Original Article

Data Organization and Management for Longitudinal DERVAN Cohort Study in KONKAN Region of India (DERVAN-3)

Suvarna Patil[1], Netaji Patil[2], Omkar Dervankar[3], Dnyaneshwar Jadhav[4], Charudatta Joglekar[5]

[1]Professor, Department of Medicine, BKL Walawalkar Hospital and Rural Medical College, Sawarde, Taluka-Chiplun, District-Ratnagiri, Maharashtra, India, [2]Consultant Radiologists, Department of Radiology, BKL Walawalkar Hospital and Rural Medical College, Sawarde, Taluka-Chiplun, District-Ratnagiri, Maharashtra, India, [3,4]Statistician, Regional Centre for Adolescent Health and Nutrition, BKL Walawalkar Hospital and Rural Medical College, Sawarde, Taluka-Chiplun, District-Ratnagiri, Maharashtra, India, [5]Biostatistician, Regional Centre for Adolescent Health and Nutrition, BKL Walawalkar Hospital and Rural Medical College, Sawarde, Taluka-Chiplun, District-Ratnagiri, Maharashtra, India.

How to cite this article: Suvarna Patil, Netaji Patil, Omkar Dervankar et. al. Data Organization and Management for Longitudinal DERVAN Cohort Study in KONKAN Region of India (DERVAN-3). Indian Journal of Public Health Research and Development / Vol. 15 No. 4, October-December 2024.

Abstract

Background: A skilful data management is the heart of any cohort study as results of such studies have far reaching impact on development national policies which will improve public health.

Methods: DERVAN cohort is a longitudinal study set up in KONKAN region of India. The study is expected to last at least 20 years. It plans to investigate impact of adolescent growth, diet and cognition on the risk of development of non-communicable diseases in the adulthood with diabetes as the main focus. The study will also investigate parents and investigate their contribution to the risk. We plan to recruit 1520 adolescent girls. MS Access was used to design data management system. Each family was given a unique identity number. At base line we have created 21 tables, 10 queries for adolescent data and 5 tables, 2 queries for parental data. The data is analysed using statistical software SPSS.

Conclusion: Though our system may not be designed using high end data base systems, it still caters to our needs in the initial stage of the project but its skilful design is expected to make a smooth adaptation to new database environment in the subsequent stages.

Key words: Cohort, Epidemiology, Data Management, Rural, MS-Access

Corresponding Author: Suvarna Patil, Professor, Department of Medicine, BKL Walawalkar Hospital and Rural Medical College, Sawarde, Taluka-Chiplun, District-Ratnagiri, Maharashtra, India.

E-mail: dr.suvarnanpatil@gmail.com

Submission date: December 24, 2023 **Revision date:** January 29, 2024 **Published date:September 20 , 2024**

Introduction

Epidemiology is the science and practice which describes and explains disease patterns in populations, and puts this knowledge to use to prevent and control disease and improve health[1]. In 1990 David Barker laid down foetal origins of adult disease (FOAD) hypothesis[2] which was subsequently came to be known as Developmental origins of Health and Disease (DOHaD) hypothesis[3] which states that exposure to under nutrition in early life can influence the risk of Non Communicable Diseases (NCD) like diabetes, hypertension and cancer in the adulthood. Since then the world including India has witnessed extensive research on DOHaDhypotheis and epidemiology of NCD. The research is of great relevance to India which is witnessing a rapid rise in the prevalence of NCDs[4-5]. Substantial evidence has been generated in the western countries[6-8] as well as India[9-10] by carrying out cohort studies many of which span across generations,[11-14] Data collected in any cohort study needs to be visualised, structured and eventually analysed statistically. In rapidly changing world of today cohort studies of intergenerational nature face constantly changing scenarios of data collection methods, data storage and security technologies, data access methodologies and data analysis techniques. Only skilful data management (DM) will make a smooth adaptation to these scenarios. The most critical part of cohort study is its ability to link data of various participants at different stages/ waves. Many research institutes involved in cohort studies especially in the developing world may not have up to date state of the art facilities for DM. The reasons are plenty but the main reasons are financial cost involved in acquiring database management system softwares, unavailability of human resources to manage data. Many cohort studies are going in remote regions where modern communication technologies for data transfer like mobile, internet may be available but the connectivity is very poor. Despite all these hurdles the task of ensuring data quality, security and accessibility needs to be done using the skilful use of available technical and human resources. Managing cohort data over the long period in itself is a challenging task.

