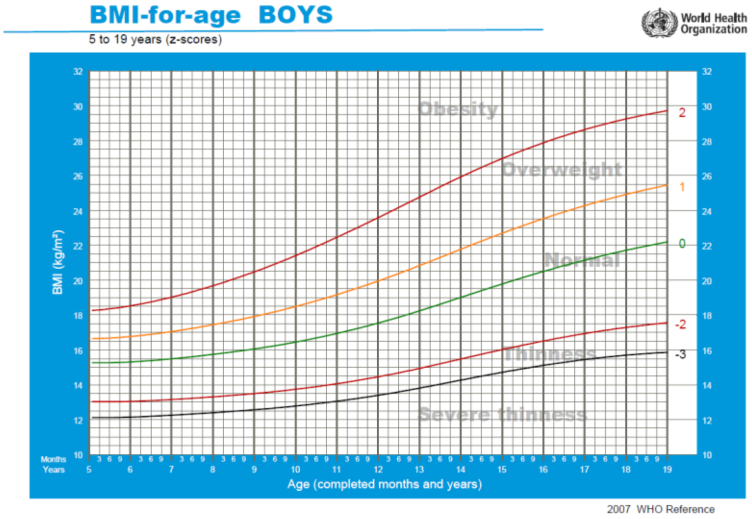
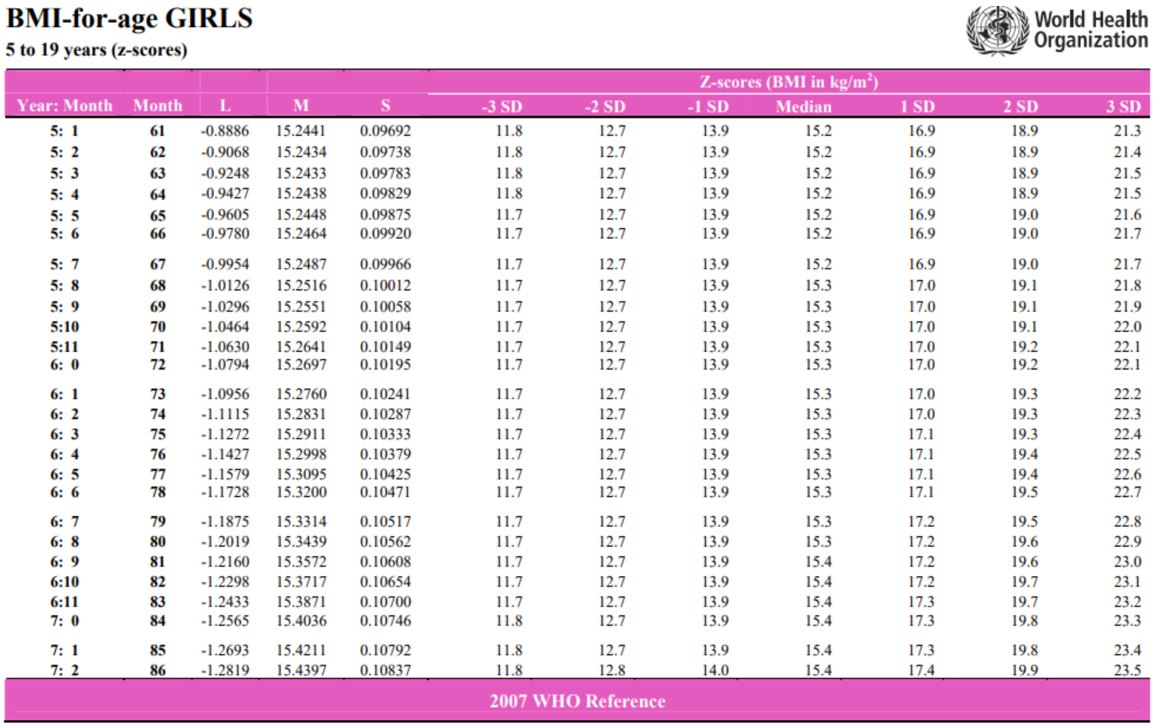


Figure 2. BMI-for-age 5 to 19 years Boys vs. Nutritional Status (z-scores chart)

On the programming level, the WHO provides z-scores tables at each threshold to accurately aggregate BMI by sex and by age in the month, which is beneficial to programming automation.



