INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa • University of Ottawa

REPORTING COMPLETENESS

OF THE

AIDS CASE REPORTING SURVEILLANCE SYSTEM

by

Jeffrey J. Whitehead M.D.

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfilment of the requirements for the
MSc degree in Epidemiology

University of Ottawa

December 1996

ABSTRACT

The Problem: The AIDS Case Reporting Surveillance System (ACRSS) is a passive surveillance system that was initiated in 1982. It is used to follow the health impact of HIV in the population at large and to determine HIV prevalence through back-calculation. An assessment of the proportion of all AIDS cases reported out of all AIDS cases diagnosed (ie reporting completeness) was required as part of the evaluation of this surveillance system.

Methods: A number of secondary data sources were considered; the use of death certificate data from the Canadian Mortality Data Base (CMDB) was chosen by a process of elimination. The limited number of overlapping personal identifiers, such as the absence of names on all and the absence of initials (in Quebec) on some ACRSS files, and the size of the mortality database necessitated the use of computerized record linkage. ACRSS data up to September 1994 and CMDB data up to the end of 1992 were utilized for this linkage. As the CMDB also does not have complete ascertainment of all AIDS cases, capture-recapture methods were used to calculate reporting completeness. Reporting completeness was calculated with respect to all reports, by gender, by province of residence, by year of death, and by size of community at death. Comparison of reporting completeness of different cohorts of certified AIDS deaths over time allowed assessment of underreporting by controlling for reporting delay. One, two, three, and five year non-reports rates were calculated to allow comparison with results from the U.K.

Results: Reporting completeness of ACRSS was about 85% (sensitivity analysis placed

it between 84.1% and 86.5%). Reporting completeness was higher for males (86%) than for females (76%), this gender difference is more marked outside the province of Quebec. Ontario and Alberta had the highest rates of reporting completeness while Quebec, P.E.I., and New Brunswick had the lowest. Decreases in reporting completeness in the most recent years may be due to reporting delay. Reporting completeness was lower in those residing in smaller communities at the time of death. Stratification by residence in province of Quebec or in the rest of Canada indicated that failure to report in Quebec is a problem independent of failure to link due to missing initials in ACRSS since November 1990. Cumulative death cohort analysis indicated an increase in underreporting at 5 years by 0.4 to 0.7% per year. One, two, three, and five year non-report rates were approximately 28%, 15%, 8%, and 5% in the late 1980s. The one year rates are in the same range as those found in the U.K.

Conclusion: Reporting completeness to the AIDS Case Reporting Surveillance System is about 85% which compares well with that seen in other countries. Reporting completeness is lower for females than males. Reporting completeness varies by province with Quebec, P.E.I., and New Brunswick having the lowest rates of reporting. Alberta and Ontario have the highest rates of reporting. Reporting is less complete in those that die in smaller communities. Reporting completeness decreased slightly during the 1980s. The most important finding is the trend to increased underreporting with time. Future validations of reporting completeness will become more difficult in Quebec due to the lack of initials on their dataset. The causes of incomplete reporting will only be elucidated with qualitative investigation.

Acknowledgments

I would like to thank my thesis supervisors Dr. Rama Nair of the University of Ottawa and Dr. Maura Ricketts of Health Canada for their assistance with this project. The help from Pierre Lalonde and Martha Fair of the Occupational and Environmental Health Research Section of Statistics Canada is also appreciated.

Table of Contents

List of Tables

List of Figures and Charts

1. INTRODUCTION

1.1. AIDS in Canada

Acquired immunodeficiency syndrome (AIDS) was first recognized following a cluster of cases of Pneumocystis carinii pneumonia (PCP) in gay men in the U.S. in 1981. The first case in Canada was reported in February 1982[1]. Further investigation led to the understanding that this disease was due to a transmissible agent, now known as the human immunodeficiency virus (HIV). This viral infection most commonly presents as an opportunistic infection or a neoplasm due to an increasing impairment of the immune system following a latent period of variable length. The terminal symptomatic period of HIV infection is known as AIDS.

The rate of increase of AIDS was most rapid in the early 1980s. This increase slowed in the latter part of the past decade but delays in reporting of cases tend to underestimate the true prevalence. In 1995 it was estimated that there were approximately 45,000 people with HIV infection in Canada[2]. As of April 1995 11,192 AIDS cases had been reported in Canada of which 7,880 were known to be deceased. Annual deaths from AIDS exceeded 1,300 in 1992. Four provinces have accounted for 95% of all AIDS cases: Ontario, Quebec, B.C., and Alberta. Mortality rates are high for men in these four provinces compared to the Canadian average. The rate for women is elevated only in Quebec compared to all others.

Risk groups for reported adult AIDS cases have obvious gender differences. In males the largest risk category is men who have sex with men (MSM) which includes homosexuals and bisexuals. This group accounts for 76% of all male AIDS cases. This risk factor partially explains the male:female disparity with the disease with 94% of cases occurring in males. Heterosexual contact is responsible for 9% of cases in males, 5% being due to sexual contact with a person at risk and 4% being due to origin in a Pattern II country. A Pattern II country is one in which heterosexual transmission is the predominant means of transmission. IDUs (intravenous drug users) account for 4% of male AIDS cases while IDUs who also belong to the MSM risk group account for 4%. In females heterosexual contact is responsible for 63% of all cases, IDU for 15%.

Most adult AIDS cases are in the 30 to 39 year old age range followed by the 40 to 49 year old in males and the 20 to 29 year old in females. About half of all AIDS cases present with PCP; KS and candidiasis are the next most common presenting illnesses although KS is rarely seen in women. Only 116 pediatric AIDS cases (<15 years of age) had been reported as of April 1995, 71% of these being due to perinatal transmission. Pediatric cases make up 1.1% of all reported AIDS cases in Canada.

1.2. AIDS Case Reporting Surveillance System

The AIDS Case Reporting Surveillance System (ACRSS) was initiated in 1982. AIDS became a reportable disease in all provinces and territories between 1983 and 1988.[3] HIV, in the absence of an AIDS defining illness, is not a reportable disease in British Columbia, Alberta, or Quebec. The national AIDS surveillance system involves the passive collection of provincial and territorial AIDS data, analysis of this data, and dissemination of the resulting information through quarterly reports. The primary analysis describes the epidemiology of AIDS (demography, risk factors for HIV infection, patterns of AIDS defining illnesses, survival trends, and temporal trends).

The health information derived from the ACRSS database has many uses. It is used to follow the health impact of HIV infection in the population at large by enumerating the most severe and definitive stage of HIV infection. Back-calculation to determine HIV prevalence complements HIV sero-prevalence surveys. This information is also useful for the development of policies on disease prevention and utilization of health care resources.

The ACRSS report form contains over 120 variables. These include patients' initials, date of birth, vital status, country of birth, date of death, and location at the time of diagnosis. Patients' initials have not been provided by the province of Quebec since November 1990. To be eligible for the database, a person

must have at least one of the AIDS defining illnesses.[4] To confirm eligibility for the database, these are listed as are the diagnostic methods. There is a section on risk behaviours associated with HIV transmission in an attempt to identify the mode of transmission. Finally, there are items related to laboratory data, TB and AIDS co-infection, and identification of the person completing the form (in case clarification is needed).

Data are collected by attending physicians and local health departments, then forwarded to provincial/territorial health authorities. Reports may be either an initial report or an update, the latter most commonly at the time of death. Note that this is a passive system of surveillance. Provinces other than Quebec or Ontario forward the data in hard copy to the Bureau of HIV/AIDS and STDs at LCDC, the latter two send in data by modem. The master AIDS database is then converted to SAS to allow analysis and dissemination of information. A Quarterly Surveillance Update is mailed to interested individuals and organizations every 3 months. It is also available through fax downloads (Faxlink) and will be posted on the LCDC, Health Canada Web site.

As with any surveillance system, it is important to maintain the quality of the data. Some variable fields are poorly completed but AIDS defining illness, diagnostic method, provincial ID number, and source are mandatory. Completing missing fields is mainly done at the local level. It is difficult to check

data validity; chart reviews are not routinely performed. Duplicate reports, due to interprovincial migration and duplicate reporting are minimized by comparing initials, date of birth, and sex of all new cases with existing files at the national and local level. Statistics Canada has also helped to clean out duplicates by doing one-file internal linkages.

Since the first report of PCP in gay men in 1981 there have been changes in the case definition reflecting improved understanding of the disease. There have been four different case definitions; starting in 1982, and refined in 1985 and 1987. The most recent revision was in 1993 when pulmonary tuberculosis, invasive cervical cancer, and recurrent bacterial pneumonia were added as AIDS-defining diagnoses. Changes in the reporting criteria have been historically demonstrated to increase the number of AIDS-defining illnesses. The changes in 1987 increased the number of reported cases by 14.8% in Canada and 30.5% in the U.S.[5] There has been no systematic attempt to identify cases retrospectively. Specifically, a search for AIDS cases that did not meet the surveillance case definition in the past has not been done.

1.3. Evaluation of Reporting Completeness to ACRSS

Data validation of a surveillance system involves assessment of reporting completeness, internal data validation, and external data validation. As a part of the evaluation of this surveillance system, it was determined that an estimate of the proportion of all AIDS cases reported would be calculated for Canada. To perform this calculation the three following types of failure to report need to be considered:

1. All AIDS cases known to be AIDS but which will never be reported to ACRSS. This is known as underreporting.
2. All AIDS cases known to be AIDS but not yet reported to ACRSS but will eventually be reported. This is known as reporting delay.
3. All AIDS cases not known to be AIDS (dead or alive). This is known as underdiagnosis.

Another term used is "reporting completeness" which most often refers to underreporting and reporting delay combined but often refers to the former alone. The lack of a standard definition of reporting delay/reporting incompleteness confuses the issue further.

At any point in time it is not possible to tell if an unreported case is delayed or withheld (underreported). There is no agreement concerning the period of time beyond which reporting delay becomes reporting incompleteness. Clearly it should not be infinite as the value of the information declines steadily with time

regardless of the role of the surveillance system.

This problem has been somewhat clarified by Evans who divided failure to report into delayed reporting and non-reporting[6]. The definitions proposed by Evans have been used in at least two studies in the U.K. These studies have been valuable because they made practical comparisons of reporting completeness possible.[7,8] In the absence of a standard definition, one is faced with uninterpretable proportions. For example, a finding of 10% reporting incompleteness does not clarify how long the system was waiting for the case reports: 6 days, 6 weeks, 6 months, 6 years? For each time frame the implications are different.

There are two other reasons that AIDS cases will not appear on a national registry which are not insignificant but will usually not be included in an assessment of failure to report: the underdiagnosis (or underrecognition) of AIDS as mentioned earlier and severe HIV disease that does not meet the case definition. Failure to make the diagnosis of AIDS has been recognized in autopsy, pathology and chart review studies.[9] This is not unexpected as AIDS in Canada is still a relatively rare disease, particularly in some geographic areas and among some populations, such as women. Because cases of AIDS which are not diagnosed will not appear on other data sources in general, this cause of failure to report cannot be included in most validations of AIDS case surveillance

systems. Secondly, almost 6% of persons with HIV infection will never develop an AIDS-defining illness but will die, nonetheless, of another illness secondary to their immunosuppression.[10] This has become less of a problem as the case definition has been broadened over the years. However, for calculation of reporting completeness, this may be a problem when the secondary data source identifies as "AIDS" a case of severe HIV associated illness which does not meet the case definition for AIDS. For example, if mortality records are used as the alternate data source, some deaths will be classified as HIV/AIDS but they will not appear on the AIDS register. An apparent increase in incomplete reporting will occur when the two records are compared if the mortality record includes as "AIDS" illnesses which do not meet the case definition. The source of this error is the use of two different definitions of severe HIV associated illness.

