METHODS

The NCR methodology follows that recommended by the WHO/IARC.

1.1 Data collection and data flow

The National Cancer Registry (NCR) is a passive pathology-based surveillance system. Copies of pathology reports confirming a cancer diagnosis are submitted voluntarily by both public and private laboratories throughout South Africa. Reports were received from 77 laboratories in 2000 and 2001(see acknowledgements).

The South African total population (Dorrington RE, Johnson L, Bradshaw D, Daniels T. The Demographic Impact of HIV/AIDS in South Africa: National and Provincial Indicators for 2006. Cape Town: Centre for Actuarial Research, South African Medical Research Council, Actuarial Society of South Africa; 2006.) is used as the denominator in calculating incidence rates; therefore all cancer cases in individuals who are clearly not resident in South Africa, for example results of specimens sent to South African laboratories by other countries, are excluded.

Data items are abstracted from the pathology reports: demographic information about the patient and tumour information (topography, morphology and date of diagnosis).

The voluntary nature of cancer surveillance in South Africa can delay data publishing, as data receipt from some of the laboratories is sporadic. Some laboratories submit only summary reports which may lead to cases of incorrect reporting as cross checks cannot be done.

1.2 Reporting of cancer

Only incident cases of primary invasive cancer diagnosed by histology, cytology or haematology are recorded each year. Doubtful*, in-situ* or borderline cancers are excluded*.* Each multiple primary cancer is recorded as an additional case, using the guidelines set out by IARC (1994). Duplicate entries are deleted. Duplicate cancers include cancers that have been diagnosed in previous years that already exist on the registry database.

Cancers are classified by anatomical site/topography using the coding convention of each laboratory, usually a Systematic Nomenclature of Medicine Topography code SNOMED-2 (Côté *et al.,* 1979) or SNOMED-3 (Côté *et al.*, 1993), and a five digit Morphology code (morphological type and behaviour, WHO 1976 and WHO 1992). Metastatic cancers are either coded to primary site of origin, if this information is available and/or is known or to primary site unknown.

From the registry's inception in 1986 to 1991, data were reported in a format compatible with the International Classification of Diseases, 1975 revision (ICD-9), (WHO 1975). Data were reported in ICD-10 format from 1992, in line with the South African Department of Health (DOH) requirements. In 1996 and 1997 incoming data were coded and checked in SNOMED-2 (Côté *et al.*, 1979), although some laboratories were coding in SNOMED-3, which is more compatible with International Classification of Diseases for Oncology, second edition (ICD-O2) and International Statistical Classification of Diseases and Related Health Problems, tenth edition (ICD-10). In 2001, the Cancer Registry followed the international registries' practice and commenced

coding in ICD-O3. The ICD-O3 represents an extension of chapter II (Neoplasms) of the tenth revision of the *International Statistical Classification of Diseases and Related Health Problems* (Percy, Van Holten and Muir, 1990).

1.3 Data quality and quality assurance

The quality of the cancer registry data has been discussed extensively elsewhere (Mqoqi, Sitas and Halkett, 2003)

