# THE DANGERS OF DATA

## Recognising the limitations of crime statistics

Antony Altbeker, Institute for Security Studies
aaltbeker@issafrica.org

It is frequently noted that police crime statistics can reflect reality badly because of under-reporting and under-recording. Less frequently noted is the fact that other sources of data can be just as problematic. This article reflects on two sources of statistics on murder – the National Injury Mortality Surveillance System and the MRC's Burden of Disease estimates – and argues that the incautious use of these data can lead to erroneous conclusions.

There are many aspects of criminal justice policy that cannot be decided purely on the basis of empirical data. One example of this is the death penalty: those who believe that putting someone to death is the appropriate response of a society to some forms of criminality will treat the findings of empirical research into whether this reduces crime or achieves any other social good, as irrelevant. Another example is prostitution and narcotics. Here, despite ample evidence that prohibition causes problems which may be greater than the evils caused directly by drugs and the sex industry themselves, many want the police to enforce the law simply because they take a moral stance against these activities.

In these kinds of debate, empirical data about the extent of the problem and the costs incurred by society in seeking to contain it may not be decisive. Conflict about the relative superiority of one approach or another is as much a reflection of contending values as it is a question of which achieves a particular end more effectively or efficiently. Data are not irrelevant to these debates, but they will seldom be decisive.

### The need for good data

There are other areas of social and criminal justice policy, however, in which good data are needed if appropriate decisions are to be made. In designing a police strategy, for instance, it would be useful to know where crime is most concentrated, how sensitive it is to changes in the level of policing, and whether or not it is significantly affected by changes to demographic, housing, welfare or any other policies. But saying that good data are needed if sound decisions are to be made means just that: the data need to be good. If they are not, they may serve to confuse matters. Worse still, they can lead to mistakes.

Given this, it is incumbent on researchers to deal faithfully with their data and to avoid stretching them beyond their limits. Perhaps the most obvious problem with data is that their presentation sometimes makes it very difficult to establish how conclusions were reached and how plausible these are. In these cases, the problem may lie with the data, with the calculations performed on them, or with their actual presentation. In other cases, conclusions drawn from data may be unsupported

by the data themselves. In all cases, however, real harm can be done when the limitations of data are not respected.

This article looks at two recent examples of these problems, both arising in discussions relating to murder rates in South Africa, and contends that, in both cases, illegitimate conclusions were drawn. Given that these errors were made on the basis of the crime data conventionally regarded as the most accurate and reliable, it suggests that researchers and policy-makers ought to be even more careful when dealing with data relating to other kinds of crime.

**Case one: The MRC's per capita murder rate**
In 2004, the Medical Research Council (MRC) released a report on how South Africans die, seeking to establish the rates of death from a wide variety of diseases, as well as from non-natural deaths like traffic accidents, homicides and suicides.[1] (Selected results were published in the *SA Crime Quarterly* No 13 Sept 2005).

The findings suggested that about 1,542 of every 100,000 people in the country in 2000/01 died that year. Of these, 628 (40%) died of communicable diseases (of which 55% were HIV/AIDS-related), 756 (49%) died of non-communicable diseases and 149 (10%) died of injuries including accidents, homicides and suicides.[2]

In all categories, men were more likely to die than women, with the differential being smallest for HIV/AIDS-related deaths and largest for injuries. There were also important variations across the provinces, with the death rate in KwaZulu-Natal being about 50% higher than that of the Western Cape.

*Arriving at the data*
It stands to reason that estimates of this sort require sophisticated statistical modelling. Nowhere in the world are the data required for these reports – which cover 131 separate categories of cause of death – generated automatically. In a developing country context, these problems are accentuated by the fact that some deaths go unreported to the authorities and, even when they are reported, errors

and omissions mean that datasets are not completely reliable.