With these definitions in mind, it should be noted that there are a number of possible reasons for underreporting. These include:

1. There is concern among physicians about the confidentiality and privacy of data held by public health surveillance systems. Since AIDS cases are reported to the same agencies which are responsible for contact tracing (which many patients want to avoid), there is considerable resistance to reporting among patients, if not physicians. Cases not reported until after death may reflect this concern.
2. Physicians may dislike the administrative bother of filling out a reporting

form.

3. There is unawareness on the part of some physicians of the need to notify, unclear communication channels between physicians and health departments, inadequate feedback to physicians by health departments, and practical problems such as delays in providing reporting forms.[11]
4. A U.S. study looking at AIDS reporting under ideal circumstances of well motivated physicians showed a 9.5% underreporting rate, virtually all of it inadvertent.[12]

Underreporting rates vary widely by geographic area. In addition, it appears that active surveillance systems have much better compliance with reporting than passive systems such as we have in Canada. Underreporting is typically at 10 to 15% in areas with active surveillance and roughly twice that level in locations with passive surveillance.[13]

2. LITERATURE REVIEW

MEDLINE was searched from 1982 to March 1996. Search terms used were (report:) as a text word and "AIDS" and "population surveillance" as MeSH terms. This was supplemented by a manual review of the Journal of AIDS, AIDS, and the American Journal of Public Health from April 1995 to February 1996. Further references were obtained with the help of the professional librarians at the LCDC library by a search of AIDSLINE. The references of articles obtained were reviewed to obtain any studies missed in the above search strategies. Additional references were found in the files provided by Dr. Maura Ricketts.

Review of Previous Studies of AIDS Underreporting

Underreporting of AIDS cases to surveillance systems is a significant problem worldwide. European estimates in 1988 varied from 0 to 20%.[14] In the U.S., which has an active surveillance system, the proportion unreported for the same year was estimated to be 8%.[15] The published studies investigating underreporting to ACRSS in Canada are discussed below. There have also been calculations of reporting delay and reporting incompleteness done by the Division of HIV/AIDS at LCDC.

A CMAJ article in 1988 by Losos et al noted that the reporting delay followed a right skewed distribution with a median reporting delay of 1.6 months in 1985-87.[16] No estimation of reporting completeness was made. A Royal Society of

Canada paper by Wells et al in the same year noted that the reporting delay curve in Canada was skewed to the right with only 3 to 5% of all cases being reported after a delay of over 14 months for the years 1984-87.[17] No correlation was found between the date of the report and duration of reporting delay.

Johnson et al in a 1989 article used death certificates to investigate reporting completeness.[18] These authors looked at a 3 year period of time from 1985 to 1987 at which time 40% of all reported AIDS cases in Canada were notified from Ontario. They selected all deaths for the same period of time coded specifically for AIDS as an immediate or underlying cause; case-defining illnesses were not included. It appears that no time lag was allowed for death to occur from the time of diagnosis nor was there an allowance for administrative delays. Manual matching was done using first and last initials, date of birth, sex, and the regional municipality or county of residence. Completeness was calculated by dividing the number of deceased reported cases of AIDS by the total number of death certificate cases for each of the three years. Death certificate cases were only matched to those listed as dead in the AIDS reporting system. Overall completeness of AIDS reporting was 75.2%. PCP, pneumonia unspecified, and KS were the most common causes of death. The authors noted that this was only an estimate of completeness of AIDS reporting because cases who died in a certain year may have been reported in a previous year. On that basis, the calculated reporting completeness seems surprisingly high.

A 1990 study by Calzavara et al looked at the reporting completeness of AIDS among members of two cohorts in Canada.[3] These were the Toronto Sexual Contact Study (TSCS) in Ontario and the Vancouver Lymphadenopathy-AIDS Study (VLAS) in B.C. Both study groups sent initials, date of birth, and city of residence of patients with AIDS and others without AIDS to the Federal Centre for AIDS (now known as the Division of HIV/AIDS Epidemiology within LCDC). The Federal Centre for AIDS (FCA) was not aware of the clinical status of these people. All cases diagnosed in the three months before the date of the linkage were excluded to allow for delay in reporting, this time period being consistent with the median time lag in reporting at that time. VLAS sent information on 961 people, 382 of whom were seronegative. TSCS sent information on 210 people, 29 of whom were seropositive without AIDS. A manual record linkage was performed. No linkages were made to the 382 seronegative subjects in the VLAS or to the 29 seropositive subjects without AIDS in the TSCS. This demonstrated the accuracy of record linkage with only these three fields (initials, date of birth, and city of residence). The proportion of AIDS not reported was 12% for VLAS and 18% for TSCS. It was noted that underreporting from the TSCS increased from 0% in 1983-84 to 44% in 1987-88 while underreporting remained stable from the VLAS. The authors did not feel that reporting delay was a factor in this increase in underreporting as the most recently reported case to FCA had been diagnosed 10 months before the linkage was done. It is not clear to me how this precludes this possibility however.

Remis and Palmer in 1991 calculated survival among adult AIDS cases after diagnosis using Kaplan-Meier techniques.[19] Using the survival distribution stratified by time of diagnosis to AIDS case reports they estimated AIDS mortality after correcting for reporting delays. These results were compared to death certificate based mortality statistics for 1987 and 1988. The authors concluded that AIDS case reporting was relatively complete in Quebec. As the analysis was based on aggregate data with no case by case matching, the possibility of error in this study seems quite high as the extent of overlap is unknown. It is not known if those that died based on death certificates were those who were modelled to die based on estimates of reporting delay and survival.

In a 1992 study by Ricketts, all death certificates coded as AIDS (279.1, 042, 043, 044) were pulled and matched manually using sex, initials, date of birth, and date of death.[20] Sex had to match before records were compared (blocked on sex). An exact match occurred if the initials, date of birth, and date of death were identical. A "probable" match occurred if initials, month and year of birth, and date of death agreed or if last initial and date of birth and date of death agreed. Geography was not used to link. Reporting completeness was calculated to be 86% using capture-recapture calculations. There were definite limitations to this study. The extent of coding of AIDS deaths under other codes was unknown and geographical information was not used to link. More

significantly, Statistics Canada would not allow the collection of any demographic information from death certificates that did not link to the ACRSS file. Therefore, reporting completeness by gender, by province, by year of death, and by community size at death could not be calculated.

Calculations of reporting delay have been done on a regular basis by the Division of HIV/AIDS at LCDC.[21] For ACRSS, the median delay for the reporting of cases is approximately 9 months with a right skewed distribution. Less than 60% of cases are reported within one year. This analysis was based on the 9,894 adult AIDS cases reported to LCDC by June 30, 1994. The situation is quite different in the U.S. where one study showed 82% reporting within 5 months to the state level.[15]

Le et al in 1994 reported on a manual record linkage between the BC AIDS Registry and 1000 members of the Vancouver Lymphadenopathy-AIDS Study cohort.[22] Underreporting rates were 10, 5, 14, 12, 12, and 16% for the years 1987 to 1992 respectively (12% overall). No significant trend was seen with date of diagnosis nor were there differences between reported and unreported cases in terms of AIDS-defining illness or socio-demographic characteristics. The authors did not comment on the possibility that reporting delay may have increased the latter year results. They suggested that underreporting was probably higher elsewhere as this cohort of patients is looked after by a group of

family practitioners with considerable experience with AIDS.

An enumeration of known HIV positive women was conducted between September 1993 and May 1994 by Calzavera et al through contact with physicians, HIV clinics, and community organizations.[23] In February 1995 a computerized record linkage was done between ACRSS and the 183 of the 687 HIV positive women who had AIDS-defining illnesses. Fewer than 50% of the women were found in ACRSS. As median reporting delay is 9 months, it would appear that much of this underreporting was due to reporting delay. Another limitation in this study was the lack of validation of the AIDS diagnosis in this population.

In summary, there have been a number of regional and cohort studies of AIDS reporting completeness in Canada. The results of these studies cannot be generalized to the entire country. Some of these studies in retrospect have significant design flaws which would be expected in a rapidly evolving area such as this. The only national picture in the past was provided by the Ricketts study in 1992. This study was limited by its restriction to causes of death that were coded as AIDS and by the inability to calculate reporting completeness by demographic variables such as gender or province of residence.

3. AIMS AND OBJECTIVES

The aim of this study was to provide a comprehensive validation of reporting completeness on a national basis of the AIDS Case Reporting Surveillance System using an external data source.

The objectives of this study were as follows:

1. To determine the level of reporting completeness to ACRSS nationally, by gender, by province of residence, by year of death, and by community size at death.
2. To estimate the level of underreporting using cumulative death cohorts.
3. To estimate one, two, three, and five year non-report rates.

4. METHODS

4.1 Definitions

For this paper, the following definitions will be used:

1. Underreporting will be defined as the failure to ever report a diagnosed case of AIDS.
2. Reporting delay is defined as the time period between the diagnosis of AIDS and the receipt of case information by the surveillance system.
3. Reporting completeness will refer to the proportion of AIDS cases reported out of all AIDS cases diagnosed. Those cases not reported will include those that are due to underreporting and reporting delay. It will not include cases of underdiagnosis.

I suggest that the expected period for reporting, describing the most reasonable time before "delay" can be said to occur, should be based on three factors:

1. The administrative time required for the information to pass from physician to the federal level of government.
2. The time at which the overall purpose of the surveillance system is compromised.
3. The time at which further reporting is unlikely.

These three factors will be discussed below.

1. Evans argues that the administrative time required (ie doctor's office to

Health Canada) to obtain AIDS reports should not take more than three months. The fact that a minority of cases of AIDS are now reported within a three month period may be due to the unique sociopolitical nature of AIDS (compared to all the other reportable diseases). Specifically, public knowledge of a person having AIDS can have a much more severe impact on that person than public knowledge of a person's Salmonella infection. Although the level of reporting of other communicable diseases tends to be lower than for AIDS, reporting of other diseases is relatively prompt. Therefore, one can argue that three months should be the initial limit after which a case can be said to be "reporting delayed". While Evans refers to reporting before the three month limit as "prompt reporting", the use of this term has not been found elsewhere in the literature.

2. As noted previously, one of the purposes of the ACRSS database is to follow trends in HIV infection in the population at large to assist in the development of policies on disease prevention and utilization of health care resources. AIDS data are particularly useful because seroprevalence surveys do not cover the entire population and AIDS data tends to be less biased. A disadvantage is that the median incubation period between HIV seroconversion and the development of AIDS is over 10 years.[24] Reporting delay adds to this limitation of AIDS surveillance. For example, detection of an increased occurrence of AIDS in a

population subgroup is much less useful after 5 years than one year and is probably of only historical importance after 10 years. Evans suggests that delayed reporting be defined as case reports which arrive between three and 12 months after diagnosis. Evans goes on to define non-reporting as all cases that are reported more than 12 months after diagnosis. Although cases reported more than one year after diagnosis are less useful for the purpose of surveillance, the value of a reported case does not drop precipitously to zero. The value gradually declines with time. Therefore, I suggest two, three, five, and 10 year non-reporting rates would be good indicators of the efficiency of the surveillance system.

3. It appears that many AIDS cases are not reported until after death. A 10 year cutpoint would be useful as reporting after that time period would not be likely as survival times of over 10 years are very unusual. As a result, the 10 year non-reporting rate may be the closest approximation to the ill-defined reporting incompleteness (ie underreporting) found in some studies (underreporting being a diagnosis that can only be made in retrospect).

4.2. Validation by a Secondary Data Source

Evaluation of underreporting in a surveillance system requires the presence of an independent secondary data source. This can be difficult for active surveillance systems as multiple sources of information are often used in the compilation and independent sources may not exist. For passive surveillance systems such as ACRSS, independence is less of a problem. Possible alternate sources of information are listed below:

1. Registries of reportable diseases that are AIDS-defining illnesses
2. Cancer registries for neoplasms associated with AIDS such as Kaposi's sarcoma and non-Hodgkins's lymphoma
3. Hospital discharge data
4. Health insurance databases
5. AIDS support group records
6. Physician charts and clinic records
7. Vital statistics registries (mortality database)

The limitations and advantages of each will be discussed below.