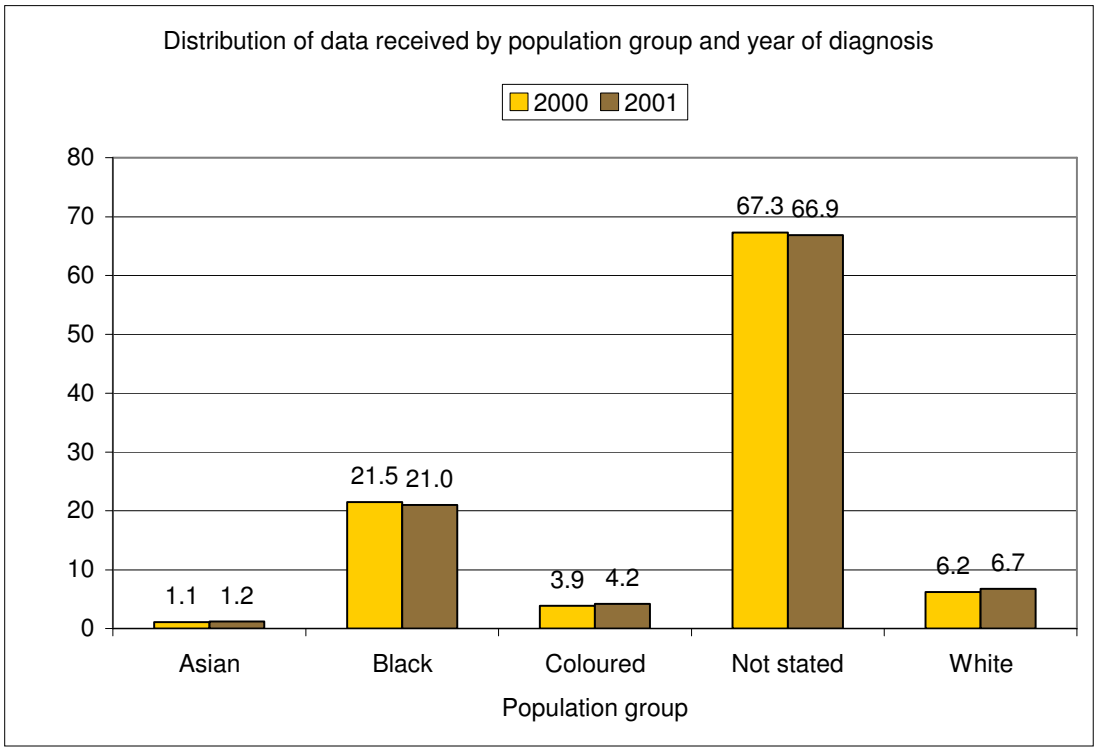
1.3.1 Completeness of data

In an attempt to ensure completeness of data on collected items, query letters on missing and/or seemingly inaccurate data are sent out. Queries include requests for clarification on all cases with unknown age, or sex, as well as unspecified site of origin of cancer, and cases with unlikely/impossible combinations of primary site and sex (e.g. male with cancer of the cervix).

Data analysis is stratified by sex, population group, and age group in order to get a clearer picture of disease patterns and indications of possible risk factors. Exposure to lifestyle factors including dietary choices, socio-economic status, sexual and reproductive health behaviour, tobacco smoking and alcohol consumption among others, are all known to impact on the risk of developing a cancer. These factors often vary by population group.

Misuse of population grouping for political ends in the past has led to overshadowing its epidemiological significance; from 1992 increasing numbers of reports were received without race group information (see Mqoqi, Kellett *et al,* 2003). The proportion with this information missing was approximately 67% for 2000/1.

Figure 1: Received cancer cases by population group



In consultation with the Data Management and Statistical Analysis Unit (DMSA) of Witwatersrand University, a hot-deck imputation method (Little and Rubin, 1990) was used to allocate population group to cases without this information. This method replaces missing values by suitable estimates; it correlates cancer cases with missing population group values against a reference database containing surnames with known group. This method has proved to be reasonably accurate and its results compare well with the previous registry statistics. Surnames which do not appear on the database remain in the group with population group unknown.

1.3.2 Unique identification and use of names

Patient information can mainly be used at two levels:

1. At a primary or clinical level where health providers use patient information for patient management. This level is important for the individual.
2. At a secondary level at which patient information is collated, analysed and extrapolated to make general statements about the health status of the communities or groups of individuals. This level is important for public health purposes and informs policy decision-making and the planning of health services.

At both levels, confidentiality of information is critical. The benefits to communities must be weighed against possible harm to individuals. Usually the secondary level does not require collection of patient names. Unique identification (ID) numbers, where available, are used for surveillance purposes in an effort to protect patient confidentiality. Systems can be put in place to make patient information available beyond the level of care without jeopardising and/or infringing on the ethical rights of the patients. Health promotion programs need to educate both health providers and communities about the importance of, need for, and use of patient information for surveillance and public health purposes.

In the absence of ID numbers being supplied on all reports, use of names cannot be avoided in the NCR for the following reasons:

a. To eliminate the ~20,000 duplicate entries where an individual has several laboratory tests for the same cancer or where a cancer has recurred. The registry reports only the earliest occurrence.
b. To be able to identify individuals with multiple primary cancers. Epidemiology of multiple primaries could be important in identifying associations between diseases, treatment regimens, etc.
c. In the absence of or poor reporting of a descriptive variable like population group, names remain the solution to devised methods such as hot-deck imputation which are used to circumvent the lack of data (*see section 1.3.1*).