BKL Walawalkar Hospital is located at a village of Dervan in Ratnagiri district of the western Indian state of Maharashtra midway between Mumbai the state capital, business hub of India and southern Indian state of Goa, an international tourist destination. In June 2019 institute launched a DERVAN cohort study[15] which will longitudinally follow more than 1500 adolescent girls (16-18 year old) for next 20 years. The study will test hypothesis that poor physical growth and poor nutrition in adolescence will increases the risk of NCD, in particular the risk of diabetes in their adulthood and in their offspring.

This manuscript describes the data organization and data management protocol for the current stage (stage-1) and also proposes protocols for the subsequent stages.

Material and Methods

Ethics and consents:

Informed and written consent was obtained from all the participants to use the data. For the adolescents below 18 years of age, informed written assent was obtained from adolescents and in addition parental consent was taken. Re-consent procedure will be followed for adolescents after completing age of 18 years. The study was approved by the Institute Ethics Committee of BKL Walawalkar Rural Medical College and Hospital. Our institute ethics committee is registered with the Government of India. Registration code is EC/755/INST/MH/2015/RR-18.

Pro-forma design: Initially the project team members involved in data collection activity were briefed about the project. The data which was going to be collected in the first stage of the cohort could be broadly classified into the 8 categories which were registration, body composition, clinical examination, nutrition, socio-economic, cognition, physical activity and laboratory. Details of individual categories are displayed in (Fig.1). We are also collecting data on parents and details are displayed (Fig. 2). Teams were formed for each category and were given the task of pro-forma design. After initial design few pilot runs for the study were conducted. Pro-forma were refined and improved upon after each pilot run and eventually finalised. We kept pro forma design as simple as it can be with minimal number of skips and

jumps, yet it could extract all the relevant information needed to carry out a particular measurement task. All the proformas as were assembled together to form a booklet so that all hard copy of the data could be kept at one place.

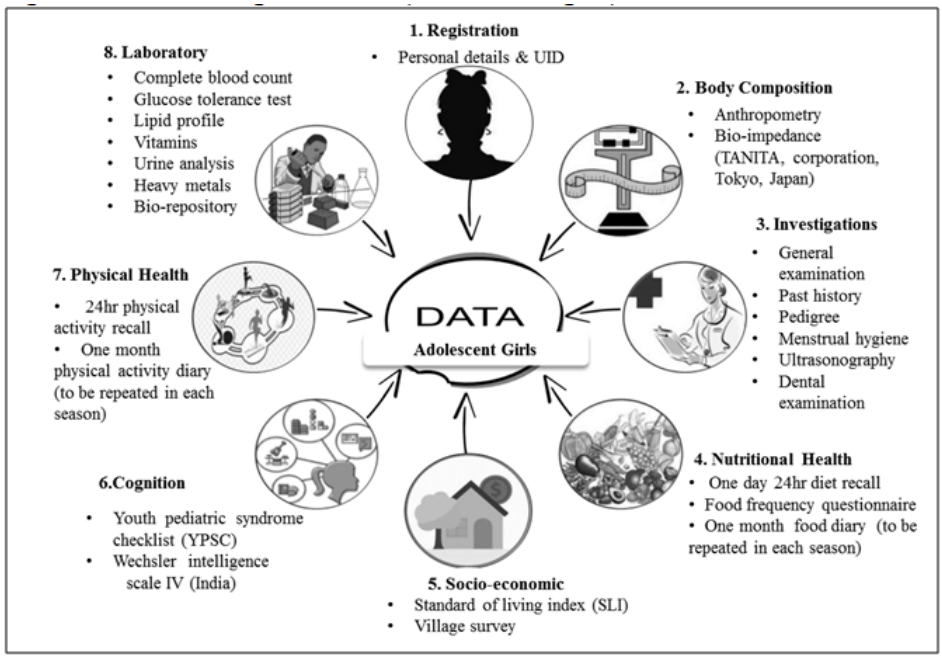

Figure 1: Different categories of data (for adolescent girls)