1. Registries of reportable diseases that are AIDS-defining illnesses

   Few reportable diseases are AIDS-defining illnesses. An exception is tuberculosis; however only 4.8% of all AIDS cases in Canada are reported to have co-infection with TB.[25] More common infectious diseases associated with AIDS such as Pneumocystis carinii pneumonia (PCP) infections are not reportable diseases, making these unsuitable

sources of data.

2. Cancer registries for neoplasms associated with AIDS such as Kaposi's sarcoma and non-Hodgkins's lymphoma

Certain neoplastic conditions in combination with HIV infection are diagnostic of AIDS and should be reported to provincial cancer registries. The two most common neoplasms in AIDS are Kaposi's sarcoma (KS) and non-Hodgkin's lymphoma (NHL). Kaposi's sarcoma is becoming less common as a complication of AIDS. However, it is listed as a risk factor in 21% of AIDS cases overall and these are clustered in men who have sex with men. The majority of NHL reports are not associated with AIDS and only 5% of people with AIDS in Canada have this disease listed as an AIDS defining illness.[21]

3. Hospital discharge data

Hospital discharge data are reported to the Canadian Institute for Health Information (CIHI) in an aggregate form. Each report is a hospital separation and may include diagnoses but will not include personal identifiers which would allow one to count cases. The same person can be admitted to hospital several times; each interaction with the hospital will be counted as a separation. This will be recorded similarly to several persons admitted once to hospital with AIDS. The reason for admission can also be obscured. Finally, not all patients with AIDS are admitted to hospital, hence there will be no record. The increased management of

AIDS patients on an out-patient basis may limit the usefulness of this data source.[26]

4. Health insurance databases

Provincial health insurance databases could potentially provide ideal alternate data sources as they would capture both inpatient and outpatient activity. Unfortunately, the quality of diagnostic information is notably poor and accessibility is very restricted in general.

5. AIDS support group records

Another possible source of information would be through community based organizations such as AIDS support groups. These are present in large urban centres and provide assistance to those with HIV infection and AIDS. The disadvantage is that these centres often serve cause-specific groups such as those with transfusion-related disease or members of the gay population. There is also considerable resistance to cooperate with health authorities from many of these groups, particularly in providing personal identifiers. Case definition is also a problem as these groups do not verify the claimed health status of their clientele.

6. Physician charts and clinic records

Comparison with physician and clinic records could be done on an individual basis but with an estimated 16,000 AIDS cases diagnosed in Canada to date, this would be very time consuming. Most physician and clinic records in this country are not computerized other than for billing

purposes. Local clinical databases such as the Vancouver Lymphadenopathy-AIDS Study cohort in B.C. can estimate reporting completeness regionally but the results will not be generalizable to Canada.

7. Vital statistics registries (mortality database)

Mortality databases in developed countries are generally well established databases containing death certificates for all deceased residents. As this database was chosen through a process of exclusion, the advantages and disadvantages of using a mortality database is discussed in the next section.

In summary, seven alternative secondary data sources have been considered for validation of the AIDS Case Reporting Surveillance System. The disadvantages of these alternate data sources include incomplete listing of the majority of people with AIDS, poor quality information (eg. bias in collection), and limited access to the information. As a result, death certificates were chosen to validate the reporting completeness of ACRSS. The advantages and limitations of the use of death certificates are discussed in detail in the following section.

4.3. Validation by the Canadian Mortality Database

The main advantage of this type of database is that all persons with AIDS die with a median survival of less than 2 years, hence are quickly captured in the mortality database. The CMDB has no significant problem with reporting delay or completeness for deaths occurring in Canada. Cases of death are generally always reported within three months. A two year delay is required to allow for electronic data entering. Therefore, the most recent available CMDB file was from 1992 for this study. Although mortality databases are believed to capture virtually all deaths that occur within a country, they do have limitations which will be discussed below:

1. *Not all persons with HIV or AIDS die from AIDS.* Intercedent events such as trauma and other diseases can be expected to lead to death. Accidental cause of death should not be uncommon as expected in this younger age group where HIV infection is more prevalent. High rates of suicide were found in one U.S. study[27], another found that suicide was the cause of death as certified by death certificate in only 0.3% of AIDS patients on a register.[28] Drug abuse was given as the cause of death in 1.3% of all cases of AIDS and total deaths due to causes unrelated to HIV were less than 3% overall in the same study.
2. *Not all deaths from AIDS are coded as AIDS on the death certificate.* The cause of death given may be one of the opportunistic infections or neoplasms which occur secondary to AIDS without reference to the HIV

infection.

3. *AIDS may be given as an antecedent or associated cause.* In Canada only the underlying cause of death is entered into the computer database. The immediate cause of death is listed in Part I on the death certificate followed by conditions which preceded and contributed to the immediate cause of death. These are known as antecedent causes. The underlying cause of death should be listed last. Nosologists follow rules established by the WHO published in the ICD-9 ACME decision tables produced by the National Centre for Health Statistics[29]. The antecedent cause will be given as the underlying cause of death provided that it is listed in the causal pathway leading to the immediate cause of death. If an error in coding is made, the antecedent cause may not be listed as the underlying cause of death and will not appear in the electronic format. HIV/AIDS may be listed in Part II of the death certificate for associated conditions which refers to conditions which contribute to the death but are not causally related to the immediate cause of death. In this event, HIV/AIDS will not be listed as the underlying cause of death.

4. *Physicians may purposefully complete certificates inaccurately.* While death certificates are not public documents in Canada and are therefore not accessible without appropriate authorization, there may nonetheless be reluctance to report diseases with social stigma. The death certificate is seen by a select group of people in any community and there may be

reluctance to allow access to even this group. This reluctance to certify AIDS deaths may be extended to a reluctance to report AIDS cases to surveillance systems. The literature is contradictory and scarce. One U.S. study showed that those with HIV/AIDS noted on their death certificate were also more likely to be in the AIDS registry.[30] However, a study from the UK demonstrated a negative correlation between death certification and notification of AIDS.[31]

5. *Physicians may complete death certificates inaccurately due to a lack of understanding of the form.* In general, they receive no instruction on the proper completion of death certificates. Hence, causes of death such as "cardiac arrest" or "respiratory failure" are commonly seen. As these events are common to all deaths, they do not provide new or useful information.

6. *Ontario has used information from the mortality database to identify previously unreported AIDS cases.* This means that the two databases are not independent in Ontario. The death certificate does not contain sufficient information to complete a report form. Up to 25% of all AIDS reports in Ontario in the past have followed identification of a non-reported AIDS case from death certificates. The majority of these cases meet the case definition required by ACRSS.[32] The cases that were identified post-mortem can be flagged but one can only estimate the number that would have been eventually reported as lengthy reporting

delays are not uncommon in Canada.

7. *Deaths outside of Canada may contribute to the absence of a death certificate.* Most, but not all, American states will forward reports of death of Canadian citizens within their boundaries to Canadian authorities. No such practice is noted from any other country in the world.

8. *Lists of AIDS cases in mortality databases are biased towards cases that occurred earlier in the epidemic and perhaps cases that are more severe (as one has to die to get on the database).* Death certificates allow the validation of reporting among persons with AIDS who are also dead, not among those living with the disease. Those who were infected early in the epidemic are more likely to be dead as are those with more severe disease. This death bias is a form of selection bias.

9. *Reporting completeness to the AIDS surveillance system for the diagnosis of AIDS may be relatively complete but may be quite incomplete for notification of death.* That is, the physician may report upon diagnosis of AIDS but not inform the surveillance system of the patient's death. This bias must be considered carefully to ensure that one is measuring underreporting of AIDS and not underreporting of deaths due to AIDS. As a result, it is necessary at times to match death certificates to people listed in the AIDS report system as living.

10. *Insufficient supplementary personal identifier information on the two databases can make a link difficult.* This leads to overestimation of

underreporting resulting from a failure to link, a form of misclassification bias.

11. *AIDS on the death certificate may not meet the case definition required by the surveillance system.* This was more likely to be a problem with the earlier case definitions. For example, pulmonary TB was not listed as an AIDS-defining illness prior to 1993. Therefore, an HIV positive person who died of pulmonary TB in 1990 may be considered to have died of AIDS by the physician. However, as this person with HIV did not meet the surveillance case definition at the time, the physician could not report this as a case of AIDS to ACRSS.

4.4. Capture-recapture

As discussed previously, the Canadian Mortality Database remains the best secondary data source for assessing ACRSS reporting completeness at a national level. The CMDB does not capture all AIDS cases. There are two methods to consider in estimating the reporting delay/completeness of AIDS in Canada. One could find all the cases which are missing from the ACRSS database but which are present in the CMDB and add them to the denominator. This would be inadequate as the CMDB is incomplete. The second method must be considered in the presence of two incomplete sources of data and no gold standard. Capture-recapture techniques, as the name suggests, were first developed by ecologists in the last century who were attempting to estimate the numbers of various species in the wild. This situation, with incomplete data, has been a problem as well in demography in underdeveloped countries and in measuring elusive populations such as the homeless.

The fish in the lake example is often used to describe this technique. To estimate the number of trout in a lake a net would be dropped in the lake, some trout caught, counted and tagged, and released back into the lake. After allowing time for dispersal and mixing, the net would be thrown in a second time and again the number of trout caught would be counted, noting those which had been tagged from the first catch. The total number of trout in the lake could then be estimated based on the number of fish caught with the net each time and the

number that were caught twice. Intuitively, if nearly all the fish caught in the net the second time are tagged, it would be reasonable to conclude that there are not many other fish out there. If very few tagged fish appear in the second net, the population estimate will be much higher.

This method does require some assumptions[33]. Probably the most important assumption is that the two samples are independent. The chance of a trout being caught in the net the second time should be statistically independent of its chance of getting caught the first time. Lack of independence can be positive or negative and is known in ecology as "trap fascination" and "trap avoidance". The interaction with the net the first time could change the manner in which the fish reacts to seeing a net the second time. It may be that only slower or less intelligent fish are caught by the net. This could be considered the Achilles heel of capture-recapture. This has led to the development of multiple-record ($k>2$) methods for estimation of population size such as the use of Bernoulli census or log-linear methods.

A second assumption is that the population is a closed one. This is virtually never the case with biological systems as fish will be born and die and it is also a problem in epidemiology. Time limitations must be set around the times that the two nets can be cast, so that this assumption is nearly satisfied.

A third assumption is that individual identifiers should not be lost or overlooked. In the case of the trout, this would mean that the tags do not fall off the trout, the tags are seen by the ecologist, and that the ecologist can tell a trout from another species of fish.

In the situation with ACRSS and CMDB the situation is slightly different. Using the fish analogy, the first net only captures very sick fish who will soon die. The second net only scoops up dead fish. Thus, the two samplings may appear to be from different populations. However, with a median survival of less than two years, one can argue that it is the same population that is being sampled twice.

Considering the problem with assessing AIDS reporting completeness, the purpose of the record linkage is to find which people are listed on both databases. Although both databases undercount the number of true cases, capture-recapture techniques allow one to estimate the number of cases missing overall by using the underascertainment rate of each data source. For the ACRSS-CMDB linkage, the capture-recapture assumptions must be considered:

1. *The sources should be independent.* The chance of a person being included on the mortality database should be independent of their chance of being included in the ACRSS database. This assumption is not considered to be a problem when the capture rate of either source is

high.[34,35] As noted earlier, there is really no adequate third data source available for AIDS on a national basis. Despite the restriction to two data sources, there are means of assessing the lack of dependence to some extent. Sekar and Deming suggest stratification of the data by a third variable such as county.[36] Within each strata an estimate of the total number of cases and the degree of completeness for both systems would be calculated. A correlation coefficient between the completeness of reporting of both systems weighted by the number of cases in each stratum is then calculated. Independence can be assumed if the correlation coefficient does not differ from zero. Desenclos and Hubert proposed comparing the sum of strata estimates to the crude unstratified estimate.[37] Recently, in some of the epidemiological literature, the independence assumption has been considered as two separate assumptions: a lack of list dependence and a lack of apparent dependence (also known as heterogeneity).[38] This difference is subtle to the point that a number of articles do not discuss it. Nevertheless, it appears that the methods of Sekar-Deming and Desenclos-Hubert only address heterogeneity. Sekar-Deming's method was used in this study by stratifying for year of death.