BMI-for-age GIRLS

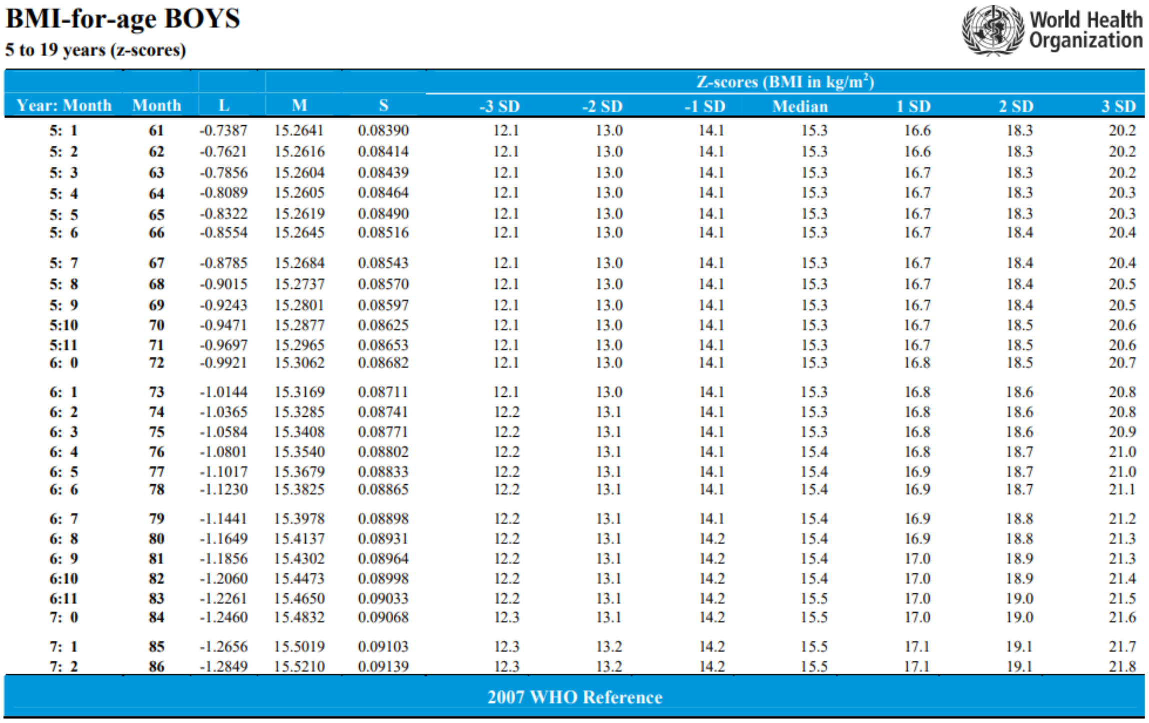
5 to 19 years (z-scores)