These estimates, so the writers explain, are, therefore, the result of a number of exercises aimed at calculating the number of people who died in 2000 and from what causes. Sources included:
• the estimates of HIV/AIDS-related deaths computed by the Actuarial Society of South Africa's model of the epidemic, a model that also predicts overall death rates;
• historical data on the causes of non-HIV/AIDS-related deaths based on data compiled from official sources, including a review of 12% of all death certificates submitted to the Department of Home Affairs between 1997 and 2001;
• data from the National Injury Mortality Surveillance System (NIMSS) on the causes of non-natural deaths.

Each of these sources of data provides only a partial and, therefore, flawed picture of reality. As a result, statisticians and demographers have to hammer the data into shape before it will produce the kinds of results that are needed. It is in this process, one in which assumptions must inevitably play a large role, that dangers lurk. And it is here that the MRC's efforts led to a large overstatement of the number of murders that took place in South Africa in 2000/01.

*Counting death*
The MRC's estimate of the number of murders that took place in 2000/01 is derived from three sources. The first is the estimate of the number of all deaths in the country, which is derived from the Actuarial Society's model, ASSA2000, with some modifications. This produced an estimate of about 557,000 deaths.

Then, to calculate the number of deaths as a result of non-natural causes, an estimate of the proportion of all deaths resulting from these causes, established in a separate study, was used.[3] This study looked at a sample of 12% of all death certificates issued between 1997 and 2001, and found, coincidentally, that in 12% of these cases, the cause of death was non-natural. Thus, we have a conclusion that about 12% of all 557,000 deaths was non-natural. This resulted in an estimate of about 67,000 non-natural deaths.

Having established that figure, the MRC then calculated the number of deaths attributable to homicide on the basis of NIMSS data. These are compiled every year on the basis of a survey of all bodies arriving at about 35 mortuaries around the country and include data on the time, place and cause of death as well as various demographic details.

Using these data, which suggest that in 2000/01 murder was the leading non-natural cause of death of bodies presented to NIMSS mortuaries, the MRC calculated that there were 26,683 murders committed in SA in that year at a rate of 59.1 per 100,000 people.[4]

After the age standardisation process, this number became 30,069 murders at the rate of 66.6 murders per 100,000. This is also the figure that appears in the MRC's report. Both figures, however, differ markedly from the number (and rate) of murders reported by the SAPS, namely 21,785 (or 49.8 per 100,000).

One immediate comment about these data is that the MRC's reporting of age standardised rates, as opposed to using the absolute number of estimated cases directly, exaggerates the difference between the MRC calculations and the number of murders reported by the SAPS. The reason for doing this is that South Africa's relatively young population means that when estimates are made of the causes of death, those that affect the young are increased relative to those that affect the old. Even without this adjustment, however, the absolute values of the number and rate of murders predicted by the MRC are, respectively, 23% and 19% higher than those of the SAPS[5] (Figure 1 and Table 1).