2. *The population should be closed.* In the present case, losses from these databases were clearly not a problem, however there were regular additions. In order to simulate a closed population, cutoffs were

determined based on different levels of acceptable reporting delay and on the expected survival after AIDS diagnosis.

3. *Individual identifiers should not be lost or overlooked.* Effectively, this means that the record linkage should be accurate with minimal false positive or false negative matches.

4.5. Record Linkage with CMDB

Record linkage is a process by which records of one individual in separate databases are brought together. This can be done manually but computer record linkage is most commonly used. The following discussion is in two sections: a general description of computerized record linkage followed by the specific methodology used in this study.

4.5.1. General Concept of Computerized Record Linkage

The first step in a computerized record linkage is to confirm that there are adequate overlapping fields in the two databases such that a linkage will be possible. Even with fully nominal personal identifiers, variability in spelling of common names (ie Jeff or Geoff), the use of common substitutes/nicknames (ie Robert or Bob), and transcription errors will hamper record linkage. Therefore, although exact or deterministic matches are ideal, in practice it is not possible and the use of probabilistic linking is required in which the probability of linkage versus non-linkage must be estimated based on summing probability weights.[39] The use of probabilistic linkage has allowed the linkage of databases lacking personal names[40]. Probabilistic linkage will be described in detail.

The overlapping fields must be coded in the same manner in both files (eg date fields YYMMDD and YYYYMMDD must be coded similarly). Duplicates within each database must be identified and removed. Records are then divided into pockets or blocks based on one or more of the more reliable fields such as date

of birth (DOB), date of death (DOD), sex, initials, etc. This saves considerable computer time by limiting matching to those pairs of records which are potentially linkable. As there is always a chance of error in any block, more than one pass is usually done using different blocking fields. As long as the blocking fields are independent, it is unlikely that a true match will be missed.

Some fields have higher predictive value than other fields. For example, a match on day of birth is much more significant than a match on sex as there are 366 possibilities for the former and only two for the latter.

Each field has two probabilities associated with it, known as the *m* and *u* probabilities. The *m* probability is the probability that a field agrees given that the paired records being tested are a true match. This is equivalent to one minus the error rate of the field. For example, if the year of birth is wrong 5% of the time, then the *m* probability will be .95 (the chance of a true positive). The *m* probability is initially estimated but refinement is possible after some linkage is done to achieve a record of matched pairs. This allows the calculation of a series of specific *m* probabilities for that dataset. A different *m* probability is possible for each value of every field. For example, the probability of the 11th day of a month making a true match is greater than that of the 21st doing so, as crossed numerals in the first case does not affect the value (misplacing the two ones still gives 11 while crossing the two and one in 21 gives 12).

The *u* probability is the probability that a field agrees given that the record pair is not a true match (the chance of a false positive). This probability is effectively the probability of random agreement as there are so many more unmatched pairs possible than matched pairs[41]. For example, in a perfect match between a file with two million records and a file with one hundred records, there will be 100 matched pairs and nearly 200 million unmatched pairs. Frequency tables can be created for the data set to calculate individual *u* probabilities for each value of every field. For example, the surname Smith would have a higher probability of random agreement than the surname Schickelgruber.

In the linkage program, weights for each of the comparison fields for each pair are calculated. Weight has been defined as the logarithm to the base two of the ratio of *m* and *u*. If *m* is greater than *u*, the weight will be a positive number. A composite weight can be calculated by summing up all the comparison field weights for each pair. Composite weights tend to follow a bimodal distribution as shown in Figure 1 in Appendix A. Those pairs with high weights are considered to be matches with no further examination necessary. Similarly, those with low weights are considered non-matches. The threshold weights are determined by comparing a sample of files manually over a range of composite weights. Those pairs with composite weights between the two threshold weights are considered potential matches and must be resolved manually. Potential links are handled by "manual resolution" in which one uses additional information, if available,

from the data sources to make a decision on whether there is a true link. If no additional data are available and no rules for resolving indeterminate links, the value of manual review is limited. This process will be discussed in greater detail later.

4.5.2. The ACRSS-CMDB Link

The general record linkage methodology used for this study is a highly sophisticated program utilized in many studies. The program used by Statistics Canada is called the Generalized Iterative Record Linkage System (GRLS; the I has been dropped for political correctness). This was a more difficult linkage than most due to the limited number of personal identifiers. It should be noted however that Calzavara et al in 1990[3] demonstrated the accuracy of record linkage using only initials, date of birth, and city of residence on one file. The ACRSS file included the 10,391 reports of AIDS cases received at Health Canada between January 1982 and September 1994. The CMDB file contained a list of all certified deaths from January 1982 to December 1992, approximately 2.0 million deaths in total. These were the most recent files available from each source. The additional 20 months available from the ACRSS file was useful as it allowed for linkage with records with considerable reporting delay. It is the practice of Statistics Canada to generate one file for each surname, consequently married women with both maiden and married names have two records. This duplication is done for the purpose of record linkage. The two million deaths generated 2.8 million death records. The actual process of the linkage involves preprocessing each file, planning the passes, exploding the files, calculating the rules for weighting, and finally running the required program. These steps are described below.

4.5.2.1. Preprocessing ACRSS

Preprocessing of the file is required to create compatible overlapping fields. There were 25 fields included without modification (other than name changes) and 27 new fields created. The modifications are shown in Table 1 in Appendix B. The final file contained 52 fields. These are listed in Table 2, also in Appendix B.

4.5.2.2. Preprocessing CMDB

The CMDB had 52 fields and contained 2.8 million death records. This file was also preprocessed. Seventeen fields were used unchanged; 13 new fields were created. The AIDSFLG code divided the CMDB into five large groups based on the ICD-9 coded cause of death (referred to as AIDS Flag 1 to 5 in decreasing order of likelihood that the deaths were due to AIDS). It was unknown prior to the linkage what proportion of deaths due to AIDS were coded as 279.1 prior to 1987 and as 042-044 from 1987 on. As will be discussed later, the improved efficiency of CMDB at finding AIDS deaths by including other codes allowed us to minimize any errors due to lack of independence between the two data sources using capture-recapture. To increase the likelihood of finding deaths due to AIDS, specified infections or specified malignant neoplasms under 042 as identified by the ICD-9 addendum establishing the codes 042-044[42] were coded as AIDSFLG 2. An example of this would be a cause of death such as pneumocystosis (136.3). "Other specified conditions" included under 043 which are much less likely to be due to AIDS were given the AIDSFLG code 3. An

example of this would be infectious diarrhea (009). The coding is shown in Table 1.

Table 1 - AIDSFLG Coding in CMDB

AIDSFLG Code	ICD-9 Codes
1	codes highly predictive of AIDS (ICD 042, 043, 044, 279.1, 795.8)
2	codes moderately predictive of AIDS (ICD 003.1, 007.2, 007.8, 007.9, 010, 011, 012-018, 031, 038, 039, 046.3, 054, 078.5, 112, 114, 115, 117.3, 117.5, 117.7, 127.2, 130, 136.3, 173, 180, 200.0, 200.1, 200.2, 200.8, 202.8, 290.1, 290.8, 290.9, 294, 298.9, 320, 321.0, 322.2, 322.9, 331.6, 348.3, 363.2, 481, 482, 484.1, 516.8, 711, 730, 790.7)
3	associated conditions as listed in MMWR Dec 25/87, Vol 36/S-7[42]
4	all deaths under age 15
5	all other causes of death

This allowed comparisons of deaths that were unlikely to be due to AIDS (AIDS Flag 5) and gave additional weight to matches with AIDS Flag 1 or 2. The final file contained 30 fields. They are listed in Table 3 in Appendix B.

4.5.2.3. Planning the Passes

The linkages were done in four independent runs: males with initials, males without initials, females with initials, and finally females without initials. Approximately 16.1% of ACRSS records lacked initials. These records were mainly Quebec ACRSS cases. Quebec has not recorded initials on new AIDS cases since November 1990. The four runs were required as different comparison rules were needed for files with and without initials due to the decreased discriminating power of the program in the absence of initials. In files lacking initials, only the date of birth, date of death, and geographic information could be used to link records. Females were linked separately from males as

AIDS is much more uncommon in females and errors in this variable field are highly unlikely. The order of blocking was decided by looking at a previous linkage between the CMDB and ACRSS. The order of blocking is shown in Table 2.

Table 2 - Blocking Variables used in each Pass

Pass Number	Blocking Variables
1	Full date of birth (YYMMDD)
2	Full date of death (YYMMDD)
3	Last initial, first initial, birth year (LFYY)
4	Last initial, birth month and day (LMMDD)
5	Last initial, first initial, death year & month (LFYYMM)
6	Birth year & month (birth year +1/-1, birth month & day crossover)

The first pass was based on full date of birth(YYMMDD). If there was an error in the date of birth then a true match could not occur. This was the only pass in which the residue file (AIDSFLG code 5 representing all causes of death unlikely to be due to AIDS) was used. As the residue file was very large and the probability of true matches was very low, chances for linking to these other causes of death were limited to the one pass. The passes that were run are shown clearly in Table 3 on the next page.

Table 3 - AIDSFLG Code and Passes

AIDSFLG CODE	PASSES					
	1	2	3	4	5	6
1-codes highly predictive of AIDS	*	*	*	*	*	*
2-codes moderately predictive of AIDS	*	*	*	*	*	*
3-codes associated with AIDS	*	*	*	*	*	
4-age< 15yrs	*	*	*	*	*	
5-all other causes of death (suicide, accidents, etc)	*					

If a CMDB record was coded AIDSFLG 1 or 2 then a copy of this record would be run through each of the six passes. This would maximize the chances of a true match if there was an error in one of the fields in either database. If multiple matches were made to one record, the match with the highest composite weight would generally be accepted. Note that only ACRSS records lacking initials were run through Pass 6 which was an attempt to compensate for errors in birth years and crossovers of birth month and day.

4.5.2.4. Exploding ACRSS

The preprocessed file was exploded (reproduced) to produce six copies to allow for six passes to occur. Note that a record was only exploded if the blocking variable to be used was present for that observation. For example, if death date was not recorded on the ACRSS record, the observation was not duplicated for passes 2 and 5. This exploded file was then divided into males and females with 38,000 and 2,100 records respectively.

4.5.2.5. Exploding CMDB

This master file of 2.8 million death records was exploded (reproduced) to give

six copies for AIDS Flag codes 1 and 2, five copies for AIDS Flag codes 3 and 4. The residue file (AIDS Flag 5) was only used for one pass with full date of birth so copies were not needed. This residue file was divided into males and females once more giving 1.5 million and 2.2 million observations respectively. There were more female records as records were produced with both last name and maiden name. Approximately 8-9% of the records in the CMDB were coded AIDSFLG 1-4 although most of these were in level 3. The linkage process is shown in Figure 2 in Appendix A.

4.5.3. Running the Linkage

Given that L is the weighting for linkage based on the *m* probability and U is the weighting for unlink based on the *u* probability; each possible matching pair of records was run through a series of comparison rules which assigned U and L weights depending on whether the pair met the conditions stated. A rule was a set of criteria by which a pair of records was compared and thereby given weights. Each possible matched pair was run through 25 comparison rules which are shown in Table 4 of Appendix B and are discussed further in the next section. All links with a cumulative weight (discussed below) over -100 were considered as potential links, cumulative weights over 80 were considered definite links. Note that these thresholds were based on past experience with the program. At this point in the linkage, the weights were called basic or global weights. For example, a match on the first initial Z was given the same weight as a match on the first initial B. Further processing assigned value-specific weights and is described in the following section.

