SESSION 8: PROCESSING DATA: QUALITY AND ANALYSIS

Aim of the session:

Session 8 aims to promote an understanding of the issues that impact on data quality and to explore the analysis of data using indicators.

LEARNING OUTCOMES:

By the end of this session participants should be able to:

- ✧ Discuss the importance of data quality
- ✧ Define good quality data
- ✧ Describe ways of assessing data quality
- ✧ Suggest ways of improving data quality
- ✧ Calculate selected Comprehensive Plan indicators.

SESSION CONTENTS:

- ✧ Identifying data quality problems
- ✧ What is good quality data?
- ✧ Assessing and improving data quality.

8.1 Identifying data quality problems

The use of **poor quality data can lead to poor decisions** because the decisions are based on incorrect information about the situation. Furthermore, if decision-makers become aware of data quality problems, they may lose confidence in the information system and resort to other ways of making decisions. If data quality is not acceptable, the work involved in collecting the data may be wasted.

Common data quality problems include:

- ✧ **gaps*** (missing data): a data element should never be left blank; either a zero or N/A (not applicable) should be entered.
- ✧ unusual month to month **variations** (that cannot be explained by, for example, seasonal variations, disease outbreaks or the impact of campaigns), i.e. values outside the normal range
- ✧ **unlikely or absurd values** (often a result of data elements not being understood, or poor data collection with many patients not counted)
- ✧ **internal inconsistencies** (e.g. the number of CD4 counts > 200 cannot exceed the total number of CD4 counts done in a month)
- ✧ **duplication** of values (e.g. the same set of values entered for two consecutive months)
- ✧ data **present where there should not be**
- ✧ **writing or typing errors**
- ✧ **arithmetic problems** – poor calculation
- ✧ data entered in **wrong boxes**
- ✧ **preferential end-digits** (when people are "fudging" data, they often tend to use numbers ending in 0 or 5)
- ✧ **counts** of data in the register **do not match** the figures in the month end summary

*Note: * ideally, all data collection tools should be customised to include only the data elements relevant for that reporting unit. In practice, we often end up with generic tools because they are easier to design and reproduce. Forms may therefore contain data elements that are not collected in each reporting unit. In such cases, "NA" should be entered.*

Table 8.1: Examples showing data quality problems

Data from Clinic A:

Data element	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Grand Total
HIV test done ANC	0	0	0	1	1	0	0	2
HIV+ ANC new	0	0	0	1	1	0	0	2
NVP dose in labour	0	4		3	3	0		10
NVP new born	3	4	31	3	3	0		44
Infant HIV+ formula	0	4		3	0	4		11
Infant HIV+ breastfeeding	0	4		1	3	6		14
HIV 1st test baby	0	0		0	0	2		2
HIV 1st test pos	0	0		0	0	0		0
Live birth HIV		4		3	3			10

Quality issues: examples

- Gaps: June, August, November, December
- August: NVP new born: unlikely value, far outside normal range even if one consider the traditional delivery "peak" (10-20% higher than normal) caused by babies conceived during the festive season
- November: Infant HIV+ formula, Infant HIV+breast-feeding: documented, but no NVP or live birth HIV information (possible, but must be investigated)

Quality issues: examples

- Gaps: Gauteng: unlikely zeros and blanks
- Value outside normal range: Male urethritis syndrome - Western Cape 2003
- Unusual fluctuation: STI partner notification rate - Western Cape 2000 - 2004
- Unlikely (impossible) value: VCT HIV + rate - Mpumalanga
- Unlikely value: VCT testing rate: Free State

Table 8.2: Selected indicators – South Africa: 2001 – 2004 (2004 data not complete, and no correction of obvious errors made)