One possible reason for the disparity is that the SAPS and the MRC use slightly different definitions of the year 2000/01. For the SAPS, this is from April 2000 to March 2001. The MRC, on the other hand, uses the period July 2000 to June 2001. It is conceivable, in other words, that both the SAPS and the MRC are right.
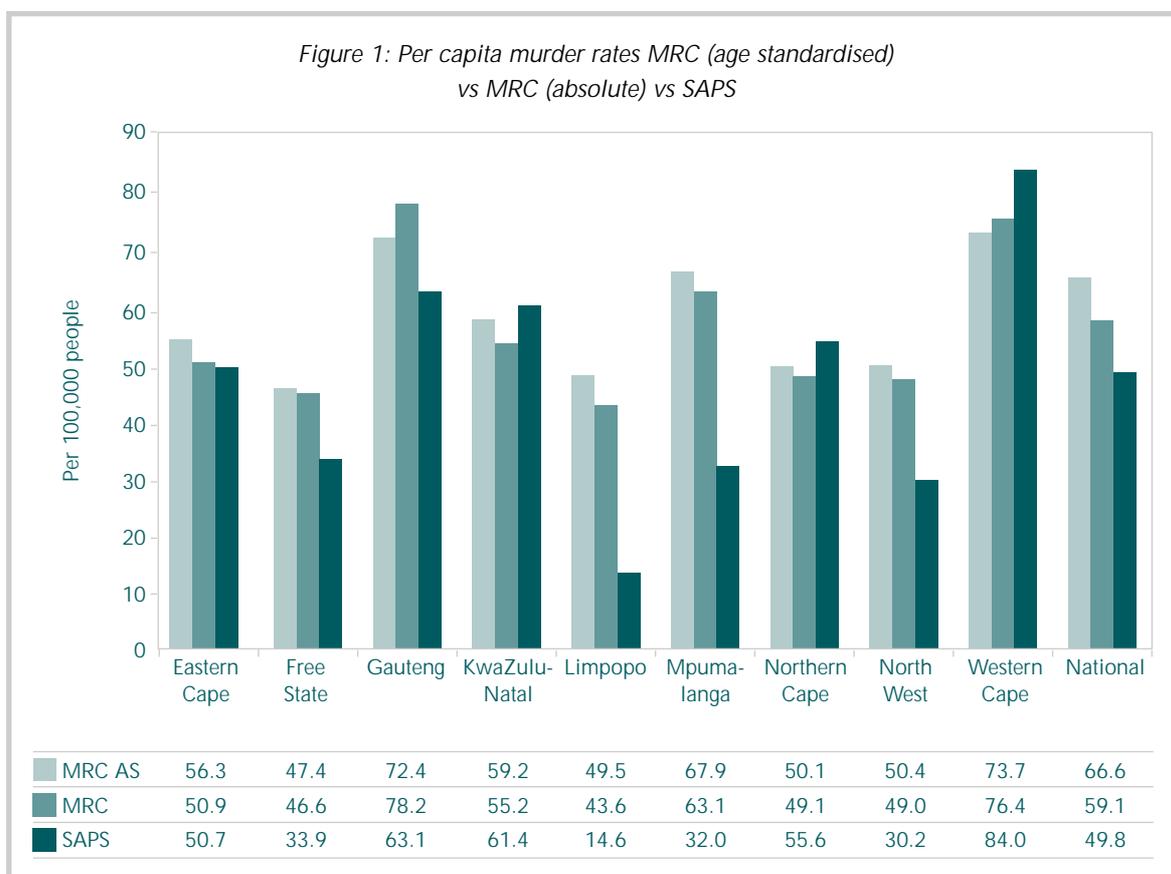


Figure 1: Per capita murder rates MRC (age standardised) vs MRC (absolute) vs SAPS

| | Eastern Cape | Free State | Gauteng | KwaZulu-Natal | Limpopo | Mpuma-langa | Northern Cape | North West | Western Cape | National |
|---|---|---|---|---|---|---|---|---|---|---|
| MRC AS | 56.3 | 47.4 | 72.4 | 59.2 | 49.5 | 67.9 | 50.1 | 50.4 | 73.7 | 66.6 |
| MRC | 50.9 | 46.6 | 78.2 | 55.2 | 43.6 | 63.1 | 49.1 | 49.0 | 76.4 | 59.1 |
| SAPS | 50.7 | 33.9 | 63.1 | 61.4 | 14.6 | 32.0 | 55.6 | 30.2 | 84.0 | 49.8 |