1.4 Analysis

Crude incidence rates per 100 000 and the percentage contribution of each cancer site to total number of cancers were calculated. To allow comparison between populations and internationally, age standardised incidence rates (ASR) per 100 000 for each cancer were calculated using the 'direct method' and the 'World population' as standard (Doll, Payne and Waterhouse 1966). Ninety five percent confidence intervals (95% CIs) for the ASR are presented and were calculated using the Poisson approximation for the

standard error of the ASR. This approximation method sometimes results in lower limit values that are less than zero. The cumulative lifetime incidence risk, the likelihood of developing a cancer in one's lifetime if one lives to age 74 (Cumrisk74), expressed as a percentage or as 1 in X number of people ("Lifetime risk", LR74), is also calculated using standard methods (Jensen *et al.*, 1991). The ASR, Cumrisk74 and the LR74 are adjusted for the proportion of cases in the unknown age category. The rate calculations for getting any cancer exclude BCC and SCC of skin cancers. Examples of all the rate calculations are given in Table 1 below:

Table 1: Direct standardisation for cancer of the cervix in black females, 2001

Age group	Observed cases	Allocated cases	Adjusted cases	Person-years	Age-specific ate	Std. world population	Expected cases	Variance for expected cases	Cumulative rate % (5-yr)
A	B	C	D	E	F	G	H	I	J
00-04	0.00	0.00	0.00	2231934	0.00	12000	0.00	0	0.00
05-09	0.00	0.00	0.00	2048377	0.00	10000	0.00	0	0.00
10-14	1.00	0.00	1.00	2035303	0.05	9000	0.00	209491	0.00
15-19	5.00	0.00	5.00	2014265	0.25	9000	0.02	1069447	0.00
20-24	22.00	0.81	22.81	1819867	1.25	8000	0.10	4723405	0.01
25-29	82.00	5.57	87.57	1652676	5.30	8000	0.42	21984249	0.03
30-34	191.00	6.16	197.16	1350657	14.60	6000	0.88	41684234	0.08
35-39	340.00	14.07	354.07	1221366	28.99	6000	1.74	91545366	0.16
40-44	423.00	15.78	438.78	1006106	43.61	6000	2.62	167187299	0.23
45-49	508.00	12.51	520.51	792276	65.70	6000	3.94	319829237	0.35
50-54	497.00	17.69	514.69	587537	87.60	5000	4.38	399349109	0.47
55-59	359.00	17.99	376.99	430145	87.64	4000	3.51	349267849	0.47
60-64	439.00	23.38	462.38	404930	114.19	4000	4.57	483386829	0.61
65-69	311.00	15.38	326.38	314401	103.81	3000	3.11	318372499	0.56
70-74	230.00	9.58	239.58	241324	99.28	2000	1.99	176300181	0.53
75-79	96.00	3.37	99.37	138155	71.93	1000	0.72	55777063	
80-84	68.00	3.28	71.28	98074	72.68	500	0.36	19848018	
85+	17.00	0.74	17.74	55184	32.15	500	0.16	15602128	
Sub-total	3589.00	146.32	3735.32	18442577	21.70	100000.00	28.52	2466136403	
UNK	251.00	15.57	266.57						
TOTALS	3840.00	161.89	4001.89						

Expected cases adjusted for age unknown	30.56
Variance	0.247
Standard error	0.497
95% LCL	29.58
95% UCL	31.53
Cumulative incidence rate (%) 0-74	3.49
Cumulative risk 0-74	3.43
Lifetime risk 0-74	29

Crude rates

Using data for cancer of the cervix for 2001 in Table 1 above, the crude rate is calculated by dividing the total number of cervical cancer cases by the population at risk (4 001.89/18 442 577) and expressing this per 100 000 women ($\times$ 100 000) = 21.70 per 100 000. Note that the calculation of this rate includes those cancers that had missing information about population group, sex or age.

Age-specific incidence rate (ASIR)

Age-specific rates are calculated by dividing the number of cancers in each age category by the population at risk in that age group (column D/E) and multiplying by 100 000. For example, in 60-64 year old women the age-specific rate is:
462.38/404 930 = 114.19 per 100 000 women. Note that the estimation of these age-specific rates included only those cases for which the age was known.

Age standardised rate (ASR)

Any two or more populations will differ somewhat in age structure. If a disease like cancer is related to old age then a comparison of crude rates may be misleading because an older population will show elevated crude cancer rates because of its age structure. To take age differences between populations into account, a standard population of fixed age structure and the 'direct' method of standardisation are commonly used. Several possible populations can be used but, for international comparisons, the World population (Column G) is the commonly used standard for cancers. The direct method involves calculating from each of the age specific rates (Column F, 0-4, 5-9, ...) the expected number of cases that would occur in this World population (e.g. in the 60-64 year age group $114.19 \times 4\ 000/100\ 000 = 4.57$ in Column H). Because the World population adds up to 100 000, the sum of the expected cases (0.01 + 0.01 + 0.01 + 0.02 + ... + 2.08) is the age-standardised rate of this population (28.52 per 100 000).

