P22   Measuring the quality of data collection in a large observational cohort of HIV and AIDS

Mariska Hillebregt [1], Elly de Lange-de Klerk [2], Dirk Knol [2], Frank de Wolf [1,3], Colette Smit [1]
[1] Stichting HIV Monitoring, Amsterdam, NL;
[2] Department of Epidemiology and Biostatistics VU University Medical Centre Amsterdam, NL; [3] Imperial College, London, UK

Stichting HIV Monitoring
Meibergdreef 9
1105 AZ Amsterdam, The Netherlands
Phone: 020-566 6473
Email: m.m.hillebregt@amc.uva.nl

Background

The Stichting HIV Monitoring performs the prospective collection of data of 14960 (2008) HIV-infected persons. In the treatment centres, data collectors obtain data directly from the patients' medical files and enter the data online into the database of ATHENA national observational cohort, using specific protocols to standardize the data collection and to minimize errors.

Objective

To determine whether the data collectors collected the information that should be collected.

Methods

- The *outpatient clinic charts* from two anonymous patients were selected. **1**

  The *first clinic chart* consisted of data from the first visit to the outpatient clinic of an HIV treatment centre (once-only data) and three follow-up visits (longitudinal   data). The *second clinic chart* consisted of nine follow-up visits.

- Of the 42 data collectors eligible to participate, *38 data collectors* took part. **2**

- A *standard* for the data to be collected was determined and consisted of 168 items. Because the data monitors normally check the data's accuracy and completeness, and check for the correct use of the specific protocols, they were able to determine the standard for the collected information. **3**

- The items collected by the data collectors were *compared to the standard* and were coded as "correct" or "incorrect". The *percentages* of correct and incorrectly collected data were calculated. **4**

- The *statistical tests* used for the normally distributed variables were Student's t-test and univariate linear regression analysis; those used for the non-normally distributed variables were Wilcoxon rank sum test, spearman's rank correlation coefficient, and Fisher's exact test.
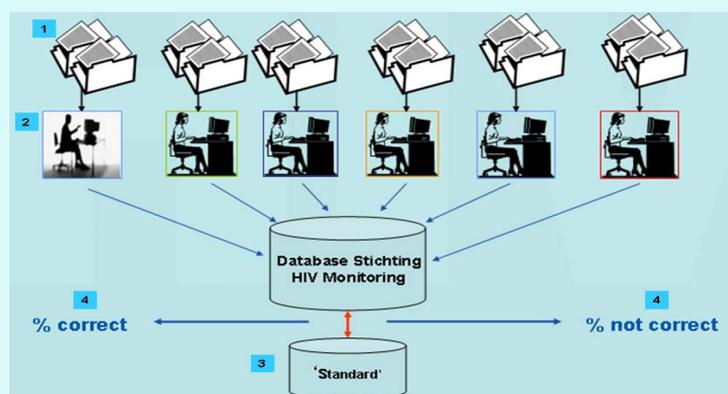
Results

- The percentage "correct" was higher for clinic chart 1 (*mean: 83% correct, SD 7%)* than for clinic chart 2 (*mean: 78% correct, SD 8%)* (figure 1 +2).

- All categories contained incorrectly collected data. The incorrectly collected data were divided into *missing data*, *incorrect start-stop dates*, and *surplus collected data* (data not shown).

- Almost all start-stop dates would *change into "correct"* if "mmyy" was considered correct (instead of the standard "ddmmyy") (data not shown) .

- Not all data collectors used *specific protocols,* although there was no significant difference in the percentages "correct" when the protocols were used and when they were not used (figure 1 +2).

- Sources other than the written comments were not always checked (medical letters, diagnostic information, laboratory results) (data not shown).

Figure 1: Percentages "correct" collected categories, clinic chart 1

Chart 1 N=38 data collectors	Percentages 'correct'						
	Mean	SD ***	Minimum	Maximum	Using protocol Mean (N)	Not using protocol Mean (N)	P-value
Overall:							
correct chart 1 (%)	83	7	65	94	85 (N=16)	82 (N=22)	0.16 *
Categories:							
correct data, only collected at first visit (%)	84	8	68	100	85 ^(N=16)	84 ^(N=22)	0.67 **
correct AIDS-defining events (%)	60	13	33	80	60 ^(N=34)	65 ^(N=4)	0.46 **
correct adverse events (%)	56	31	0	100	57 ^(N=22)	55 ^(N=16)	0.93 **
correct Antiretroviral Therapy (%)	92	11	63	100	86 ^(N=26)	88 ^(N=12)	0.31 **
correct co medication (%)	84	14	50	100	85 ^(N=36)	66 ^(N=2)	0.09 **
correct visits (%) ****	96	7	75	100	-	-	-

* Student's t-test   ** Wilcoxon rank sum test   *** Standard deviation   **** No protocol

^ The 'mean' is given instead of the 'mean rank score' to better understand the differences.

Figure 2: Percentages "correct" collected categories, clinic chart 2

Chart 2 N=36 data collectors	Percentages 'correct'						
	Mean	SD ***	Minimum	Maximum	Using protocol Mean (N)	Not using protocol Mean (N)	P-value
Overall							
correct chart 2 (%)	78	8	58	91	81 (N=6)	77 (N=30)	0.40 *
Categories:							
correct AIDS-defining events (%) ****	100	0	100	100	-	-	-
correct adverse events (%)	75	20	14	100	74 ^ (N=16)	75 ^ (N=20)	0.90 **
correct Antiretroviral Therapy (%)	72	11	47	92	73 ^ (N=29)	69 ^ (N=7)	0.31 **
correct co medication (%)	65	18	33	100	67 ^ (N=30)	57 ^ (N=6)	0.23 **
correct visits (%) *****	91	8	76	100	-	-	-

* Student's t-test   ** Wilcoxon rank sum test   *** Standard deviation   **** No AIDS-defining events had to be collected   ***** No protocol

^ The 'mean' is given instead of the 'mean rank score' to better understand the differences.

Conclusions

- High proportion of correctly collected data (clinic chart 1: **83% correct**, clinic chart 2: **78% correct**)

- Incorrectly collected data were divided into **missing data**, **incorrect start-stop dates** and **surplus collected data.**

Recommendations

- **First**, the specific protocols should be easy to follow, with clear descriptions, but because of the complexity of the HIV disease and the variation in the content of the clinic charts, the protocols remain difficult.

- **Therefore**, data collectors should be more knowledgeable about HIV disease, so they can better recognize what data to collect.

- Data collectors should **also** be trained in the use of difficult protocols to improve their ability to collect data.

- **In addition**, all data collectors need to be trained to check sources other than written comments, especially if the characteristics of the patient contain "data collected at first visit only" and/or events ("adverse events" and "CDC events").

- **Finally**, there should be more agreement among physicians about what information to record in the clinic charts to facilitate data extraction for the data collectors.