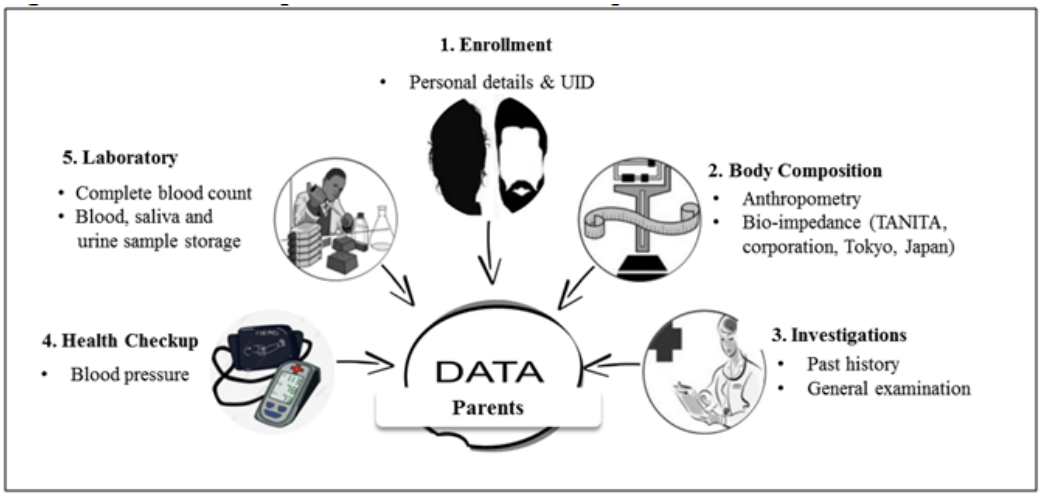

Figure 2: Different components of data collected on parents

Data design and entry:

Being a longitudinal study, the most crucial part was to identify the adolescent girl and her related information at any time point of the study. A permanent Unique IDentity number (UID) was generated. Parents (father and mother) are identified with the same UID so that entire family information at any stage/wave of the cohort can be extracted. We designed data tables using MS Access (MSAC). Details of all the tables in each of the data category are shown in Table-1. Each section of pro-forma has separate table structure. All tables are joined with relational database system. Data entry forms (Graphical User Interface) have been designed to look exactly similar to pro-forma page being entered(Fig.3). It eases the data entry task and minimises errors during data entry. Just now there are five users accessing this database for data entry.

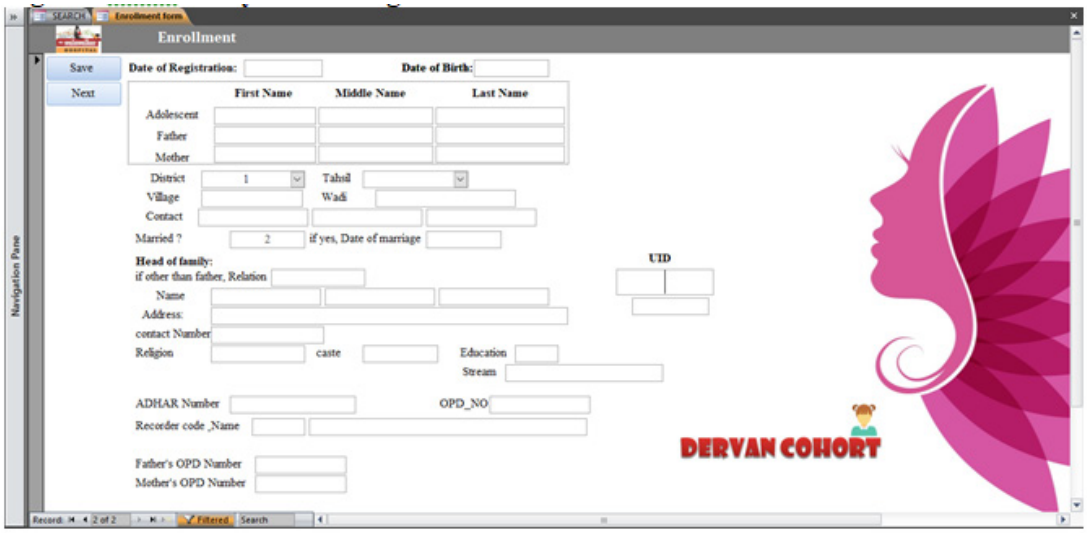

Figure 3: Data Entry form Using MSAC.