4.5.4. Rules for Weighting

The rules for weighting were based on the probability of a true match versus a non-match. Some of these probabilities were quite exact, others were based on previous experience. A summary of the comparison rules follows:

1. The first weights adjust for file size which changes slightly depending on the year of death. Age-specific likelihood of death was added as a weight reflecting the lower probability of death in the younger ages versus the elderly population.

2. Negative weights were assigned for logical errors. These are listed below:

   (a) date of onset in ACRSS being after the date of death in the CMDB.

   (b) negative weights were proportional to the difference in birth days, months, or years between the two data sources.

   (c) crossed birthdates such as month and day reversed in one file.

   (d) partial agreements on dates of death were progressively weighted as with birthdates.

   (e) surnames and given names were compared with initials with increasing negatives weights when initials were crossed, disagreed, or were missing.

3. Geographic agreement was weighted at the level of the province, census division, and census subdivision. Census divisions are quite large, for

example Ottawa-Carleton is a single census division. Ottawa or Nepean are examples of census subdivisions. This comparison weighting was complex as ACRSS had three sources of geographic information (place of diagnosis, residence, and hospital) and CMDB had two (place of residence and place of death). The best two agreements out of the six possibilities were weighted.

4. Matches on diagnosis and cause of death were sought using the AIDSFLG levels. Positive weighting was given if the cause of death agreed on AIDSFLG levels 1 or 2.

5. More specific agreements on specific diagnoses and causes of death were then weighted. An agreement on an unusual disease of low probability of occurrence, such as cryptococcosis, would receive the same weight as a more common disease of higher probability (ie pneumonia). That is, diseases were not assigned value-specific weights. Theoretically, a match on cryptococcosis should receive extra weighting as the probability of this occurrence was much less. One concern was that some physicians have coded the same diagnosis twice when there is a relapse. This may have allowed overweighting because a PCP cause of death could be matched with both diagnosis 1 and diagnosis 2. Another concern is that although people with AIDS get more than one disease, ACRSS would generally receive the first diagnosis only while CMDB only receives the last.

4.5.5. Manual Resolution

To determine the rules for a priori manual resolution a pilot study was performed on one year's data. Data from 1990 was linked to ACRSS using global weights to ensure that there were no major problems with the linkage. Global weights were later replaced by value-specific weights once it was clear that there were no logical problems with the linkage program. The following observations were made by reviewing all of these potential links:

1. The lack of value-specific weighting and other factors explained where our manual resolution differed from the program output as follows:

   (a) Country of birth was very influential in decision-making.

   (b) City/community size should be a factor, ie matches on small cities such as Moncton should have been weighted higher.

   (c) Rare dates of birth for diagnosis of AIDS should be given higher weighting. For example, a match involving a person with a date of birth in 1919 who died from AIDS should have a high weight as AIDS is a rare disease in the elderly. There were concerns that the value-specific weights to be used were based on frequencies found in the mortality database as this was the larger of the two databases. A 1955 D.O.B. for a death in CMDB is less common than a 1919 D.O.B. but the opposite frequency would be seen in ACRSS.

2. Differences in date of death by one day may have been penalized too

much. Deaths during the night often lead to this discrepancy.

3. If the place of birth was not given, the name could give evidence of geographical origin.

4. If there were no initials, then DOB and DOD were used to match but the cause of death and AIDSFLG became very important especially if DOD was missing from ACRSS. If the cause of death was AIDSFLG 5 (ie the residual file), one should be reluctant to make a match based only on date of birth and date of death because the residual file was so large that false positive matches were likely to occur. It seemed unlikely that we would link any AIDS patients who died of other causes (such as suicide or MVAs) in Quebec.

5. The most difficult resolutions involved files without initials (from Quebec). Two examples are listed below:

    (a) no initials, same D.O.B., in Montreal, AIDS diagnosis

    (b) no initials, same D.O.B., different D.O.D., AIDS diagnosis.

6. Matches within suburbs such as Halifax-Dartmouth, Quebec-Ste Foy, etc were given too little weight as the program did not account for their proximity.

7. The cause of death was not found to contribute to decision making if it was not AIDSFLG 1 or 2 because non-specific diagnoses such as pneumonia (486) and unspecified nonspecific illness (799.9) were often listed.

8. AIDS-defining illnesses in the ACRSS file did not contribute to decision making as the cause of death is just as likely to be some other opportunistic infection due to the multiple illnesses seen in people living with AIDS.

9. Geographic codes became more significant with age as the younger gay population is quite mobile between the major Canadian cities of Montreal, Toronto, and Vancouver. There is also the anecdotal mobility of AIDS cases from their city of acquisition of HIV infection/residence back home to smaller centres to die. In keeping with this, if a case was identified as alive in ACRSS and the place of diagnosis was different from the place of death, then the geographic difference was ignored. Older age cases, which were more likely to be transfusion-recipients, were expected to be less mobile.

10. If the code was over 799, the cause of death was manually reviewed as it may have been suicide related.

11. When initials were present and matched; the following rules were used:

    (a) If the CMDB record was AIDSFLG 5, other fields such as DOB, DOD, and geography would have to be perfectly matched to link.

    (b) If the CMDB match record was an older person and AIDSFLG 1 (ie AIDS on the death certificate), this was sufficiently unusual that other variables would not have to be exact matches.

    (c) If the DOB agreed but DOD was missing in ACRSS and CMDB

cause of death was AIDSFLG 5, the frequency of death in that age group and the cause of death were carefully examined. For example, an elderly person certified as deceased due to heart disease is unlikely to be a true match even with matching initials and DOB. In these cases two reviewers had to agree for a match to be made.

Based on this pilot, our threshold for "definite" matches was estimated to be in the range of 75 to 80 at the top end and at -20 at the bottom end using global weights. It was decided that for AIDSFLG1 cases, links should be manually checked down to the lowest weight of -100. A careful search through the most recent ACRSS file (October 95) was needed if these AIDSFLG1 cases remained unlinked.

4.5.5.1. Males with initials

Following the test match and clarification of the methodology as described above, the file of all males with initials was linked to the 1982-1992 mortality file using value-specific weights. Value-specific weights were used for initials, geographic areas, birthplace codes, and age at death. This produced 5,101 potential links. Links were provided as a one ACRSS record to multiple CMDB record configuration, that is each ACRSS record was listed with at least one potential CMDB match. The vast majority of links (94.0%) were in a 1:1 ratio. In the males with initials linkage, roughly 50% of the matched files were from

Ontario, 20% from B.C., 18.8% from Quebec, and 6.1% were from Alberta. The under-representation of Quebec cases in this group was due to the absence of initials in Quebec data since November 1990.

Manual review of a 10% systematic sample (511) was done to ensure that the program was producing reasonable output and also to set thresholds a priori. The upper threshold for manual resolution using global weights had been estimated to be about 80 while the lower threshold was estimated to be approximately -20. This estimation was repeated at this time using value-specific weights. Matched records were divided into "probable" matches, "possible" matches, or no match and compared with the weight given by the GRLS program as shown in Table 5 in Appendix B. Note that "probable" means definite from a practical point of view but the term "probable" is preferred as this is a probabilistic linkage system where no match is identifiable as truly definite. The upper threshold was set at 110, the lower at -20. Above the upper threshold virtually all of the links were "probable" matches, below the lower threshold virtually all links were non-matches. It was expected that the span of weights requiring manual resolution would consist of about 260 links in the total file. All links with weights between 200 and 110 with AIDSFLG 3 or 5 would be reviewed to minimize false positive matches. The AIDSFLG1 death certificates matched with weights between -20 and -100 (ie non-matches) would be manually reviewed to avoid false negative matches. False negative matches would be

recorded as death certificate only (DCO) cases which would falsely elevate underreporting.

Nearly 93% of all links were "probable". Having set the thresholds for manual resolution, the full file of links for males with initials was printed out. All links in the indeterminate range were reviewed independently by myself and by M.R. ACRSS records and death certificates were obtained, as needed, to assist in resolving these indeterminate cases manually. There were 99 matched pairs in AIDSFLG 3 and 5 cases with weights between 200 and 110. Death certificates from Ontario for the years 1990 to 1993 were not readily available to Statistics Canada due to problems related to their conversion to optical disk technology. This decreased the number of accessible death certificates to 55. Of these 55 death certificates, 33 (60%) indicated definite HIV or AIDS. An additional three were very probable with diagnoses of overwhelming viremia, hemophilia, and immune suppressive illness/sarcoma. Of the 19 remaining, 10 gave no additional information or were interim reports prior to a final coding of suicide. For weights between 110 and -20, 53 cases were AIDSFLG 3 or 5, 13 Ontario charts were unavailable. Of the remaining 40, 17 (42%) were definitely HIV or AIDS on the death certificates. The lowest weight at which an AIDS case was identified by this manner was at -2. This raised the question of whether we should pull death certificates at even lower weights. This was done by taking a systematic sample of AIDSFLG 3 and 5 death certificates with low weights

between -20 and -100. There were 124 of these potential links but checking the latest ACRSS database (October 95) allowed the elimination of 33 of these links on the basis of recent death dates that were not in agreement with those on the death certificate. There were a total of 91 remaining of which 60 were not Ontario deaths between 1990 and 1993. Every second death certificate was then pulled to assess the possibility of missing true matches even at this low weight. Of these 30 death certificates, there was no evidence that any of the deaths were associated with HIV/AIDS. Clearly, the linkage program was working as it should in that the probability of an AIDSFLG 5 death being actually due to HIV/AIDS decreased as the weight decreased as shown in Table 4 below.

Table 4 - Linkage Weight with AIDSFLG 3 and 5 and HIV/AIDS on Death Certificate (Males with Initials)

Linkage Wt with AIDSFLG 3 and 5	200 to 111	110 to -20	-21 to -100
Total Links	99	53	91
Death Certificate Available (not Ontario)	55	40	60
HIV/AIDS on death certificate	33 (60%)	17 (42%)	0 (0%)

For this reason, death certificates for the links with high weights between 504 and 200 which were AIDSFLG 3 or 5 were not examined in hard copy (this included 35 and 100 links respectively).

Most of the cases missed by the use of the underlying cause of death were due to coding of AIDS in Part 2 of the death certificate, which contains contributing but not causal factors. There were instances in which the physician gave an

underlying cause of death of shorter duration than the immediate cause. The nosologist cannot code the cause of death due to the underlying cause in these circumstances. In another case, a physician gave the immediate cause of death as being due to bilateral pneumonia. This was due to or as a consequence of Hodgkins disease which was listed as due to or as a consequence of AIDS. As AIDS is not on the causal pathway to Hodgkins disease, the cause of death stopped at Hodgkins disease which was then coded.

The main focus of the manual resolution was the potential matches with weights between 110 and -20. This included a total of 230 potential matches. These were assessed individually by the two investigators blind to each other's decisions and then compared. The results are shown in Table 5.

Table 5 - Comparison of Manual Resolution Results from Two Investigators (Males with Initials)

Weights between 110 and -20 (males with initials)		MR		Totals
		+	-	
JW	+	171 p_o=.743 p_c=.689	24	195 (.848)
	-	16	19 p_o=.083 p_c=.028	35 (.152)
Totals		187 (.813)	43 (.187)	230 (1.000)

p_o=observed proportion of agreement
p_c=chance expected proportion of agreement
kappa=$(p_o-p_c)/(1-p_c)$=0.38

The overall agreement was approximately 82%. The kappa value was calculated to be .38, that is 38% agreement over and above chance. It was agreed that this kappa was too low leading the two observers to an unblinded

comparison of discordant matches. There were 40 matches in which the two investigators disagreed on the outcome (17% of the 230 total between weights 110 and -20). These disagreements were resolved in one of two ways:

1. One person agreed that the evidence favoured the other decision and a change was made (n=12).
2. The investigators could not reach agreement and a third category of link was created, to be known as a "possible" match (n=28), the other two to be known as "probable" (as close to definite as possible) and non-matches. "Possible" matches fell into 9 groupings which are listed in Table 6 in Appendix B.

