Analysis of Data Quality:
Key HMIS Indicators for Maharashtra F.Y. 2018-19

In this Analysis, issues in data quality are analyzed across indicators and districts. Once an issue is identified, its type is defined along with a small description. An instance of the issue in the data is also highlighted. Once the issue is well understood, causes as well as solutions are reflected upon.

Data Quality Issue - 1:

Type: Validity

Description: Certain indicators expressed as percentages exceed 100% and hence violate the validity of the data. Following are examples for a few indicators recording erroneous percentages:

Indicator	Year	Maximum Value (%)
% Institutional Deliveries to total ANC registration	2018-19	107.8
	2017-18	110.4
% Live Births to Total Deliveries	2018-19	101.4
	2017-18	101.2
% Newborns breast fed within 1 hour of birth to Total live birth	2017-18	100.8

Causes/Solutions:

- Data collected on the field is aggregated so that it can be reported as a percentage. Instead of going through this process manually, facility-wise data entry should be made in HMIS Application [1].
- Dynamic forms can be used to automatically aggregate the data on entry [2].
- Such errors can also arise due to data entry or typing errors. To reduce the chance of such errors making it into aggregated indicators, validation checks should be performed [1]. Here is an example of a validation check:
  Number of live births ≤ Number of deliveries;
  which means, live births should always be less than or equal to the number of deliveries registered.

Data Quality Issue – 2:

Type: Completeness

Description: Data for a few indicators was not reported. If such an error is isolated, it was most likely caused at the time of data entry on-field. Here is an instance of such an error:

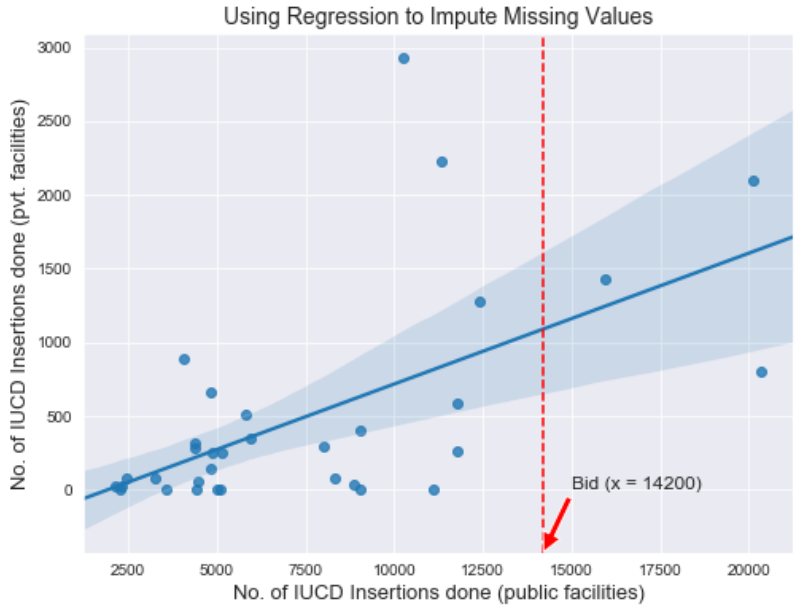
District	Year	IUCD insertions done (pvt. facilities)	IUCD Insertions done (public facilities)
Bid	2017-18	*NaN*	14202

Causes/Solutions:

- HMIS System in India largely relies on simple UNIX/DOS based programs to store and record the data [3]. Because of this, the load of processing/calculation is shifted to the field. This can be addressed by the use of

Databases. Computers have been supplied to almost all the districts which means the hardware is readily available. Using databases helps enforce data quality constraints such as Primary Key and Not Null constraints.

- If the error cannot be prevented on field, we can use Regression to impute the missing data. We can take related indicators to the one with the missing entry as features to our model. Here is an example:

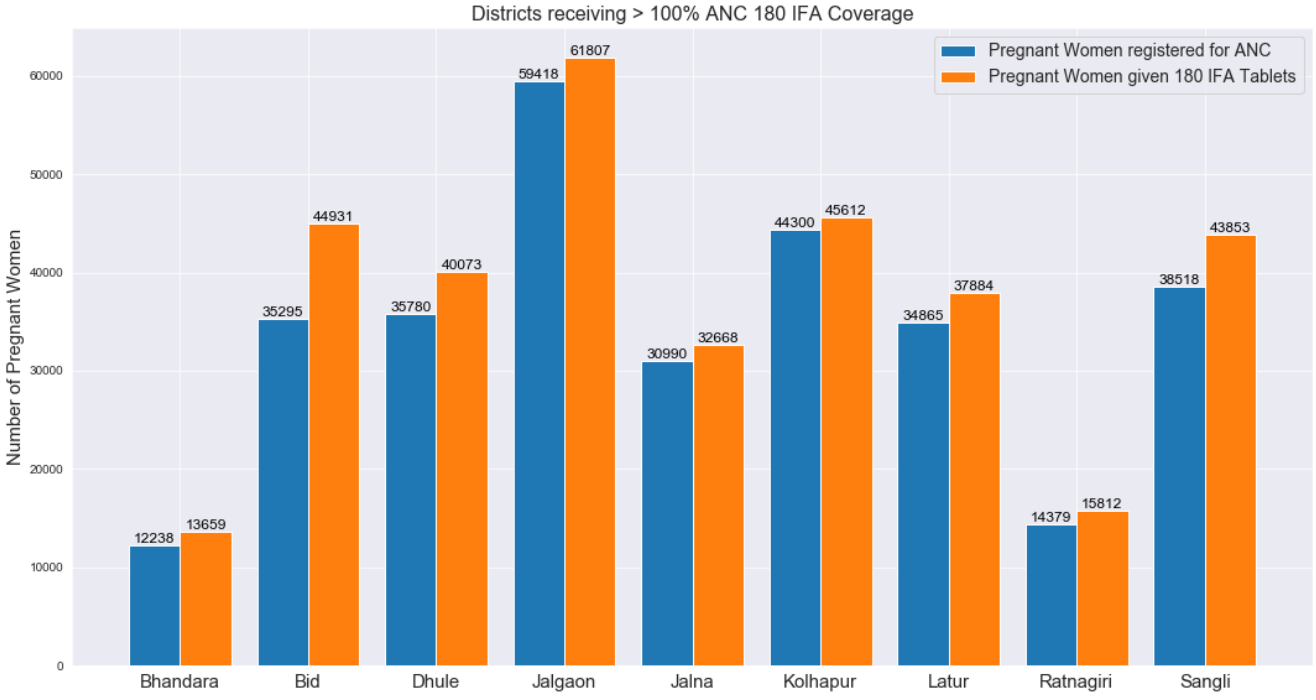


By fitting a model to two co-related features, we can estimate our missing entry. Here, given the value of number of IUCD Insertions done for *public facilities* for the district of Bid, we can estimate the number of IUCD Insertions done for *private facilities* to be along the red line in the shaded region representing the confidence interval. Other methods include simply imputing the data from previous years [1].

Data Quality Issue – 3:

Type: Accuracy

Description: A systematic error was observed in reporting the number of pregnant women given 180 IFA Tablets across a few districts. These districts showed a greater number of women given 180 IFA Tablets than were actually registered for Antenatal Care. Here are the instances of the error:



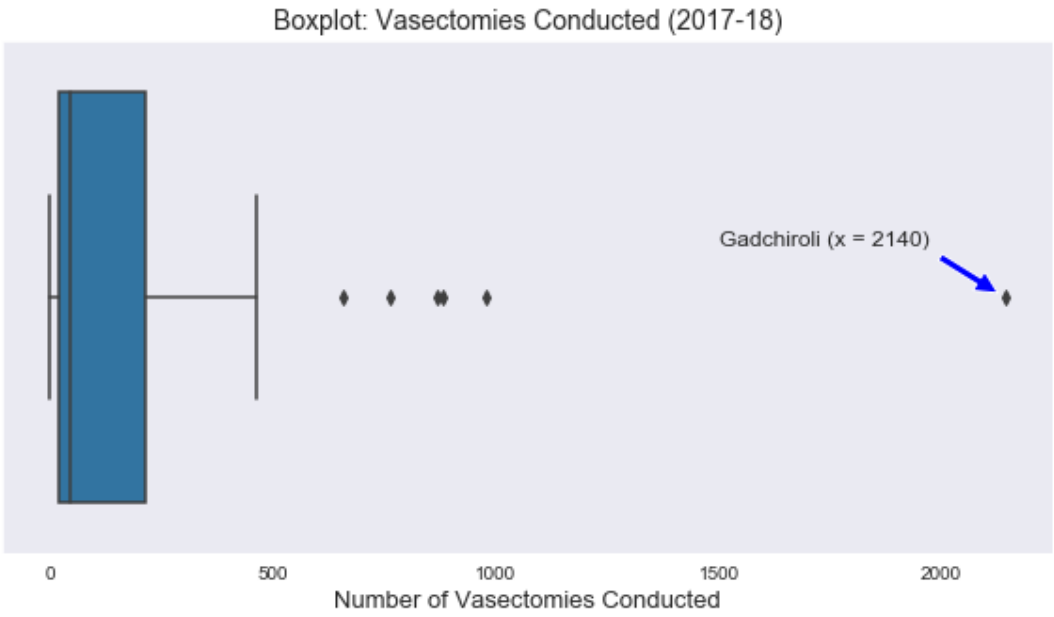
Causes/Solutions:

- Such errors occur due to multiple or poorly designed registers. Recording of information in multiple forms and formats adds to “data weariness” [1]. A compact register should be made to carry to the field. Once the collection process is over, the data should be transferred to the main register at the PHC. Visual inspections or data profiling should be carried out often by Data Stewards.
- Such errors also occur due to duplication of data. We can use Machine Learning algorithms to address this issue by detecting duplicates at the time of data entry by calculating the similarity between two entries. Our algorithm should be able to train on limited, high-dimensional data. Support Vector Machines can be used effectively here, to rank pairs of entries according to their similarity [4].

Data Quality Issue – 4:

Type: Accuracy

Description: Certain indicators showed existence of statistical outliers. An outlier is a data-point that lies beyond 1.5 standard deviations of the mean of the distribution. In the case of number of Vasectomies conducted for the year 2017-18 across districts, we can see 6 outliers in the following box-plot. Although we cannot definitively say that these outliers are errors, we can examine that the data for Gadchiroli lies beyond 3 standard deviations and has a high probability of being an erroneous data point.



Causes/Solutions:

- We can use Deep Learning to detect statistical outliers. For a given indicator, we can calculate the error between the value predicted by our model and the real value observed in the data. If the difference is out of the Upper or Lower Control Limit – which lie 3 standard deviations on either side of the mean – the data-point is marked as an outlier [5]. The obvious limitation here is that we need sufficiently large data.
- This technique can identify data entry errors or large computational mistakes which can be costly at the policy-making level.

Data Quality Issue – 5:

Type: Consistency

Description: An inconsistency was observed in reporting 0's and NaN's. 0's cannot be interpreted as NaN's or vice versa. A value of 0 for the number of women discharged under 48 hours of delivery indicates that no women were discharged while a NaN indicates that the data was simply not reported. Here is an instance of such an error:

District	Year	Number of Women Discharged under 48 hours of delivery in private institutions
Wardha	2017-18	0
	2018-19	0
Jalgaon	2017-18	*NaN*
Latur	2017-18	*NaN*

Causes/Solutions:

- In the absence of consistency, it becomes difficult to procure good quality data. In the above example it is difficult to determine whether no women were discharged under 48 hours of delivery at private institutions or the data is simply missing for the district of Wardha.
- Such inconsistencies arise when there is a lack of guidelines and protocols regarding what to do when data is not reported/received [1]. Facilities should be made aware of the difference between reporting a 0 and not reporting a value.
- Dynamic Forms [2] can be used to auto-fill N/As for the fields where data is not entered or notify the Data Steward about the same so that appropriate action can be taken as per the guidelines.

References:

1. *Health Programme Managers' Manual – Vol. II* (2011)

2. Chen, K., Chen, H., Conway, N., Hellerstien, J. M., & Parikh, T. S. (2011). *Usher: Improving Data Quality with Dynamic Forms. IEEE Transactions on Knowledge and Data Engineering, 23(8), 1138-1153.*

3. Bodavala, R., *Evaluation of Health Management Information System in India. Need for Computerized Databases in HMIS*. pp. 2–5. Available from: hsps.harvard.edu

4. Bilenko, M., & Mooney, R. J. (2003). *Adaptive Duplicate Detection using learnable string similarity measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '03*

5. Dai, W., Yoshigoe, K., & Parsley, W. (2017). *Improving Data Quality through Deep Learning and Statistical Models. Information Technology – New Generations, 515-522*

Prepared by:

Dhaval S. Potdar

dhavalspotdar@gmail.com