Indicator Name	Period	Eastern Cape	Free State	Gauteng	KwaZulu-Natal	Limpopo	Mpumalanga	North West	Northern Cape	Western Cape
Male condom distribution rate (annualised)	2001	6.84	5.25		6.47	8.09	6.49	5.52	3.42	5.94
	2002	7.28	5.18		6.45	8.84	7.34	4.85	3.55	8.29
	2003	8.40	7.42	0.54	7.24	8.98	2.80	5.08	4.65	10.12
	2004	8.65	5.23	4.36	6.56	8.43	4.14	5.16	4.00	13.39
Male Urethritis Syndrome rate	2001	28.83	28.80	23.35	21.51	26.32	25.96	25.40	31.35	28.43
	2002	31.52	29.38	19.98	23.76	25.75	34.05	23.17	31.68	27.66
	2003	27.34	27.05	20.77	22.69	24.42	44.14	23.30	31.50	44.80
	2004	25.78	25.10	23.97	25.85	24.12	40.43	24.10	33.20	28.61
STI partner notification rate	2001	81.18	12.15		40.06	59.83	64.32	61.85	0.90	65.41
	2002	87.76	13.82		86.05	64.75	88.89	57.11	24.77	66.66
	2003	87.75	43.61	15.72	92.82	75.66	101.53	60.85	53.40	106.84
	2004	90.66	68.91	84.76	104.16	83.32	105.03	69.90	56.12	74.55
STI partner tracing rate	2001	39.35	40.90		26.35	41.59	26.47	45.39	21.47	23.86
	2002	33.32	39.80		28.75	37.74	31.97	45.28	43.89	23.12
	2003	32.02	42.82	39.53	27.97	33.77	33.01	45.08	30.80	25.93
	2004	28.30	34.64	32.24	25.26	31.64	29.51	44.34	32.07	25.68
VCT HIV positive rate	2004	36.15	45.12	43.91		33.02	124.16	37.74	25.38	
VCT testing rate	2004	71.88	10.68	74.50		59.75	67.24	58.28	91.36	

Quality issues (note that only *gross* deviations/gaps will be apparent when data are aggregated up to provincial level):

- Gaps: Those are not *quality* issues in the strict sense, but a result of some provinces not collecting specific national data elements. Gauteng, which uses 40% of the 300+ million condoms distributed annually, only started collecting data on male condoms distributed and STI partner slips in Aug-Sep 2003. This also explains the very low values seen for 2003, when data collection systems were not yet up well established.
- Value outside normal range: Male urethritis syndrome - Western Cape 2003 (Cause: data error – value after correction 28.5%!!)
- Unusual fluctuation: STI partner notification rate - Western Cape 2000 - 2004 (Cause: data error – value after correction 67.9%!!)
- Unlikely (impossible) value: VCT HIV + rate – Mpumalanga (Cause: data element confusion – value after correction ~ 50%!!))
- Unlikely value: VCT testing rate: Free State (Cause: data element confusion – value after correction 35-40%!!)

So while nearly all the seemingly "gross" outliers in the table above have logical explanations and/or can be easily corrected, the core lesson is this: *if you do not scrutinise and use your own data, you will not be able to pick up and correct mistakes*. Remember Murphy's law: "If anything can go wrong, it will!!" (or the cynics version: "Even if nothing can go wrong, it will!!" ☺)

Table 8.3: Selected PMTCT data Clinic B: January – June 2004

Data Element	January	February	March	April	May	June
Antenatal first visits	247	169	231	244	244	250
Antenatal client tested for HIV		102	117	113	113	125
Antenatal client tested HIV positive – new		7	12	8	8	10
Total births in facility	181	581	167	176	176	170
Live birth to woman with HIV	4	7	4	9	9	5
Nevirapine dose to woman at labour	3	5	4	6	6	5
Nevirapine dose to baby born to woman with HIV	8	7	3	9	9	5

Quality issues: examples

- ✧ Gaps: January
- ✧ Value outside usual range: ANC first visit – February. Investigate: is this an error or was there a problem with the service?
- ✧ Total births – February, clearly incorrect: should the number perhaps be 185 or 158?
- ✧ Same values entered for two consecutive months: April and May
- ✧ Preferential end-digits: June (0 and 5)
- ✧ Internal inconsistencies: January. NVP to baby = 8; live birth to woman with HIV = 4. Investigate also for March: why did the baby not receive NVP?

Note the difference between validating facility raw data and validating provincial indicators. With detailed raw data we can pick up many mistakes that would be invisible in aggregated indicators. Many small mistakes also obviously impact on provincially aggregated indicators, but the error will not result in gross outliers that are easy to spot. Therefore, the closer to the source (both time-wise and geographically) data quality is checked, the better.

8.2 What is good quality data?

The information cycle consists of four stages: collection, processing, presentation and use. Each of these stages involves important components. So far we have focused on stage 1 of the information cycle: collection. Now we go on to look at stage 2: the processing of data.

Processing involves two issues: data quality checks and data analysis. Before data can be analysed for use, the quality of the data must be acceptable, or the indicators will be meaningless. If data quality is not acceptable, the other stages of the information process have little value.

Garbage in → no quality checks → garbage out

Data in → quality checks → information out → decisions based on a true reflection of the situation

Note that calculating indicators as a "trial run" in order to assist with identifying data errors and/or gaps is often a good idea – indicators are often sensitive to errors and they will highlight problems not easily seen by inspecting the raw data. Such "trial runs" must not be confused with the calculation of indicators for reporting and management use.