Table 1: Details of database (for adolescent girls) with categories

Sr. No	Category	Name of table	Total No of fields	Numeric fields	Text fields	Date/ time	Combo fields	Other	Remark
1	Registration	Enrollment	40	6	29	3	9	2	
2	Body Composition	Subject Anthropometry	35	30	3	2	3	0	
		TANITA	Machine generated data						
3	Investigations	Past History	64	52	10	2	12	0	*
		General Examination	59	47	12	0	40	0	
		Pedigree (for history of NCDs)	32	13	1	0	0	18	*
		Menstruation	24	21	1	1	0	0	
		Ultrasonography	60	55	2	3	22	0	
		Polycystic Ovary Syndrome	36	25	2	1	17	6	
		Dental Examination	74	72	2	0	6	0	
4	Nutritional	Nutrition1 pattern	27	27	0	0	23	0	
		Nutrition3	29	2	27	0	10	0	
		Nutrition values	34	30	4	0	0	0	
		Food diary	7	6	0	1	0	0	*
5	Socio-economic status	Standard of living index (SLI)	49	43	6	0	13	0	*
		Village survey	36	24	12	0	0	0	*
		Village facilities	56	56	0	0	0	0	*

Continue.....

6	Cognition	Youth Pediatric Symptoms Checklist	41	40	1	2	1	0	
		Wechsler's Adult Intelligence Scale	32	19	1	2	0	0	
		Wechsler Intelligence Scale for Children	32	19	1	2	0	0	
7	Physical health	Physical Activity	19	7	12	0	0	0	
8	Laboratory	Lab data	94	91	3	0	19	0	*
		Bio-repository	-	-	-	-	-	-	

*Partially filedon filed or at home visit

Update of reference files

Some data (registration, nutrition and physical activity diaries) gets updates daily as the study progresses. So whenever we update such data, relevant queries in databases will automatically get updated.

Data Security

Database is encrypted with a password. Monthly backup of all databases is set up locally as well as on institute server.

Data export and cleaning

As MSAC uses ODBC (Open Database Connectivity) drivers, it is easily imported in statistical software such as SPSS, STATA. Double data entry is not possible because of limited human resources. But we minimize the errors by exporting data in statistical tools where we have designed some utilities (checking for impossible values, genuine or spurious outliers) to identify errors which are then corrected in the main database system at the same time.

Data analysis

Before going to actual analysis it is necessary to make all files in standard format such as ordering by UID, variable widths, recoding variables etc. Initial ready syntaxes are made for each file. Keeping separate syntax file for each table is always better. Master file and its master syntax file is generated after cleaning of all separate tables.

Data confidentiality

We do not disclose subjects' identity. They are always identified by UID. Their identities (name, address, photo, contact details etc.) are stored in the registration database. None of the other databases have this information. We are not sharing our data with any outside research institutes. In long term we might have to share the data where our subjects' identities will be anonymous.

Results and Discussion:

We have described the data management protocol for an intergenerational longitudinal cohort study in the resource limited KONKAN region of India. Data organization in epidemiological research has been discussed in the literature[16-18]. Also we were able to identify few studies which have used MSAC for data organization [19-21].We preferred MSAC to store data. The foremost reason was the cost involved. It is available as a part of MS Office even in a desktop setup. It is a good system to store relational database. We can collect data from multiple tables in one place and change it too. We are able to enforce several validations to ensure that only data which meets the norms set by our project protocol is entered. It also lets multiple users open a single database at the same time; thereby allowing users the ability to edit different records without conflict. It prevents orphaned records and has variety of ways to view the data. It allows the use of Structured Query Language (SQL) to quickly retrieve just the rows and columns of data contained in one table or many tables. It is an excellent application for creating modest databases for users who may not be familiar with technical language. Though SQL databases have far more capacity and are widely used in cohort research data management[22] systems used are much more technical. MSAC database is more suitable for