World Health Organization

					Z-scores (BMI in kg/m^2)						
Year: Month	Month	L	M	S	-3 SD	-2 SD	-1 SD	Median	1 SD	2 SD	3 SD
5: 1	61	-0.8886	15.2441	0.09692	11.8	12.7	13.9	15.2	16.9	18.9	21.3
5: 2	62	-0.9068	15.2434	0.09738	11.8	12.7	13.9	15.2	16.9	18.9	21.4
5: 3	63	-0.9248	15.2433	0.09783	11.8	12.7	13.9	15.2	16.9	18.9	21.5
5: 4	64	-0.9427	15.2438	0.09829	11.8	12.7	13.9	15.2	16.9	18.9	21.5
5: 5	65	-0.9605	15.2448	0.09875	11.7	12.7	13.9	15.2	16.9	19.0	21.6
5: 6	66	-0.9780	15.2464	0.09920	11.7	12.7	13.9	15.2	16.9	19.0	21.7
5: 7	67	-0.9954	15.2487	0.09966	11.7	12.7	13.9	15.2	16.9	19.0	21.7
5: 8	68	-1.0126	15.2516	0.10012	11.7	12.7	13.9	15.3	17.0	19.1	21.8
5: 9	69	-1.0296	15.2551	0.10058	11.7	12.7	13.9	15.3	17.0	19.1	21.9
5:10	70	-1.0464	15.2592	0.10104	11.7	12.7	13.9	15.3	17.0	19.1	22.0
5:11	71	-1.0630	15.2641	0.10149	11.7	12.7	13.9	15.3	17.0	19.2	22.1
6: 0	72	-1.0794	15.2697	0.10195	11.7	12.7	13.9	15.3	17.0	19.2	22.1
6: 1	73	-1.0956	15.2760	0.10241	11.7	12.7	13.9	15.3	17.0	19.3	22.2
6: 2	74	-1.1115	15.2831	0.10287	11.7	12.7	13.9	15.3	17.0	19.3	22.3
6: 3	75	-1.1272	15.2911	0.10333	11.7	12.7	13.9	15.3	17.1	19.3	22.4
6: 4	76	-1.1427	15.2998	0.10379	11.7	12.7	13.9	15.3	17.1	19.4	22.5
6: 5	77	-1.1579	15.3095	0.10425	11.7	12.7	13.9	15.3	17.1	19.4	22.6
6: 6	78	-1.1728	15.3200	0.10471	11.7	12.7	13.9	15.3	17.1	19.5	22.7
6: 7	79	-1.1875	15.3314	0.10517	11.7	12.7	13.9	15.3	17.2	19.5	22.8
6: 8	80	-1.2019	15.3439	0.10562	11.7	12.7	13.9	15.3	17.2	19.6	22.9
6: 9	81	-1.2160	15.3572	0.10608	11.7	12.7	13.9	15.4	17.2	19.6	23.0
6:10	82	-1.2298	15.3717	0.10654	11.7	12.7	13.9	15.4	17.2	19.7	23.1
6:11	83	-1.2433	15.3871	0.10700	11.7	12.7	13.9	15.4	17.3	19.7	23.2
7: 0	84	-1.2565	15.4036	0.10746	11.8	12.7	13.9	15.4	17.3	19.8	23.3
7: 1	85	-1.2693	15.4211	0.10792	11.8	12.7	13.9	15.4	17.3	19.8	23.4
7: 2	86	-1.2819	15.4397	0.10837	11.8	12.8	14.0	15.4	17.4	19.9	23.5

2007 WHO Reference

Figure 3. BMI-for-age 5 to 19 years Girls (z-scores table)



BMI-for-age BOYS

5 to 19 years (z-scores)

World Health Organization

					Z-scores (BMI in kg/m^2)						
Year: Month	Month	L	M	S	-3 SD	-2 SD	-1 SD	Median	1 SD	2 SD	3 SD
5: 1	61	-0.7387	15.2641	0.08390	12.1	13.0	14.1	15.3	16.6	18.3	20.2
5: 2	62	-0.7621	15.2616	0.08414	12.1	13.0	14.1	15.3	16.6	18.3	20.2
5: 3	63	-0.7856	15.2604	0.08439	12.1	13.0	14.1	15.3	16.7	18.3	20.2
5: 4	64	-0.8089	15.2605	0.08464	12.1	13.0	14.1	15.3	16.7	18.3	20.3
5: 5	65	-0.8322	15.2619	0.08490	12.1	13.0	14.1	15.3	16.7	18.3	20.3
5: 6	66	-0.8554	15.2645	0.08516	12.1	13.0	14.1	15.3	16.7	18.4	20.4
5: 7	67	-0.8785	15.2684	0.08543	12.1	13.0	14.1	15.3	16.7	18.4	20.4
5: 8	68	-0.9015	15.2737	0.08570	12.1	13.0	14.1	15.3	16.7	18.4	20.5
5: 9	69	-0.9243	15.2801	0.08597	12.1	13.0	14.1	15.3	16.7	18.4	20.5
5:10	70	-0.9471	15.2877	0.08625	12.1	13.0	14.1	15.3	16.7	18.5	20.6
5:11	71	-0.9697	15.2965	0.08653	12.1	13.0	14.1	15.3	16.7	18.5	20.6
6: 0	72	-0.9921	15.3062	0.08682	12.1	13.0	14.1	15.3	16.8	18.5	20.7
6: 1	73	-1.0144	15.3169	0.08711	12.1	13.0	14.1	15.3	16.8	18.6	20.8
6: 2	74	-1.0365	15.3285	0.08741	12.2	13.1	14.1	15.3	16.8	18.6	20.8
6: 3	75	-1.0584	15.3408	0.08771	12.2	13.1	14.1	15.3	16.8	18.6	20.9
6: 4	76	-1.0801	15.3540	0.08802	12.2	13.1	14.1	15.4	16.8	18.7	21.0
6: 5	77	-1.1017	15.3679	0.08833	12.2	13.1	14.1	15.4	16.9	18.7	21.0
6: 6	78	-1.1230	15.3825	0.08865	12.2	13.1	14.1	15.4	16.9	18.7	21.1
6: 7	79	-1.1441	15.3978	0.08898	12.2	13.1	14.1	15.4	16.9	18.8	21.2
6: 8	80	-1.1649	15.4137	0.08931	12.2	13.1	14.2	15.4	16.9	18.8	21.3
6: 9	81	-1.1856	15.4302	0.08964	12.2	13.1	14.2	15.4	17.0	18.9	21.3
6:10	82	-1.2060	15.4473	0.08998	12.2	13.1	14.2	15.4	17.0	18.9	21.4
6:11	83	-1.2261	15.4650	0.09033	12.2	13.1	14.2	15.5	17.0	19.0	21.5
7: 0	84	-1.2460	15.4832	0.09068	12.3	13.1	14.2	15.5	17.0	19.0	21.6
7: 1	85	-1.2656	15.5019	0.09103	12.3	13.2	14.2	15.5	17.1	19.1	21.7
7: 2	86	-1.2849	15.5210	0.09139	12.3	13.2	14.2	15.5	17.1	19.1	21.8