| Table 1: Comparative murder rates: MRC vs SAPS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eastern Cape | Free State | Gauteng | KwaZulu Natal | Limpopo | Mpuma-langa | Northern Cape | North West | Western Cape | National |
| MRC pop estimate | 6,897,865 | 2,862,088 | 8,765,262 | 9,211,922 | 5,277,432 | 3,054,973 | 955,010 | 3,753,128 | 4,399,414 | 45,177,094 |
| SAPS pop estimate | 6,846,154 | 2,787,611 | 7,871,632 | 8,982,085 | 5,500,000 | 3,040,625 | 872,302 | 3,566,225 | 4,192,857 | 43,659,491 |
| Per cap: MRC | 50.9 | 46.6 | 78.2 | 55.2 | 43.6 | 63.1 | 49.1 | 49.0 | 76.4 | 59.1 |
| **Per cap: MRC (Age Standardised)** | **56.3** | **47.4** | **72.4** | **59.2** | **49.5** | **67.9** | **50.1** | **50.4** | **73.7** | **66.6** |
| Per cap: SAPS | 50.7 | 33.9 | 63.1 | 61.4 | 14.6 | 32.0 | 55.6 | 30.2 | 84.0 | 49.8 |
| Difference (SAPS v MRC) | 0% | 37% | 24% | -10% | 199% | 97% | -12% | 62% | -9% | 19% |
| **Difference (SAPS v MRC AS)** | **11%** | **40%** | **15%** | **-4%** | **239%** | **112%** | **-10%** | **67%** | **-12%** | **34%** |
| MRC murders | 3,514 | 1,333 | 6,858 | 5,083 | 2,303 | 1,927 | 469 | 1,838 | 3,359 | 26,684 |
| **MRC murders (AS)** | **3,881** | **1,356** | **6,342** | **5,455** | **2,614** | **2,075** | **479** | **1,891** | **3,241** | **30,069** |
| No. murders SAPS | 3,471 | 945 | 4,967 | 5,515 | 803 | 973 | 485 | 1,077 | 3,522 | 21,758 |
| Difference (SAPS v MRC) | 1% | 41% | 38% | -8% | 187% | 98% | -3% | 71% | -5% | 23% |
| **Difference (SAPS v MRC AS)** | **12%** | **43%** | **28%** | **-1%** | **226%** | **113%** | **-1%** | **76%** | **-8%** | **38%** |

Conceivable, perhaps, but unlikely. If this difference were to account for the disparity, it would imply that the months April, May and June 2000 (which appear in the SAPS figures, but not in the MRC's) would have had unusually low murder rates, while the April, May and June 2001 rates (which appear in the MRC's figures, but not in the SAPS's) would have been unusually high. While we have no monthly data against which to test this possibility, it seems highly unlikely, since the SAPS records suggest that the number of murders fell in 2001/02 relative to 2000/01.

*Irreconcilable differences*
If this is not the reason for the disparity, there must be another explanation. One possibility is that the police are mistaken, that for reasons of inefficiency, or of inadequate systems, or of political expediency, they have failed to record all the murders committed in 2000/01. This cannot, of course, be dismissed as inconceivable, especially after the finding, reported in a separate MRC study into intimate femicide, that:

in 6.9% of probable homicides identified at mortuaries there was no police case number.

This conclusion was drawn after many months of exhaustive searching. There was thus no evidence of a police investigation. Attempts to find these numbers revealed that victims of homicide could not be traced via their names or ID numbers in the SAPS computerised database, even when these are known.[6]

If police error or inaccessibility accounted for their under-recording of murders, it might explain why the MRC estimate of murders in Limpopo is nearly three times higher than the number reported by the SAPS. It does not explain, however, why the MRC predicts neatly 40% more murders in Gauteng than SAPS reports, but 8% *fewer* in KwaZulu-Natal. This is the exact opposite of what would be expected if police systems were to blame for an undercount of murders.

Still, even if this were the case, it would only account for a portion of the difference between the MRC's projected figures and those of the SAPS. We must, therefore, explore the possibility that the MRC's approach has led to an overstatement of the

number of murders. This turns out to be a distinct possibility, and for two reasons:

- The first problem with the MRC's calculations probably led to an over-estimation of the number of people who died of non-natural causes.

- Within the category of non-natural deaths, the second problem may have led to an overestimation of the number of murders.