desktop use with a small number of users accessing it simultaneously. It is more compatible than SQL. It is possible email copy of our database (in scenarios like work from home, work at different project sites). MSAC interface is mainly for end users unfamiliar with more complex database interactions. SQL does not offer the forms and drag and drop query creation that access uses. MSAC allows users to create tables and queries by manipulating icons and using wizards. SQL is more for the experts and only gives the users and command line interface so it is less intuitive and takes a longer time to run. Considering all these issues we felt that MSAC is an affordable solution for our research study that do not need huge number of records for storage at least for now. Syntax and structures are simple. MSAC has some limitations too. All the information from your database is saved into one file. Even though Microsoft states that MSAC is able to support 255 concurrent users, it is a more practical choice to select MSAC when the database will be used by only 15 to 20 simultaneous users. It is recommended that all MSAC users operate with the identical operating system but this is not always possible. Almost all textual information in our data is coded in numbers so that it eases import in statistical tools for further analysis but there are situations where coding is not possible. Analysing such data statistically is challenging. One such situation (food diary data) is elaborated in (Fig. 4). Food items consumed were listed in diary. But the name of the food item could be spelt differently within the same diary or between diaries of individuals making even simple data analysis like calculating frequency of certain food item a difficult task. Also there is no limit to food items consumed by an individual. The poor data design creates unnecessary blank columns in the datasheets. We overcame this by uniquely identifying all the food items using food code list which was updated immediately whenever the new food item is discovered. It takes care of spelling mistakes. We encoded the data using combo box (drop down list) which is a useful tool. It requires less space to store data. It is easier to store in a single column than to have multiple columns for various options.

Figure 4: Food diary data and coding

Subject	Food items noted by the subject	Food code given in system
Subject 1	**Chapatti**	**4**
	Rice (plain)	1
	Dal(Pigeon pea)	**97**
	Milk	52
Subject 2	**Chapati**	**4**
	Plain Rice	1
	Curry (Pigeon Pea)	**97**
Subject 3	**Roti (wheat)**	**4**
	pickel	113
	Pulav (Veg)	149
	Paneer masala	309

Data generation using various devices is becoming increasingly common in healthcare informatics. This eliminates transcription errors and also reduces paper work. But in many research setups these devices are part of the hospitals where similar data is generated for clinical purpose. In such scenarios filtering out the data for a particular research study needs to be done skilfully. System designed to guarantee the uniqueness of the subject (UID in our case) may not be compatible with the subject identification system of the device. Our body composition data was generated using MC-780 body composition analyser (TANITA Corporation, Tokyo, Japan). The data generated was stored on the device itself. We were able to generate UID while measuring the body composition, but since parents were also identified by the same UID we modified our body composition data storage protocol by ordering the data by UID and further grouping by subjects, their parents. It helps us to export such data using a single query as per our requirement. It creates an excel worksheet which can be imported in the statistical software. We used ultrasonography device (Philips 11HD)to measure organ sizes but it does not have the backend facility to store the data, hence we were compelled to use transcription. There are some public domain health care applications for data entry and storage like Epi Info, OPEN EMR, OPEN MRS2, OPEN CILNICA.

They do have some good features like user friendly form designs but they may not be able to take care of research study protocols. Main disadvantages are difficulties in set up without significant knowledge of programming, limited analysis options, difficult data exporting mechanisms. Many of these platforms require either a strong in-house staff with significant technical skills and experience to install and maintain the system.

Our cohort study is in the initial stage of recruitment and generating the baseline data. But as the cohort matures data volume will increase tremendously. Many of the data items in the future stages will be images, audios, videos requiring large storage space. But MSAC can hold only upto 2GB data. Thus a new data management strategy will need to be evolved for long term use. Data organization and DM are very crucial to any research but any published research in epidemiology there is hardly any mention or discussion of DM component. Research journals should start demanding a section on DM methods which could be placed after a section on statistical methods in the manuscripts.

Comments

In long terms studies it is necessary to build a healthy relationships with the community from where our subjects come. This will help us in minimising the dropouts and data attrition. Just now as a part of public relations we generate identity cards and participation certificates for our subjects. We send them birthday reminders to enrich the bonding with our research team. Health care tips and nutritional counselling are provided through chat groups. Subjects coming to the hospital with health issues get guidance from our project study members.

Overcoming of hurdles

Since our cohort investigations are done on hospital campus, subjects are needed to be registered on hospital management system by generating a unique outpatient department (OPD) registration number. Our cohort DM system is completely separate from hospital investigations. We are storing the OPD registration number as well as unique identity number (UIN) given by the government of India as data items in our system. OPD number helps us to extract any future hospital visit data if needed. UIN helps the subjects in availing the benefits of various government schemes.

Because of unexpected calamity of COVID-19 it has affected the enrolment. We are unable to do some investigations (blood tests, body composition). But the data like diet and physical activity recall is collected through phone interviews and message chat with subjects. Online health care sessions are conducted but again this is constrained by poor internet connectivity.