2007 WHO Reference

Figure 4. BMI-for-age 5 to 19 years Boys (z-scores table)

IN HYBRID STUDIES

The most complicated scenario lies in a study of the population regardless of age, if the minimum age is over 5 years old, the practical approach is to define study specified nutritional status aggregation combining the references above organically. With reasonable interpretation in supportive documents and validation with risk assessment, study specified aggregation hence is acceptable. Here is an example:

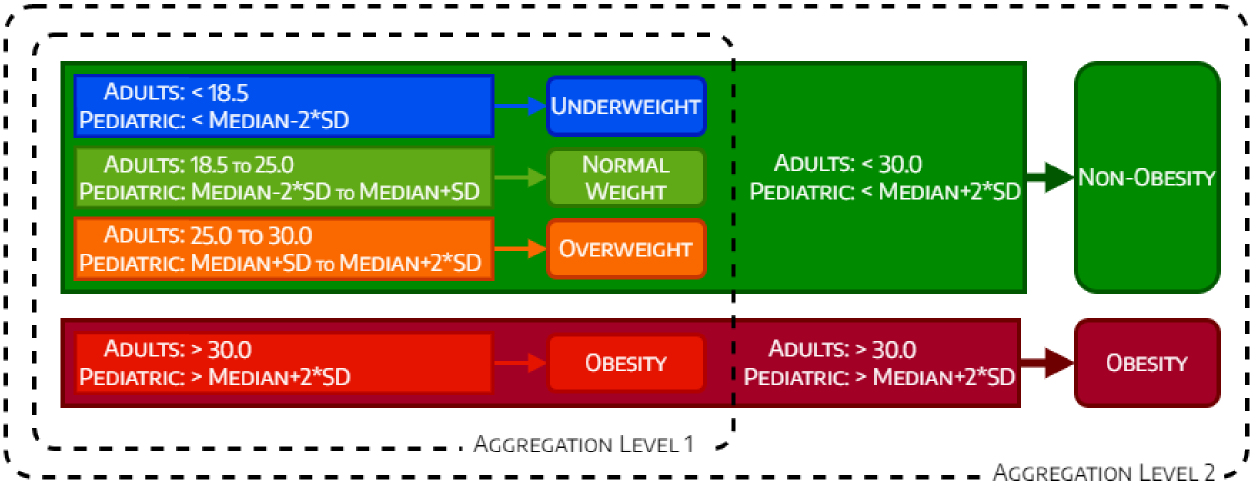


Figure 5: Example of BMI aggregation to Nutritional Status in Hybrid Study

Furthermore, when making a study a specified aggregation, it should always be noted that case-by-case risk assessment shall be required by the data provider to determine the appropriateness of disclosing study information or particular data in special circumstances such as demographic details, rare diseases, tiny population, single-center trials, and low-frequency events.