Adjustment for age unknown

Because the age standardised rate only uses data from those with a known age, a small adjustment has to be made for the proportion of people in the age unknown category. The formula is:

ASR $\times$ (total cases / cases with known age)

$= 28.52 \times 4001.89/3735.32$

= 30.56 per 100 000.

Standard error (s. e.) for ASR and 95% confidence interval

An age-standardised incidence rate calculated from real data is taken to be an estimate of some true value, which could be known only if the units of observation were infinitely large.

A standard error (s.e) gives a measure of precision of the estimated rate and is also used to calculate a confidence interval. The 95% confidence interval represents a range of values within which it is 95% certain that the true value of the incidence rate lies (Kirkwood 1988, Jensen *et al.*, 1991).

The first step is to calculate the variances for each age-specific expected number of cases based on the world standard population. The adjusted numbers of cases (including the cases with age unknown, allocated to age categories *pro rata*) were used. This was carried out using the following formula for women aged 60-64:

$$Var_{60\text{-}64} = \frac{\left(IR_{60\text{-}64} \cdot W_{60\text{-}64}{}^2 \cdot 100000\right)}{N_{60\text{-}64}}$$

Where $Var_{60\text{-}64}$ is the variance of the estimated number of cases; $IR_{60\text{-}64}$ is the age-specific incidence rate per 100 000 for the cancer; $W_{60\text{-}64}$ is the world standard population for the age category 60-64 and $N_{60\text{-}64}$ is the mid year population at risk in that age group.

For cancer of the cervix in black women aged 60-64 during 2001 the age-specific incidence rate was found to be 122.34. This is higher than the value of 114.19 given in column F because that (column F) age-specific rate did not include an allocation of cases from the age unknown group. *W* was 4000, *N* was 404 930, and so the variance was calculated as 483 386 829.

The second step is to sum all the age-specific variances presented in column I to obtain a total of 2 466 136 403.

The variance for the age-adjusted rate for cancer of the cervix (World standard population) is then determined as 2 466 136 403/(100 000*100 000) = 0.247.

The Standard Error is simply √0.247 = 0.4966.

The upper and lower 95%CIs are obtained by adding and subtracting (1.96*0.4966), and were calculated to be 31.53 and 29.58 respectively for cancer of the cervix in black women during 2001.

Cumulative rate and cumulative risk

The cumulative cancer incidence rate can be used to calculate the cumulative lifetime risk (LR), that is, the probability of developing a cancer in one's lifetime (here defined as 0-74 years). The cumulative rate is:

$$\text{Cum. rate} = \sum_{i=1}^{A} a_i \; t_i$$

where A = age class, i = indicator of an age class, a = age specific incidence rate in that age class and t = number of years in each age class (Jensen *et al.*, 1991). If data are presented in five-year age groups the cumulative rate is the sum of five times the incidence rate of each age-specific group of interest (in this case between 0-74 years). So the cumulative incidence (Column J) is 0.00 + ... + 0.61 + 0.56 + 0.53 = 3.49%. Note that the values in column J were calculated using the adjusted numbers of cases including allocations from the age unknown group, and so they are not directly calculated from the data in Table 1.

A small adjustment is now needed to take account of the sequential removal from the population at risk of people who developed cancer of the cervix. This is achieved by using the following formula:

Cumulative risk0-74 (CUMRISK74) = $100 \times [1 - \exp (- \text{cum. rate}/100)]$

$= 100 \times [1 - \exp (- 3.49/100)]$

$= 3.430\%$

The Lifetime risk expressed as risk over the years 0-74 is calculated as 100/CUMRISK74, or 1 in 29.

The assumptions made in calculating lifetime risks are that no other causes of death or disease are in operation, and that no significant changes in exposure occur over time. The interpretation of the lifetime risk is that (because of under-reporting), at least 1 in 29 black women have a lifetime risk of developing cancer of the cervix.

1.5 Presentation of cancer incidence report

This report presents statistics of cancer cases that were newly diagnosed in 2000 and 2001 only. Cancer cases that are already on the registry database and with the same type of cancer are regarded as duplicates and are not reported. Cases already existing in the registry database but presenting with a new and different type of cancer are regarded as new cancers and are classified as multiple primary cancer cases. Cancer incidence rates are at times confused with cancer prevalence which is the number of existing cancer cases at a point in time, irrespective of the diagnosis date. It is important for the cancer report users to differentiate between these two disease measures. Health providers in a clinical setting are likely to see both new cases and those previously diagnosed elsewhere and therefore already on our database.

BCC and SCC of skin are excluded in determining the total rates and risks for all cancers combined.

Details about the cases with missing information about age, sex or population group are presented as appendices.