Future system updates

At initial stage all the subjects are unmarried. We are planning to follow adolescent girls annually with limited number of investigations. The details are laid down in (Fig. 5). But as the cohort girls reach adulthood, new investigations will be planned. New databases will need to be created. Subjects will be exposed to different events of life cycle (for e.g.: marriage, divorce, migration, pregnancy, delivery, post- delivery, abortions and deaths). We are planning to develop event driven updates in our system.

In future we plan to use automated reminders via text messages to subjects. These reminders will have birthday greetings, upcoming visit notices, health care tips etc.

This will keep alive their interest in the project. Also we are in a process to develop a project website where subjects will be able to update the crucial information (like date of marriage, date of delivery, address/contact change if any etc.).

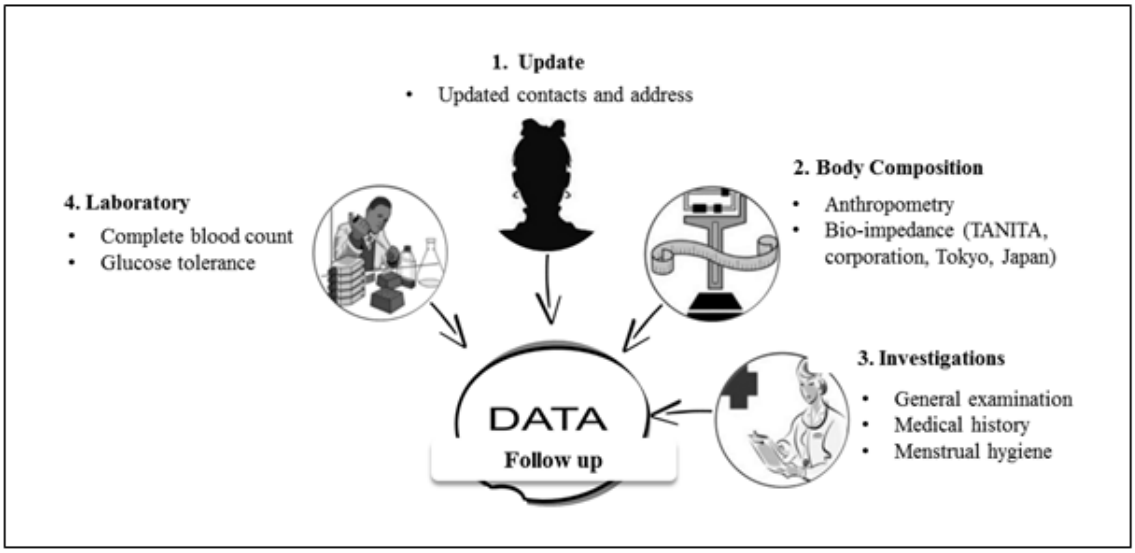

Figure 5: Annual follow up data.

Conclusions

We were able to do the skilful design of data management system for a cohort study using MS-Access which was cost effective. Our methods can be used as a guiding template for any resource limited region where extensive epidemiological studies of longitudinal nature are carried out.

Conflict of Interest: None of the authors have any conflict of interest to declare.

Acknowledgements: We would like to thank our cohort subjects for participation.

Funding

Rajiv Gandhi Science and Technology Commission, Government of Maharashtra, India, funds the study. Funding reference is https://rgstc. maharashtra.gov. in/ sanctioned- projects. Grant Award sanction reference is RGSTC/File-2018/ DPP-195/CR-38.

Author's contributions:

Suvarna Patil is the principal investigator and made valuable suggestions, Netaji Patil supervised the radiology data collection, Omkar Dervankar designed the databases and developed queries as needed and also wrote the initial draft, Dnyaneshwar Jadhav was involved involved in data acquisition and subsequent statistical analysis, Charudatta Joglekar supervised data management as well as statistical analysis and manuscript development.