4.5.5.2. Males without initials

A special linkage run was necessary as all Quebec cases from November 1990 forward do not have initials. There were a total of 957 "possible" links weighted between 343 and the threshold of -80. All links with weights less than -80 were considered non-matches. In the previously established range of weights requiring manual resolution between 110 and -20 there were 351 potential links (36.7%). In comparison, recall that for males with initials there were 230/5101 (4.5%) in the same weight range. This illustrates the loss of discriminating power by the linkage program which occurs with the loss of personal identifiers from the ACRSS database.

It became apparent during the manual resolution of these cases that a large

number of matches would have to be placed in a "possible" category; these were flagged for the purpose of inclusion and exclusion in a sensitivity analysis. Rules were developed to assist with the matching process as described earlier. The rules are listed below:

1. If the case was AIDSFLG 1, then either DOB or DOD had to be exact with only a minor error allowable in the other date (either DOB or DOD) as in the day, month, year, or a crossover to be considered a "probable" match.
2. If the DOD was missing for an AIDSFLG 1, then a "probable" match would be confirmed only if the geographical variables matched perfectly and were outside of Montreal or the birthplace was outside Canada and matched. Montreal as a place of diagnosis was not significant for matching as the vast majority of all AIDS cases in Quebec are diagnosed in that city.
3. A DOD off by one day was considered equivalent to exact agreement for this field as deaths during the night are often misclassified.
4. If the DOB or DOD was missing but the other fields matched in AIDSFLG 3 or 5, then it was not a match unless review of the death certificate indicated HIV or AIDS which elevated the match to a "possible".

As a check, similar to the males with initials match, death certificates on all matches with AIDSFLG 3 and 5 between weights 200 and -20 were pulled. The percentage of HIV/AIDS on the death certificate decreased as the linkage weight

decreased:

Table 6 - Linkage Weight with AIDSFLG 3 and 5 and HIV/AIDS on Death Certificate (Males without Initials)

Linkage Weight	200 to 111	110 to 61	60 to 21	20 to -2
HIV/AIDS on death certificate	4/7 (57%)	6/9 (67%)	2/7 (28%)	7/52 (13%)

As in the run "males with initials", the manual resolution was done independently by the two investigators. It was decided in advance that matches would be designated as "probable", "possible", or non-links as the lack of initials limited discriminating power. To assess the objectivity of the manual resolution, weighted kappas were calculated using weight 0 for agreement, weight 1 for "probable"-"possible" or "possible"-nonlink, and weight 2 for "probable"-nonlink. The results are shown in Table 7.

Table 7 - Comparison of Manual Resolution Results from Two Investigators (Males without Initials)

Weights between 110 and -20 (males without initials)		MR						Totals
		"probable"		"possible"		non-link		
JW	"probable"	36	wt=0 p_o=.102 p_c=.020	14	wt=1 p_o=.040 p_c=.055	0	wt=2 p_o=0 p_c=.087	50 (.142)
	"possible"	10	wt=1 p_o=.028 p_c=.061	108	wt=0 p_o=.308 p_c=.170	35	wt=1 p_o=.100 p_c=.205	153 (.436)
	non-link	3	wt=2 p_o=.008 p_c=.059	15	wt=1 p_o=.043 p_c=.164	130	wt=0 p_o=.370 p_c=.198	148 (.422)
Totals		49 (.140)		137 (.390)		165 (.470)		351

p_o=observed proportion of agreement
p_c=chance expected proportion of agreement
kappa=1-(sum $w_i p_o$/sum $w_i p_c$)=0.69

This gave a weighted kappa of .69 indicating that the agreement between investigators over and above that due to chance was relatively good. The relatively few disagreements were resolved by examining the matches together, looking for errors in logic or reviewing the rules stated above. If agreement could not be reached, then the match was considered a "possible".

4.5.5.3. Females with initials

The manual resolution on females with initials was uncomplicated. There were 260 potential link groups for females with initials (ie one ACRSS file to many CMDB files such as 1:1, 1:2, 1:3, 1:4). Of these, only 18 had weights less than 110. Most of the links were from Quebec (n=117) reflecting the high proportion of women with AIDS from that province. Ontario had the second largest number (n=89). Of the ten death certificates for those coded AIDSFLG 3 or 5 between weights 200 and -20, one was unavailable, three indicated HIV/AIDS and were reclassified as AIDSFLG 1. All were classified as "probable" or non-links, it was not necessary to use the "possible" designation.

4.5.5.4. Females without initials

There were 71 potential link groups (one ACRSS:many CMDB) for females without initials. 32 (45%) of these 71 had weights less than 110, again reflecting the loss of discriminating power of the linkage program when personal identifiers were missing. 66 of the 71 deaths were from Quebec, the remainder were from Ontario and B.C. Rules were applied to distinguish "probable", "possible", and non-matches. As before, the most recent ACRSS file and hard

copies of death certificates were pulled to attempt resolution of gray area linkages. Six death certificates with AIDSFLG 3 or 5 were pulled, of these, three certificates noted HIV/AIDS.

The final analysis file was then created with the linked cases identified as "probable" or "possible". 98.2% of all links had a weight greater than 110. 94.3% of all links were to AIDSFLG 1 or 2 death files. A separate death certificate only (DCO file) was also produced containing death records from CMDB identified as AIDS that did not link to any ACRSS case. Both the linked file and the DCO file were merged with Statistic Canada's Geographic Tape File ("ACRSS only" cases could not be merged because of the absence of specifying geographical information). Analysis of migration and reporting completeness with respect to community size based on census metropolitan areas or census subdivisions was conducted using this file.

4.6. Analysis

4.6.1. Representativeness

To assess bias, the ACRSS file was compared with the linked file (as "probables" and "possibles") and the DCO file in terms of gender, province of residence, year of death, community size at death, risk group, and age group at diagnosis. In the ACRSS file, the province of attribution for the AIDS report was used as the province of residence. Province of residence is relatively incomplete in ACRSS because it is assumed to be the same as the attributed province unless specifically stated otherwise. Province of attribution is defined as the province where the first signs or symptoms of HIV infection appeared. Size of community at death was measured by census metropolitan area (CMA). The presence and absence of initials in the files was also compared for each of the demographic variables.

Length of survival and date of onset in the linked and ACRSS files were compared to assess the presence of a "death bias". As the secondary data source in this study is a file of deaths, it is possible that those listed in the CMDB will be more likely to have severe disease with shorter survival. Those listed in the CMDB may also have developed AIDS earlier in the epidemic. The result of a death bias would be that the linked file would be less representative of the ACRSS file in terms of length of survival and date of onset, as well as any other factor which may change over time such as risk group, gender distribution. Date of onset was defined as the date of onset of the earliest AIDS defining illness.

4.6.2. Calculation of Reporting Completeness from the Linkage

The calculations required follow from Table 8 below using notation often found in the demographic literature.[36]

Table 8 - Calculation of Reporting Completeness

$N=(n_1n_2)/n_{12}$		ACRSS		Total
		+	-	
CMDB	+	n_{12}	N_1 (DCO)	n_1
	-	N_2	x	
Total		n_2		N

n_1=total number of deaths in CMDB
n_2=total number of deaths in ACRSS
n_{12}=number of deaths found in both ACRSS and CMDB (linked file)
N_1=deaths found only in CMDB (DCO or death certificate only cases)
N_2=deaths found only in ACRSS
N=total deaths=$(n_1n_2)/n_{12}$

N as given in the last equation above is an estimate as it is calculated by solving the following equations which contain estimates:

$E_1=n_1/N$ where E_1 is the probability that a person that died of AIDS is identified by CMDB. (Equation 1)

$E_2=n_2/N$ where E_2 is the probability that a person that died of AIDS is identified by ACRSS. (Equation 2)

Since $n_{12}=n_2 \times n_1/N=(n_2/N) \times (n_1/N) \times N=E_1E_2 \times N$ assuming independence of sources:

i.e. $n_{12}=E_1E_2N$. (Equation 3)

Solving Equation 3 for N (replacing E_1 and E_2 with Equations 1 and 2 respectively), one obtains:

$N=(n_1n_2)/n_{12}$.[34] (Equation 4)

Rearrangement gives:

$n_2/N=n_{12}/n_1$ (Equation 5)

The left side of Equation 5 gives the reporting completeness of ACRSS which is

the total number of deaths in ACRSS divided by the total population size. The right side of Equation 5 is equivalent to $n_{12}/(n_{12}+DCO)$ and is actually used for the calculation. Other variables to be considered are N_1 and N_2 which represent the deaths that are only found in the respective databases and are not shared. Note that $n_1 = N_1 + n_{12}$ and $n_2 = N_2 + n_{12}$.

The linked file provided n_{12} by linking ACRSS 1982-Sept 94 and CMDB 1982-92. The residue (death certificate only or DCO) file gave the deaths that were found only in the CMDB which is N_1. From this, n_1 and reporting completeness were calculated. Note that the variable n_{12} will have CMDB + ACRSS(death notified) and CMDB + ACRSS(death not notified) if a death variable is not used for all the passes in the linkage. In this way, we examined the reporting completeness of ACRSS by looking at whether cases were reported (ignoring mortality status). The other option would require the use of a death variable for the linkage such that n_{12} would only include CMDB + ACRSS(death notified). In this case, one would be examining the completeness of reporting of death, among those with AIDS reported to ACRSS, which would be a less useful exercise.

Reporting completeness was calculated using the capture-recapture equation for: all reports, by gender, by province of residence, by year of death, and by size of the community where the individual resided at death. Repeat calculations were done to allow sensitivity analysis by counting "possible"

matches as DCO cases for Canada as a whole and for Quebec alone. Calculations were also done for Canada excluding Quebec. This latter calculation was done with "possible" cases excluded as there were very few "possibles" outside of Quebec. Sekar-Deming's method of testing for independence was utilized by stratification by year of death. The analysis at this point did not account for reporting delay.

4.6.3. Change in Reporting Completeness with Time

To investigate any change in reporting completeness over time, it would be valuable to ascertain the reporting completeness of each cohort of AIDS cases diagnosed in each subsequent year. This was not possible as we did not know the date of diagnosis of death certificate only (DCO) cases; DCO case numbers are required for the calculation of reporting completeness. Therefore we decided to compile cohorts by the dates of death allowing an assumption that date of death predicts date of diagnosis. Reporting completeness of cohorts of AIDS deaths certified by CMDB by the quarter of each year were plotted over time. The initial curve plotted took advantage of the fact that ACRSS data was collected up to September 1994 while the mortality database ended on December 31, 1992 as it allowed for 20 months of reporting delay after death. It was anticipated that the shape of the curve would give some estimate of underreporting (ie AIDS cases that would never be reported) as the effect of reporting delay was minimized with time.

As noted above, this analysis was improved by censoring mortality data successively backwards at 3 month intervals such that reporting completeness could be calculated at quarterly intervals since the year of death. For example, all cases of AIDS in which the death was certified by CMDB prior to December 31, 1985 were totalled. Reporting completeness was calculated with linked cases having been reported to ACRSS by year end; DCO cases were either

never linked or were linked to an ACRSS report after the last day of 1985. Reporting completeness was then recalculated at the end of each subsequent quarter. Some of the DCO cases during each year would link to a newly reported ACRSS case. Each link would decrease the DCO count by one and increase the linked count by one. In this way, cumulative reporting completeness was calculated over the following years. The shape of each cohort's curve was compared to permit projection of recent cohorts whose data was still censored.

Reporting completeness at equivalent years after date of death certification by CMDB were then plotted against the date by which the AIDS death had been reported. The slopes of these curves were then calculated by chi square test for trend on Epi Info. These calculations were repeated with "possible" matches included as DCOs to allow a sensitivity analysis.