What is good quality data?

Good quality data has the following characteristics:

- **Correct**: the data is accurate, i.e. the numbers reflect what actually occurred.
- **Complete**: all required data elements are recorded (no gaps) and all reporting units have submitted their reporting forms
- **Consistent**: the data is stable and shows no unexplained variations over time, i.e. the values are in the same range or follow the same trends as previous months. The data is also consistent with that of other similar facilities.

Note: Large variations in data do not always point to quality problems, e.g. a large increase in the number of people seeking health care could point to an outbreak of disease, a successful advertising campaign, or perceived improvement in service delivery and/or drug availability. A sudden decrease in numbers of clients accessing a service could point to a problem with the service.

How is data quality assessed?

Data quality is assessed through two mechanisms, specifically looking for the potential data quality problems already mentioned:

- **Visual scanning**, i.e. looking at registers, summaries, forms and printouts
- **Computerised data quality checks**: e.g. maximum/minimum values, validation rules, indicator trial-run, using graphs and/or maps to experiment with various scenarios.

With some experience, it becomes possible to rapidly identify problems even when quickly scanning (eye balling) the data.

Quality should be acceptable at **all the points** along the path of data flow, i.e. individual patient forms, registers, forms and printouts. *(NB: It is important to emphasise this. If there is a data quality problem, the source of the problem must be identified and corrective action taken there, otherwise there will be a repeat of the problem the following month.)*

8.3 Improving data quality

Data quality must be acceptable at all points along the data flow path, or the information generated will be less useful or even meaningless. We have looked at ways of assessing data quality. But how do we go about setting up systems to make sure that the data we collect is of good quality right from the start?

Ways of improving data quality:

- **Train** staff in data collection, data quality checks (validation) and the use of information for action
- Ensure that data element and indicator **definitions** are understood
- Look for possible **weaknesses in the system**, resulting in double counting or missing of entries
- Make data collection as **easy** as possible: user friendly tools; limited dataset; limited number of forms and registers; limited duplication of entries
- **Pre-test** any new data collection tools before introducing them
- Have clearly **defined responsibilities** at every step in the information cycle
- Have **procedures** in place to formally check data quality
- Provide **feedback** to staff on the quality of the data they submit
- Help staff to understand **why** they collect data: provide feedback on how the data is used by managers, and how they can use data/information themselves either to take local decisions or to lobby for specific management decisions/actions (e.g. better equipment, more support staff and/or lay counsellors, funding for a community campaign)
- If possible errors are identified, look for the **source of the error** and correct where possible
- Identify **gaps in staffing** (for instance lack of data entry clerks) and motivate strongly for vacant posts to be filled

ANNEX TO SESSION 8:

Table 8.4 Fantasia clinic with data errors

	Fantasia ARV clinic			Jun 2004		Jul 2004		Aug 2004		
Row	**Data element**	**Patient group**		**M**	**F**	**M**	**F**	**M**	**F**	**Total**
1	Patients staged	over 14		66	427	495	53	74	519	1634
2		0 – 5		19	22	24	17	16	26	124
3		6 to 14		0	0	0	0	13	17	30
4		Total		75	449	519	70	103	562	1788
5	Patients assessed eligible for treatment	over 14		50	379	39	424	46	456	1394
6		0 – 5		13	17	16	14	12	15	87
7		6 to 14		0	0	0	0	9	11	20
8		Total		63	396	55	438	67	482	1501
9	Patients completed drug readiness training (this month)	over 14		12	170	15	175	20	178	570
10	Registered ART patient: ART start	over 14	Reg 1a	49	134	40	148	42	157	570
11			Reg 1b	0	6	0	9	0	13	28
12			Reg 2	0	0	1	0	1	1	3
13		Preg women	Reg 1a	0	0	0	0	0	0	0
14			Reg 1b	2	0	0	3	0	5	10
15			Reg 2	0	0	0	0	0	0	0
16		0 – 5	Reg 1	1	1	2	5	1	3	13
17			Reg 2	1	2	1	4	1	1	10
18		6 to 14	Reg 1	0	0	0	0	3	7	10
19			Reg 2	0	0	0	0	0	0	0