References

1. Last JM. A dictionary of epidemiology 4th ed. New York: Oxford University Press;2001.
2. Barker DJ. The fetal and infant origins of adult disease. BMJ. 1990;301(6761):1111.
3. GluckmanPD, Buklijas T, Hanson MA. The Developmental Origins of Health and Disease (DOHaD) concept: past, present, and future.Cheryl R, ed. Boston, MA: Academic Press; 2016.
4. Abhinav RP, Williams J, Livingston P, Anjana RM, Mohan V. Burden of diabetes and oral cancer in India. J DiabCompl. 2020;34(11):107670.
5. Mohan V, Anjana RM, Unnikrishnan R, Venkatesan U, Uma SG, Rahulashankiruthiyayan T, Samhita SK, Coimbatore SS. Incidence of hypertension among Asian Indians: 10 year follow up of the Chennai Urban Rural Epidemiology Study (CURES-153). J DiabCompl. 2020;34:107652.
6. Eriksson J. Developmental pathways and programming of diabetes: epidemiological aspects. J Endocrinol. 2019;0680.
7. Plant DT, Pawlby S, Sharp D, Zunszain PA, Pariante CM. Prenatal maternal depression is associated with offspring inflammation at 25 years: a prospective longitudinal cohort study. Transl Psychiatry. 2016;6(11):155
8. Huang RC, Prescott SL, Godfrey KM, Davis EA. Assessment of cardiometabolic risk in children in population studies: underpinning developmental origins of health and disease mother-offspring cohort studies. J Nutr Sci.2015; 4:69.

9. Yajnik CS, Deshmukh US. Maternal nutrition, intrauterine programming and consequential risks in the offspring. Rev Endocr MetabDisord. 2008;9(3):203-11.

10. Krishnaveni GV, Yajnik CS. Developmental origins of diabetes-an Indian perspective. Eur J ClinNutr. 2017;71(7):865-69.

11. Sinha S, Aggarwal AR, Osmond C, D Fall CH, Bhargava SK, Sachdev HS. Intergenerational Change in Anthropometric Indices and Their Predictors Among Children in New Delhi Birth Cohort. Indian Pediatr. 2017;54(3):185-92.

12. Sachdev HS, Fall CH, Osmond C, Ramakrishnan L, Biswas SK, Leary SD L, Reddy KS, Barker DJ, Bhargava SK. Anthropometric indicators of body composition in young adults: relation to size at birth and serial measurements of body mass index in childhood in the New Delhi birth cohort. Am J ClinNutr. 2005;82(2):456-66.

13. Antonisamy B, Raghupathy P, Christopher S, Richard J, Rao PS, David JP Barker, Fall CH. Cohort Profile: the 1969-73 Vellore birth cohort study in South India. Int J Epidemiol. 2009;38(3):663-9.

14. Raghupathy P, Antonisamy B, Geethanjali FS, Saperia J, Leary SD, Priya G, Richard J, Barker DJ, Fall CH. Glucose tolerance, insulin resistance and insulin secretion in young south Indian adults: Relationships to parental size, neonatal size and childhood body mass index. Diabetes Res ClinPract. 2010;87(2):283-92.

15. Patil S, Patil N, Joglekar C, Yadav A, Nilwar A, Banawali U, Bhat R, Domable V, Warape B, Mohite R, Joshi K.aDolescent and prEconception health peRspectiVe of Adult Non-communicable diseases (DERVAN): protocol for rural prospective adolescent girls cohort study in Ratnagiri district of Konkan region of India (DERVAN-1). BMJ Open. 2020;10(9):e035926.

16. Meyer J, Ostrzinski S, Fredrich D, Havemann C, Krafczyk J, Hoffmann W. Efficient data management in a large-scale epidemiology research project. Comput Methods Programs Biomed. 2012;107:425-35.

17. Youngblut JM, Loveland-Cherry CJ, Horan M. Data management issues in longitudinal research. Nurs Res. 1990; 39;188-9.

18. GoldingJ. Data organisation and preparation for statistical analysis in a longitudinal birth cohort. Paediatr Perinat Epidemiol. 2009;23:219-25

19. Schneider JK, Schneider JF, Lorenz RA. Creating user-friendly databases with Microsoft Access. Nurse Res. 2005;13:57-75.

20. Lee H, Chapiro J, Schernthaner R. How I do it: A practical database management system to assist clinical research teams with data collecting, organization, and reporting. AcadRadiol. 2015; 22:527-33.

21. Dennert K, Friedrich L, Kumar R. Creating an affordable, user-friendly electronic inventory system for lab samples. SLAS Technol. 2020; 11.

22. Carmichael D. Information technology for longitudinal birth cohorts. PaediatrPerinatEpidemiol. 2009;23:213-8.