4.6.4. Non-reporting Rates

Non-reporting rates were calculated for the one, two, and five year periods after diagnosis. The calculations required are detailed below:

1. Numerator - The reporting delay was calculated for each case in the ACRSS database by subtracting the date of onset from the date that the report was received by the surveillance system at the federal level. These were then aggregated into yearly intervals (Table 1 of Appendix C). The number of cases that were not reported within a set time interval (eg one year) was calculated by subtracting the one year reports from the total number of AIDS with onset in that year from ACRSS.

2. Denominator - A few assumptions were necessary to calculate the denominator which should represent all AIDS cases diagnosed in Canada in any particular year. The denominator would include cases that were reported to ACRSS, cases that were certified on death, cases that were captured by both data sources, and cases that did not appear on either. This is a clear indication for the use of the capture-recapture equation $N=n_1n_2/n_{12}$ (N is the total number of AIDS cases, n_{12} is the number of linked cases, n_1 is the total number of cases in CMDB, and n_2 is the total number of cases in ACRSS). One problem is that although the date of diagnosis is known for n_{12} (linked file) and n_2 (ACRSS), this is unknown for n_1 as this includes linked cases as well as DCO cases. Death certificates do not state the date of diagnosis. Therefore, an assumption was made concerning the time of onset of AIDS in DCO cases. The median survival

in days both overall and in those reported after death (from Table 16) was subtracted from the date of death from the death certificate. Calculated survival in ACRSS and on the linked file were quite close and the average was used in each case. Overall survival was estimated to be 344 days, for those reported to ACRSS after death it was 141 days. This provided a range of the estimated year of onset for the DCO cases. Adding these values to the linked cases gave n_1 (total in CMDB) by year of diagnosis. The other assumptions are those required for the use of capture-recapture. This includes the assumption that all cases have an equal chance of being captured by each data source. This is not the case as only those who are dead have an equal chance of appearing on both lists. By assuming that all those on the ACRSS list had died, it was possible to calculate the likely number of total cases of AIDS in Canada by year of onset as shown in the example in Table 2 of Appendix C. Note that this assumption should give a falsely elevated N and a falsely low non-reporting rate in the most recent years of diagnosis as some of the ACRSS cases cannot possibly link if they are still alive. High and low estimates of N were derived by using either overall or report after death survival times as well as by including "possibles" with the linked file or the DCO file.

All analysis was done using SAS 6.10-6.11 or EpiInfo.

5. RESULTS

5.1. Comparison of Files

The comparison between the ACRSS file, "probable" links, "possible" links, the DCO file, and the presence and absence of initials in terms of various demographic variables is seen in Tables 9 to 14 and is discussed in more detail below. As 602 (10.5%) of the 5755 linked cases were recorded as alive in ACRSS it was apparent that underreporting of death in ACRSS could not be ignored. This prevented the comparisons of year of death and community size at death to the ACRSS file. Therefore the ACRSS file is not included in Tables 11 and 12.

There are a number of possible sources of variation including those due to errors inherent in record linkage, capture-recapture, and sampling error. Sampling error was not believed to be a major source of error due to the large proportion of the population captured by these data sources. Statistical tests of comparison were not used as we were interested in finding any discrepancies, whether statistically significant or not. Furthermore, a large number of differences may appear to be statistically significant with the large numbers involved when in fact the differences have no practical importance.

5.1.1. Gender (Table 9): The percentage of females in the DCO file (10.1%) was higher than in ACRSS (5.9%), the "probable" file (5.2%), and the "possible" file (4.4%). This may be partially due to the lack of initials.

Table 9 - File Comparison by Gender

Gender by File	ACRSS 1194		"Probable"		"Possible"		DCO file		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Male	9779	94.1	5305	94.8	151	95.6	810	89.9	4864	95.1	592	92.4
Female	612	5.9	292	5.2	7	4.4	91	10.1	250	4.9	49	7.6
Total	10391	100	5597	100	158	100	901	100	5114	100	641	100

5.1.2. Province of Residence (Table 10): Note that 84.8% of the "possible" links were from Quebec. Of the linked files lacking initials, 99.5% were from Quebec as well. 66% of all DCO files were from Quebec.

Table 10 - File Comparison by Province of Residence

Prov of Res	ACRSS 1194		"Probable"		"Possible"		DCO file		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Nfld	42	0.4	23	0.4	0	0	2	0.2	23	0.5	0	0
N.S.	168	1.6	84	1.5	0	0	6	0.7	84	1.6	0	0
N.B.	76	0.7	43	0.8	1	0.6	14	1.6	44	0.9	0	0
P.E.I.	8	0.1	5	0.1	0	0	2	0.2	5	0.1	0	0
Que	3084	29.7	1480	26.4	134	84.8	596	66.1	976	19.1	638	99.5
Ont	4296	41.3	2516	45.0	13	8.2	151	16.8	2526	49.4	3	0.5
Man	122	1.2	61	1.1	0	0	8	0.9	61	1.2	0	0
Sask	82	0.8	49	0.9	0	0	4	0.4	49	1.0	0	0
Alta	664	6.4	316	5.6	6	3.8	17	1.9	322	6.3	0	0
B.C.	1840	17.7	1015	18.1	4	2.6	82	9.1	1019	19.9	0	0
N.W.T.	7	0.1	0	0	0	0	1	0.1	0	0	0	0
Yukon	2	0	0	0	0	0	0	0	0	0	0	0
Outside Canada	0	0	5	0.1	0	0	18	1.9	5	0.1	0	0
Total	10391	100	5597	100	158	100	901	100	5114	100	641	100

5.1.3. Year of Death (Table 11): Note that the absence of initials in linked files first showed in 1986 deaths and increased rapidly with time. A similar pattern is seen with the "possible" file. This is not surprising as 82.9% of the "possible" links involved ACRSS files with no initials.

Table 11 - File Comparison by Year of Death

Year of Death	"Probable"		"Possible"		DCO file		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
1982	13	0.2	0	0	1	0.1	13	0.2	0	0
1983	26	0.5	0	0	3	0.3	26	0.5	0	0
1984	73	1.3	0	0	4	0.4	73	1.4	0	0
1985	168	3.0	1	0.6	10	1.1	169	3.3	0	0
1986	323	5.8	1	0.6	36	4.0	322	6.3	2	0.3
1987	498	8.9	10	6.3	60	6.7	486	9.5	22	3.4
1988	604	10.8	10	6.3	78	8.7	584	11.4	30	4.7
1989	797	14.2	16	10.1	109	12.1	760	14.9	53	8.3
1990	883	15.8	23	14.6	155	17.2	805	15.7	101	15.8
1991	1026	18.3	38	24.1	215	23.9	893	17.5	171	26.7
1992	1186	21.2	59	37.3	230	25.5	983	19.2	262	40.9
Total	5597	100	158	100	901	100	5114	100	641	100

5.1.4. Community Size at Death (Table 12): A higher proportion of "possible" links than "probable" links died in large communities. This would be expected as there are many deaths from AIDS in Montreal but very few in Moose Jaw making the latter location easier to match with confidence.

Table 12 - File Comparison by Community Size at Death

Community Size at Death	"Probable"		"Possible"		DCO file		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
>=1,000,000	3826	68.4	136	86.1	618	68.6	3436	67.2	526	82.1
>=500,000	824	14.7	10	6.3	103	11.4	771	15.1	63	9.8
>=100,000	601	10.7	2	1.3	80	8.9	581	11.4	22	3.4
>=9,000	228	4.1	5	3.2	68	7.6	214	4.2	19	3.0
Non-CMA	110	2.0	5	3.2	31	3.4	104	2.0	11	1.7
Other countries	8	0.1	0	0	1	0.1	8	0.1	0	0
Total	5597	100	158	100	901	100	5114	100	641	100

5.1.5. Risk Group (Table 13): There do not appear to be any major differences between the files in terms of risk group.

Table 13 - File Comparison by Risk Group

Risk Group	ACRSS 1194		"Probable"		"Possible"		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Homo/bisexual	7880	75.9	4399	78.6	118	74.7	4050	79.2	467	72.8
Homo/bi/IDU	374	3.6	185	3.3	12	7.6	172	3.4	25	3.9
IDU	310	3	95	1.7	7	4.4	87	1.7	15	2.3
Bloodprod	205	2	123	2.2	1	0.6	109	2.1	15	2.3
Transpre85	184	1.8	140	2.5	2	1.3	130	2.5	12	1.9
Endemic	374	3.6	187	3.3	1	0.6	147	2.9	41	6.4
Het-Risk	506	4.9	197	3.5	4	2.5	181	3.5	20	3.1
Occ Exposure	2	0	1	0	0	0	1	0	0	0
NIR	464	4.5	227	4.1	13	8.2	199	3.9	41	6.4
Perl-Risk	82	0.8	43	0.8	0	0	38	0.7	5	0.8
Total	10391	100	5597	100	158	100	5114	100	641	100

5.1.6. Age Group at Diagnosis (Table 14): Note that it is easier to match in the lower age groups (newborn to age 19) as deaths in these groups from AIDS are relatively uncommon.

Table 14 - File Comparison by Age Group at Diagnosis

Age Group at Diagnosis	ACRSS 1194		"Present"		"Absent"		Initials Present		Initials Absent	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Less than 1 yr	55	0.5	27	0.5	0	0	25	0.5	2	0.3
1 to 4 years	34	0.3	18	0.3	0	0	16	0.3	2	0.3
5 to 9 years	8	0.1	4	0.1	0	0	3	0.1	1	0.1
10 to 14 years	15	0.1	5	0.1	0	0	5	0.1	0	0
15 to 19 years	37	0.4	21	0.4	0	0	18	0.3	3	0.5
20 to 29 years	1886	18.2	994	17.8	36	22.8	914	17.9	116	18.1
30 to 39 years	4549	43.8	2351	42.0	74	46.8	2154	42.1	271	42.3
40 to 49 years	2723	26.2	1516	27.1	34	21.5	1367	26.7	183	28.5
50 yrs & over	1084	10.4	661	11.8	14	8.9	612	12.0	63	9.8
Total	10391	100	5597	100	158	100	5114	100	641	100

5.1.7. Date of Onset and Survival (Tables 15 and 16): The median date of onset is earlier in the linked file than in the ACRSS file by at least nine months (January 1989 versus October 1989 overall and May 1989 versus May 1990 for cases reported after death). In both files cases that are reported after death have later onset than that seen in the overall file.

Table 15 - File Comparison by Median Date of Onset of AIDS

Median Date of Onset	ACRSS 1194	Linked File
Overall	October 1989 range=Nov 1979-Aug 1994	January 1989 range=Mar 1980-Dec 1992
When reported after death	May 1990 range=Nov 1979-Aug 1994	May 1989 range=Feb 1981-Dec 1992

Survival time was minimally shorter in the linked file compared to the ACRSS file. Those cases reported after death show a shorter survival time by six to seven months compared to overall survival times in both files (150 versus 345 days in ACRSS and 132 versus 343 days in the linked file). The ranges are shown in brackets.

Table 16 - File Comparison by Median Survival in Days

Median Survival in Days (range in brackets)	ACRSS 1194	Linked File
Overall	345 (0-5127)	343 (0-4541)
When reported after death	150 (0-5127)	132 (0-3682)

5.2. Reporting Completeness

Reporting completeness was calculated for all cases, by gender, by province of residence, by size of community at death, and by year of death. This involved dividing the number on the linked file (n_{12}) by the sum of the number on the linked file and the number on the death certificate only (DCO) file. Sekar-Deming's test for apparent dependence by stratification using year of death gave a correlation coefficient of 0.36. This was not surprising as this method does not appear to correct for very low variability within the strata of these relatively efficient sources. The calculation of this correlation coefficient is shown in Appendix D. Results are shown for Quebec alone, for the rest of Canada with "possibles" deleted, and for Canada overall. Note that the analysis at this point did not account for reporting delay. The results are shown in Table 17.