As described earlier, in calculating the number of non-natural deaths that had occurred, the MRC relied on an earlier study of 12% of all death certificates issued between 1997 and 2001. It concluded that 12% of those were for non-natural deaths. A more careful study of the report, however, shows that the 12% is an average for the period, but that the proportion of all deaths resulting from non-natural causes was falling quickly, having made up 16% of 1997 deaths and only 9% of 2001 deaths. In 2000, it made up 10%.[7]

Obviously, if the number of non-natural deaths was calculated at 10% rather than 12%, the figure would fall from 67,000 to 56,000. Since this is the basis against which the proportion of murders within the category of non-natural deaths (45%) was applied, this would result in reducing the estimated number of murders by nearly 5,000. This correction, by itself, may be sufficiently large to bring the MRC's predicted number of bodies down to the SAPS's figure of 21,758.

In addition to this, however, questions must also be raised about the MRC's direct application of the NIMSS findings about the causes of non-natural deaths to the subset of all non-natural deaths.

NIMSS is a mortuary-surveillance programme that tracks the number and cause of death of bodies arriving in morgues around the country. This sounds like a plausible source of data on non-natural deaths. The trouble with NIMSS, however, is that it is heavily biased towards urban areas. This is evident from the fact that 62% of all bodies surveyed by NIMSS in 2001, for instance, were presented at Gauteng and Western Cape mortuaries, despite the fact that only 38% of the population

lives in those heavily urbanised provinces. In addition, even in less urbanised provinces, the mortuaries accessed by NIMSS tend to be in urban areas.[8]

This matters because, despite the assurance offered by the MRC that there are similarities between the NIMSS results and observations made at two rural demographic monitoring projects with which they are associated,[9] there is wide consensus in academic literature that murder rates in rural areas are lower than those of urban areas. Indeed, this is apparent in the SAPS statistics, where the murder rate in Limpopo is only about one-third that of the rest of the country. Because the MRC imposes a figure generated by a sample with a strong urban bias, however, their estimates of the number of murders in Limpopo is nearly three times that of the SAPS.

All things considered then, it is hard not to conclude that despite the genuine efforts of the MRC to calculate the murder rate off other data (the number of people who are thought to have died, the proportion of those who die from non-natural causes, and the proportion of non-natural deaths that are homicides), the result is so much greater than the SAPS reported figures, that questions must arise as to its validity. It would seem reasonable, therefore, to continue to rely on SAPS figures unless and until those can be shown to be erroneous.

### Case two: murder rates in the 'Coloured' community

Last year, the *SA Crime Quarterly* published two articles that suggested that the homicide rate in the Coloured community was significantly higher than that of the rest of the country.[10] The problem with both these pieces is that for the years after 1990, they are premised on the NIMSS data regarding the race of the victims of murderous violence.

Leggett, after citing Thomson's data for 2003, summarises the premise of both pieces, writing that "figures from the National Injury Mortality Surveillance System (NIMSS) … show Coloureds to be far more vulnerable. In both 2001 and 2002, the NIMSS recorded a disproportionately large number of Coloured homicides in the total reviewed: 14%