Row	Data element	Patient group		M	F	M	F	M	F	Total
20		Total		53	143	44	169	48	187	644
21	Patients attending 6 month follow up	over 14		17	223	21	298	24	348	931
22		0 – 5		21	23	17	21	23	26	131
23		6 to 14		0	0	0	0	0	0	0
24		Total		0	0	0	0		0	0
25	CD4 count: Patient on ART at staging	over 14	CD < 200	39	347	37	373	31	449	1276
26		over 14	CD > 200	20	101	16	111	43	135	426
27	CD4 count: Patient on ART at 6 monthly	over 14	CD < 200	3	11	2	9	3	13	41
28		over 14	CD > 200	15	204	17	284	19	337	876
29	Child: CD4 <15% at staging	0 – 5		6	7	3	0	4	7	27
30		6 to 14		0	0	0	0	3	2	5
31	Child: CD4 < 15% at 6 monthly	0 – 5		0	0	0	0	0	0	0
32		6 to 14		0	0	0	0	0	0	0
33	CD4 counts done (Total)			83	670	64	777	103	943	2651
34	Viral load < 400 (baseline)	over 14		11	212	10	320	13	330	896
35		0 – 5		0	1	0	0	1	0	2
36		6 to 14		0	0	0	0	1	3	4
37	Viral load < 400 (at 1st 6 months)	over 14		14	208	13	296	27	334	892
38		0 – 5		0	0	0	0	0	0	0
39		6 to 14		5	2	2	5	3	11	28
40	Viral loads done – Total									0

Row	Data element	Patient group	M	F	M	F	M	F	Total
41	ART patient: weight gain >10% at 1st 6 months	over 14	15	125	10	150	25	225	550
42		0 – 5	12	14	10	12	12	14	74
43		6 to 14	0	0	0	0	0	0	0
44	De-registered ART patient - known death	over 14	4	17	3	23	6	14	67
45		0 – 5	1	0	0	2	1	1	5
46		6 to 14	0	0	0	0	0	0	0
47		Total	5	17	3	25	7	15	72
48	De-registered ART patient - transfer out	over 14	0	1	1	3	0	1	6
49		0 – 5	0	0	0	0	0	0	0
50		6 to 14	0	0	0	0	0	0	0
51		Total	0	1	1	3	0	1	6
52	De-registered ART patient - loss to follow up	over 14	2	7	0	9	1	11	30
53		0 – 5	0	0	1	0	0	1	2
54		6 to 14	0	0	0	0	0	0	0
55		Total	2	7	1	9	1	12	32
56	De-registered AT patient - ART stop	over 14	0	0	0	2	0	1	3
57		0 – 5	0	0	0	0	0	0	0
58		6 to 14	0	0	0	0	0	0	0
59		Total	0	0	0	2	0	1	3
60	De-registered ART patient (excl death & transfer)		2	7	1	11	1	14	35
61	De-registered ART patient - total		7	25	5	39	8	29	113

Row	Data element	Patient group		M	F	M	F	M	F	Total
62	Full-time staff	Doctors		2.5	1	2.5	1	2.5	1	3.5
63		Nurses		0	5	0	6	0	6	5.7
64		Pharmacist		0	0.5	0	0.5	0	5	2.0
65		Nutritionist		0	1	0	1	0	0	0.7
66		Social worker		0	1	0	1	0	1	1.0
67		Counsellor		0	4	0	4	0	4	4.0
68		Data Capturer		0	0	0	0	0	0	0.0
69	Drug stock out (1=Yes; 0=No)	Stavudine	0			0		0		
70		Lamudine	0			0		0		
71		Efavirenz	0			0		0		
72		Zidovudin	1			1		0		
73		Lopinavir/ Ritonovir	0			0		0		
74		Didanosin	1			0		0		

Table 8.5

Note: This table contains corrected data (compared with table 8.4)