When it comes to the balance between data utility and risk assessment, there could be more than one aggregation on the study level, always starting from the more dense aggregation could retain as much information as possible, as long as the de-identified data could pass the risk assessment, there is no reason to over-aggregate the low-frequency events undermining data utility.

BMI INTERNATIONAL VARIATION

The WHO references are not always valid and keep evolving from time to time, country to country, making BMI aggregation somehow problematic. During clinical trials, participants from different sites may have a difference in BMI, percentage of body fat, and other health risks. Take diabetes as an example, there is a higher risk of type 2 diabetes mellitus and atherosclerotic cardiovascular disease at BMIs lower than the WHO reference for overweight, although the thresholds for observed risk varies among different populations observed in Europe, Asia, and Africa.

A practical approach is to remove BMI and aggregate height and weight by either median, quantiles, or percentiles since the date could deviate dramatically from the WHO reference.

OTHER LOW-FREQUENCY EVENT

Other low-frequency events in demographic details, such as a scar on a specific body location, on one hand, would be extremely difficult to notice when it comes to programming automation without looking into every single data. On the other hand, once such a low-frequency event was detected, removing or redacted would be the most practical way.

The general solutions to this topic remain unwritten as today, artificial intelligence technics with supervised training may throw light on it in the foreseeable future. A practical approach is to review the de-identified data carefully populating the frequency table of those open text fields by manually input.

CONCLUSION

Demographic details as height, weight, and BMI are always valuable to clinical trials and sensitive to individual privacy since most of them were directly collected during a clinical trial or derived from raw data. For the sake of data utility, sensitive data could be kept If the data is not personal anymore. Each sponsor must decide how they want to approach sensitive data. In such as dilemma, the data provider need to balance the extent of measures taken to address required individual privacy while maintaining data utility, since data such as demographic details may be the key information desired by the researcher to perform their clinical analyses.

Common de-identification technics introduced above for both sensitive and low-frequency events are more like adding noise by aggregating data, which allows the data provider to fulfill the agreement to share the data with any third party and prohibit the researcher from attempting to re-identify a certain individual from de-identification data without telling the whole truth. By applying a common de-identification approach across data providers, the utility of the de-identified data would be increased such that multiple data providers can be easier to facilitate meta-analyses.

As technology continuously advances, more and more de-identified data would become accessible, the methodology will require regular review among the pharmaceutical giants to ensure that the balance between data utility and required individual privacy is appropriately balanced. Methods for quantifying the risk of participant re-identification may need to be employed more frequently, particularly as external data sources grow and data linkage techniques improve such as differentiate privacy or synthetic data.

REFERENCES

“Terminology Harmonisation in Data Sharing and Disclosure Terms and Definitions”. Available at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/WP063.pdf.

“De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach”. Available at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+and+Anonymization+of+Individual+Patient+Data+in+Clinical+Studies+a+Model+Approach.pdf.

“Data Anonymisation and Risk Assessment Automation”. Available at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf.

“The WHO Body mass index – BMI”. Available at https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi.

“The WHO BMI-for-age (5-19 years)”. Available at https://www.who.int/toolkits/growth-reference-data-for-5to19-years/indicators/bmi-for-age.

“Computation of Centiles and Z-Scores for Height-for-Age, Weight-for-Age and BMI-for-Age”. Available at https://cdn.who.int/media/docs/default-source/child-growth/growth-reference-5-19-years/computation.pdf?sfvrsn=c2ff6a95_4.

ACKNOWLEDGMENTS

I would like to express gratitude to Veronique Poinsot, Raphael Bejuit, Jianfeng Ye for their continuous contribution and guidance in encouraging me to proceed bravely into the unknown.

RECOMMENDED READING

- *K. El Emam, S. Rodgers, and B. Malin, “Anonymising and Sharing Individual Patient Data,” BMJ, vol. 350, p. h1139, Mar. 2015.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Huan Lu
China Analytics and Reporting Tools, Biostatistics and Programming
Sanofi
39F, Chengdu Yintai in99 Tower 3, Chengdu, Sichuan, China
+86 28 6222 8024
Bryce.Lu@sanofi.com