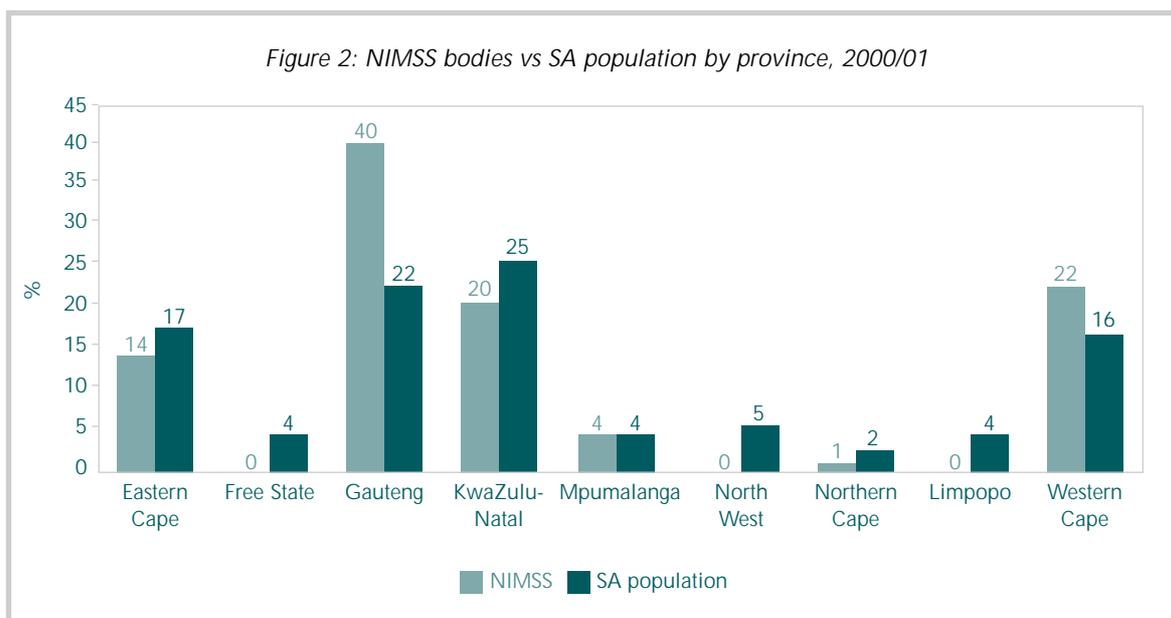in 2001 and 13% in 2002, compared to the 9% share held by Coloureds in the national population."

The trouble with this argument, however, is that NIMSS reports only raw data. It does not seek in any way to extrapolate from the data collected at its 30-odd mortuaries to the population as a whole. Thus, the only way in which the racial breakdown of victims in the NIMSS sample might correspond to that of the country as a whole would be if the catchment areas for the mortuaries participating in NIMSS were representative of the country as a whole. Unfortunately, this is very far from the case. In fact, the NIMSS data, as already pointed out are biased towards urban areas (Figure 2).

In addition, and more importantly with respect to the question of the murder rate in the Coloured community, it is also biased towards areas where Coloured people live. This is partly an effect of the urban bias, since Coloured people tend to be more urbanised than the rest of the South African population, but it is also an effect of the fact that the urban areas that dominate the NIMSS sample are also those with a large Coloured population.

Partial evidence of the effect of this distortion is revealed by calculating the number of Coloured victims one would have expected to find in the NIMSS sample, by taking the number of homicides in the provinces in which NIMSS mortuaries exist and multiplying those by the proportion of the population made up by the Coloured community. This would imply that the murder rate in that community was precisely that of the rest of the population and it would help set a par value above which we might say that Coloured people are, indeed, over-represented in the NIMSS sample. In fact, if we do this for the NIMSS sample for 2001 we get an expected number of Coloured victims of 1,684. NIMSS, however, found only 1,551 Coloured victims. Coloured people were, if anything, under-represented.

However, this test is only partial: because NIMSS has an urban bias, the proportion of the *provincial* population that is Coloured should not be used to calculate this par value. To be more accurate, it is necessary to look at the proportion of the population made up by Coloured people in the catchment area of the mortuaries concerned. For a number of reasons, this is not possible. Still, in the absence of this, it is impossible to conclude on the basis of the NIMSS data that the murder rate in the Coloured community is higher than that of the rest of the country. Indeed, when we set par values for all South African race groups, it turns out that the NIMSS sample suggests an over-representation of African victims and under-representation of all other groups (Figure 3).