Fantasia ARV clinic									
			Jun 2004		**Jul 2004**		**Aug 2004**		
Data element	**Patient group**		**M**	**F**	**M**	**F**	**M**	**F**	**Total**
Patients staged	over 14		66	427	53	495	74	519	1634
	0 - 5		19	22	24	17	16	26	124
	6 to 14		0	0	0	0	13	17	30
	Total		85	449	77	512	103	562	1788
Patients assessed eligible for treatment	over 14		50	379	39	424	46	456	1394
	0 - 5		13	17	16	14	12	15	87
	6 to 14		0	0	0	0	9	11	20
	Total		63	396	55	438	67	482	1501
Patients completed drug readiness training (this month)	over 14		12	170	15	175	20	178	570
Registered ART patient: ART start	over 14	Reg 1a	49	134	40	148	42	157	570
		Reg 1b	0	6	0	9	0	13	28
		Reg 2	0	0	1	0	1	1	3
	Preg women	Reg 1a	0	0	0	0	0	0	0
		Reg 1b	0	2	0	3	0	5	10
		Reg 2	0	0	0	0	0	0	0
	0 - 5	Reg 1	1	1	2	5	1	3	13
		Reg 2	1	2	1	4	1	1	10
	6 to 14	Reg 1	0	0	0	0	3	7	10
		Reg 2	0	0	0	0	0	0	0
	Total		51	145	44	169	48	187	644
Patients attending 6 month follow up	over 14		17	223	21	298	24	348	931
	0 - 5		21	23	17	21	23	26	131
	6 to 14		0	0	0	0	0	0	0
	Total		0	0	0	0		0	0
CD4 count: Patient on ART at staging	over 14	CD < 200	43	347	37	373	31	449	1280
	over 14	CD > 200	23	101	16	111	43	135	429

			Jun 2004		Jul 2004		Aug 2004		
Data element	**Patient group**		**M**	**F**	**M**	**F**	**M**	**F**	**Total**
CD4 count: Patient on ART at 6 monthly	over 14	CD <	3	11	2	9	3	13	41
	over 14	CD >	15	204	17	284	19	337	876
Child: CD4 <15% at staging	0 - 5		6	7	3	0	4	7	27
	6 to 14		0	0	0	0	3	2	5
Child: CD4 < 15% at 6 monthly	0 - 5		0	0	0	0	0	0	0
	6 to 14		0	0	0	0	0	0	0
CD4 counts done (Total)			90	670	75	777	103	943	2658
Viral load < 400 (baseline)	over 14		11	212	10	320	13	330	896
	0 - 5		0	1	0	0	1	0	2
	6 to 14		0	0	0	0	1	3	4
Viral load < 400 (at 1st 6 months)	over 14		14	208	13	296	17	334	882
	0 - 5		0	0	0	0	0	0	0
	6 to 14		0	0	0	0	0	0	0
Viral loads done - Total									0
ART patient: weight gain >10% at 1st 6 months	over 14		10	123	12	189	11	205	550
	0 - 5		12	14	10	12	12	14	74
	6 to 14		0	0	0	0	0	0	0
De-registered ART patient - known death	over 14		4	17	3	23	6	14	67
	0 - 5		1	0	0	2	1	1	5
	6 to 14		0	0	0	0	0	0	0
	Total		5	17	3	25	7	15	72
De-registered ART patient - transfer out	over 14		0	1	1	3	0	1	6
	0 - 5		0	0	0	0	0	0	0
	6 to 14		0	0	0	0	0	0	0
	Total		0	1	1	3	0	1	6
De-registered ART patient - loss to follow up	over 14		2	7	0	9	1	11	30
	0 - 5		0	0	1	0	0	1	2
	6 to 14		0	0	0	0	0	0	0
	Total		2	7	1	9	1	12	32
De-registered ART patient - ART stop	over 14		0	0	0	2	0	1	3
	0 - 5		0	0	0	0	0	0	0
	6 to 14		0	0	0	0	0	0	0
	Total		0	0	0	2	0	1	3

		Jun 2004		Jul 2004		Aug 2004		
Data element	**Patient group**	**M**	**F**	**M**	**F**	**M**	**F**	**Total**
De-registered ART patient (excl death & transfer)		2	7	1	11	1	14	35
De-registered ART patient - total		7	25	5	39	8	29	113
Full-time staff	Doctors	2.5	1	2.5	1	2.5	1	3.5
	Nurses	0	5	0	6	0	6	5.7
	Pharmacist	0	0.5	0	0.5	0	0.5	0.5
	Nutritionist	0	1	0	1	0	0	0.7
	Social worker	0	1	0	1	0	1	1.0
	Counsellor	0	4	0	4	0	4	4.0
	Data Capturer	0	0	0	0	0	0	0.0
Drug stock out (Y/N)	Stavudine	n		n		n		
	Lamudine	n		n		n		
	Efavirenz	n		n		n		
	Zidovudin	y		y		n		
	Lopinavir/Ritonovir	n		n		n		
	Didanosin	y		n		n		

Approved posts for the Fantasia ARV clinic		
	Doctors	4
	Nurses	8
	Pharmacist	1
	Nutritionist	1
	Social worker	2
	Counsellor	6
	Data Capturer	1

ART total patients registered end previous month (May)	over 14	567
	0 - 5	63
	6 to 14	0
	Total	630
Number of patients on waiting list end previous month (May)		6