Figure 2: NIMSS bodies vs SA population by province, 2000/01

Figure 3: Par values for population representation vs NIMSS sample, 2001
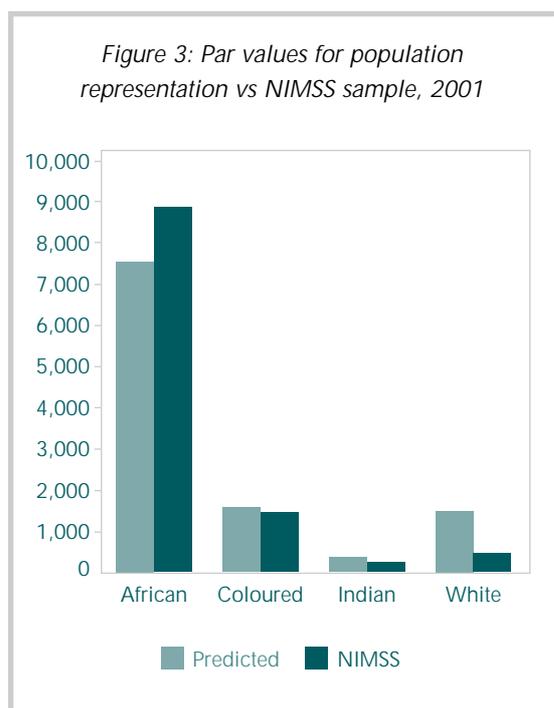
As has already been suggested, this is *not* to say that the murder rate among Africans is significantly higher than the national average or that the opposite is the case for other groups. It is to suggest very strongly, however, that it is impossible to establish how risk is distributed among population groups merely on the basis of NIMSS. To do so would require far more information about the demographics of the catchment areas for the mortuaries covered by NIMSS.

## Conclusion

This article has sought to show how the failure to pay sufficient respect to the limitations of data, however seemingly solid, can result in quite serious misjudgements about the level of crime and, indeed, the distribution of risk.

It offers no answers to the questions of how much murder there really is or whether some communities are more at risk than others. All it offers is the suggestion that, in the absence of more compelling data, we ought to accept police statistics as reflective of reality and that NIMSS data cannot be used to estimate the burden of risk without much more data about the population from which its samples are drawn.

**Endnotes**

1  D Bradshaw, N Nannan, R Laubscher, P Groenewald, J Joubert, B Nojilana, R Norman, D Pieterse and M Schneider, *South African National Burden of Disease 2000*, South African Medical Research Council, Cape Town, 2004.

2  These figures are based on a spreadsheet provided by the MRC on their website at <http://www.mrc.ac.za/bod/Age%20standardised%20rates.xls>. They do not precisely match the figures provided later in this article because the rates here have been standardised to the age structure of the South African population.  In essence, that process makes our death rates comparable with those of other countries whose populations' age structures differ from our own. Thus, because ours is a relatively young population, causes of death that disproportionately affect the young are adjusted relative to the absolute number of such deaths estimated by the model. In the rest of this paper, as far as possible absolute numbers and rates are used, rather than these age-adjusted rates. Unfortunately, these absolute figures are not provided in all cases and some have been made available only through personal communication.

3  D Bradshaw, P Groenewald, R Laubscher, N Nannan, B Nojilana, R Norman, D Pieterse and M Schneider, *Initial Burden of Disease Estimates for South Africa, 2000*, South African Medical Research Council, Cape Town, 2003.

4  D Bradshaw, personal communication, August 2005.

5  The reason for the difference between SAPS estimates and MRC estimates, depending on whether absolute or per capita rates are used, is because the SAPS uses a slightly lower population estimate than does the MRC. This has the effect of making the SAPS's per capita higher than it would be if it used the same population number as does the MRC.

6   S Mathews, N Abrahams, LJ Martin, L Vetten, L van der Merwe and R Jewkes, "*Every six hours a woman is killed by her intimate partner": A national study of female homicide in South Africa*, South African Medical Research Council, Cape Town, 2005.

7   D Bradshaw, et al. 2003. Op cit.

8   R Matzopolous, *A profile of fatal injuries in South Africa 2001: Third annual report of the National Injury Mortality Surveillance System*, South African Medical Research Council, Cape Town, 2002.

9   Bradshaw, et al, 2004, p 8.

10  JDS Thomson, A murderous legacy: Coloured homicide trends in South Africa, in *SA Crime Quarterly 7*, ISS, Pretoria, 2004 and T Leggett, Still Marginal: Crime in the Coloured community, in *SA Crime Quarterly 7*, ISS, Pretoria, 2004.