Table 17 - Reporting Completeness in Canada ("possibles" as DCO in brackets)

Geographic Region	Linked (n_{12})	DCO	n_{12}+DCO	$n_{12}/(n_{12}$+DCO)
Quebec	1614 (1480)	596 (730)	2210	73.0% (67.0)
Rest of Canada*	4117	305	4422	93.1%
Canada	5755 (5597)	901 (1059)	6656	86.5% (84.1)

*"possibles" excluded

5.2.1. Gender (Table 18): For males reporting completeness was 87.1%, for females 76.7% (84.7% and 74.9% when "possible" links were added to DCOs). This difference was less in Quebec than it was in the rest of Canada.

Table 18 - Reporting Completeness by Gender ("possibles" as DCO in brackets)

Gender	Quebec	Rest of Canada ("possibles" excluded)	Canada
Male	73.4 (67.0)	93.3	87.1 (84.7)
Female	70.0 (67.0)	86.2	76.7 (74.9)
Total	73.0 (67.0)	93.1	86.5 (84.1)

5.2.2. Province of Residence (Table 19): Reporting completeness for ACRSS by province of residence was highest in Ontario (94.4%) and Alberta (95.0%), lowest in Quebec (73.0%), P.E.I. (71.4%), and New Brunswick (75.9%). Reporting completeness in Quebec dropped from 73.0% to 67.0% when "possible" matches were counted as DCO cases. Much smaller changes were seen in the other provinces in the sensitivity analysis as seen in Table 19.

Table 19 - Reporting Completeness by Province of Residence ("possibles" as DCO in brackets)

Province of Residence	Quebec	Rest of Canada ("possibles" excluded)	Canada
Newfoundland		92.0	92.0 (92.0)
Nova Scotia		93.3	93.3 (93.3)
New Brunswick		75.4	75.9 (74.1)
P.E.I.		71.4	71.4 (71.4)
Quebec	73.0 (67.0)		73.0 (67.0)
Ontario		94.3	94.4 (93.9)
Manitoba		88.4	88.4 (88.4)
Saskatchewan		92.5	92.5 (92.5)
Alberta		94.9	95.0 (93.2)
B.C.		92.5	92.6 (92.2)
N.W.T.		0	0 (0)
Yukon		100.0	100.0 (100.0)
Other countries		21.7	21.7 (21.7)
Total		93.1	86.5 (84.1)

5.2.3. Year of Death (Table 20): Reporting completeness by year of death decreases in the most recent years but this may be due entirely to reporting delay. This does not rule out an increase in underreporting also occurring.

Table 20 - Reporting Completeness by Year of Death ("possibles" as DCO in brackets)

Year of Death	Quebec	Rest of Canada ("possibles" excluded)	Canada
82	100.0 (100.0)	80.0	92.9 (92.9)
83	100.0 (100.0)	81.2	89.7 (89.7)
84	91.2 (91.2)	97.7	94.8 (94.8)
85	91.2 (91.2)	95.9	94.4 (93.8)
86	81.5 (80.8)	94.8	90.0 (89.7)
87	80.4 (76.6)	93.7	89.4 (87.7)
88	77.7 (74.1)	94.8	88.7 (87.3)
89	75.7 (72.0)	94.9	88.2 (86.4)
90	70.9 (65.8)	93.1	85.4 (83.2)
91	62.7 (54.7)	92.5	83.2 (80.2)
92	68.9 (56.8)	90.9	84.4 (80.4)
Total	73.0 (67.0)	93.1	86.5 (84.1)

5.2.4. Community Size at Death (Table 21): Reporting completeness by community size at the time of death appears to be lower in smaller communities. For example, reporting completeness when death occurs in census metropolitan areas (CMAs) of between 9,000 and 100,000 people is only 77.4% compared to 86.5% in a CMA of over 1,000,000 population size.

Table 21 - Reporting Completeness by Community Size at Death ("possibles" as DCO in brackets)

Community Size	Quebec	Rest of Canada ("possibles" excluded)	Canada
>=1,000,000	73.3 (66.5)	94.9	86.5 (83.5)
>=500,000	77.7 (75.9)	92.4	89.0 (87.9)
>=100,000	66.2 (65.0)	91.2	88.3 (88.0)
>=9,000	59.0 (56.4)	83.6	77.4 (75.7)
Non-CMA	73.7 (63.2)	80.4	78.8 (75.3)
Total	73.0 (66.9)	93.1	86.5 (84.1)

5.3. Change in Reporting Completeness with Time

Cumulative cohorts of reporting completeness were calculated to examine the effect of reporting delay. The results are shown in Table 22 on the following page.

Table 22 - Reporting Completeness with "possibles" as Linked

AIDS death certified by	Reporting completeness in each following year								
	start	1 year	2 years	3 years	4 years	5 years	6 years	7 years	8 years
851231	75.3	84.3	87.0	87.6	89.3	92.3	93.0	94.0	94.0
860331	77.8	83.3	86.3	86.3	91.2	91.5	92.3	93.4	93.7
860630	76.5	84.3	85.8	87.2	90.7	91.2	92.3	93.4	93.6
860930	77.9	83.8	85.2	86.5	90.2	90.7	92.2	92.7	92.9
861231	76.3	81.9	83.8	86.6	89.5	90.1	91.5	91.8	
870331	75.8	82.1	83.1	86.0	88.8	89.4	91.0	91.1	
870630	75.9	81.4	83.8	87.8	88.6	90.0	90.5	90.8	
870930	75.4	81.3	83.8	87.7	88.2	89.7	90.1	90.2	
871231	74.9	81.4	85.8	88.3	88.8	90.5	90.6		
880331	76.0	81.3	87.3	88.5	89.3	90.6	90.7		
880630	74.8	82.0	87.0	88.2	89.6	90.0	90.2		
880930	75.3	81.9	86.9	87.7	89.4	89.6	89.8		
881231	76.9	84.7	87.9	88.2	89.7	89.8			
890331	77.0	85.7	87.6	88.4	89.5	89.6			
890630	77.4	85.5	87.3	89.0	89.5	89.7			
890930	77.8	85.8	87.1	88.9	89.3	89.5			
891231	79.7	86.5	87.0	88.9	89.1				
900331	80.8	85.3	86.7	88.4	88.6				
900630	79.2	85.2	87.3	88.3	88.6				
900930	79.0	85.0	87.3	88.4	88.8				
901231	81.9	84.8	87.3	88.0					
910331	78.9	84.6	86.8	87.5					
910630	79.9	84.7	86.7	87.3					
910930	78.1	84.4	86.7	87.1					
911231	78.4	85.2	86.7						
920331	79.7	85.7	86.7						
920630	80.5	85.7	86.6						
920930	80.3	85.8	86.5						
921231	81.0	85.8							
b (slope)	.008	.005	.001	-.001	-.001	-.004	-.010	-.023	-.014
Chi sq	90.52	47.61	2.89	.93	1.93	3.56	7.34	10.03	.43
p value	.000	.000	.089	.33	.16	.059	.007	.001	.51

Each column is graphed to examine the reporting completeness over time. These charts are found in Appendix C as Charts 1 to 9. There is a trend to increased reporting completeness in the first year after death certification of AIDS as the date of death certification becomes more recent. For example, 84.3% of AIDS deaths that had occurred prior to the end of 1985 were reported to ACRSS at one year after death certification. By comparison, 86.5% of AIDS deaths that had occurred prior to the end of 1989 had been reported to ACRSS one year after death certification. This trend to improved early reporting was statistically significant using chi square for trend. Late reporting, on the other hand, shows a statistically significant decrease in years six and seven after the death certification of AIDS (chi square for trend p=.007 at 6 years, p=.001 at 7 years). For example, reporting completeness six years after death certification dropped from 93.0% for the cohort which died prior to December 31, 1985 to 89.8% for the cohort which died prior to September 30, 1988. These calculations were repeated with "possible" cases transferred from the linked file to the DCO file as shown in Table 23 on the following page.

Table 23 - Reporting completeness with "possibles" as DCOs

AIDS death certified by	Reporting completeness in each following year								
	start	1 year	2 years	3 years	4 years	5 years	6 years	7 years	8 years
851231	74.9	83.9	86.6	87.3	89.0	92.0	92.6	93.6	93.6
860331	77.5	83.0	86.0	86.0	91.0	91.0	91.8	92.9	93.2
860630	76.3	84.1	85.6	86.9	90.5	90.7	91.8	92.9	93.1
860930	77.8	83.6	85.1	86.3	90.0	90.4	91.8	92.3	92.5
861231	76.2	81.8	83.6	86.5	89.2	89.8	91.2	91.5	
870331	75.6	81.7	82.7	87.6	88.2	88.9	90.5	90.6	
870630	75.5	80.9	83.3	87.2	87.9	89.3	89.9	90.1	
870930	74.8	80.6	83.0	86.8	87.3	88.8	89.1	89.3	
871231	74.2	80.7	85.0	87.4	87.8	89.5	89.6		
880331	75.2	80.5	86.5	87.6	88.4	89.7	89.8		
880630	74.0	81.1	86.1	87.2	88.5	89.0	89.1		
880930	74.5	80.9	85.9	86.6	88.2	88.4	88.7		
881231	76.0	83.7	86.8	87.1	88.6	88.7			
890331	75.9	84.6	86.5	87.2	88.2	88.3			
890630	76.4	84.4	86.2	87.8	88.2	88.4			
890930	76.8	84.7	86.0	87.7	88.0	88.2			
891231	78.7	85.3	85.8	87.6	87.8				
900331	79.8	84.1	85.5	87.0	87.3				
900630	78.2	84.0	85.9	86.9	87.2				
900930	77.8	83.6	85.8	86.9	87.3				
901231	80.6	83.5	85.8	86.5					
910331	77.6	83.2	85.2	85.9					
910630	78.5	83.1	85.0	85.6					
910930	76.6	82.7	84.9	85.3					
911231	76.8	83.4	84.8						
920331	78.0	83.8	84.7						
920630	78.5	83.5	84.4						
920930	78.2	83.5	84.2						
921231	78.8	83.5							
b (slope)	.005	.002	-.002	-.003	-.004	-.007	-.013	-.026	-.014
Chi sq	26.55	6.37	5.33	9.51	9.52	8.45	12.03	12.16	.37
p value	.000	.011	.021	.002	.002	.004	.0005	.0005	.54

These are graphed in Charts 10 to 18 of Appendix C. These calculations showed significant increases with time in reporting completeness up to one year after death certification followed by decreases in reporting completeness at years two to seven after death certification.

5.4. Non-reporting Rates

The one year non-report rates were in the 27 to 30% range in the late 1980s as shown in Table 24. The results for 1991 are lower but this may be due to large numbers of this cohort being alive. This would lead to error in the capture-recapture calculation and hence this result is likely to be less valid.

Table 24 - 1 Year Non-report Rates

Year of Dx	Low Estimate of N	High Estimate of N	ACRSS reported by year	Report in year 1	1 year non-report rate using low estimate of N	1 year non-report rate using high estimate of N
1979	-	-	1	0		
1980	6	6	6	0	1.000	1.000
1981	9	9	9	1	0.889	0.889
1982	25	25	24	14	0.400	0.400
1983	64	65	62	43	0.297	0.292
1984	162	166	160	124	0.222	0.217
1985	364	372	360	277	0.228	0.223
1986	638	663	613	430	0.287	0.276
1987	953	992	897	631	0.279	0.268
1988	1162	1195	1084	730	0.305	0.296
1989	1387	1452	1271	848	0.305	0.291
1990	1516	1603	1305	814	0.324	0.306
1991	1700	1892	1357	874	0.284	0.255

Two year, three year, and five year non-report rates are approximately 15%, 8%, and 5% in the late 1980s respectively as seen in Tables 25 to 27 on the following pages.

Table 25 - 2 Year Non-report Rates

Year of Dx	Low Estimate of N	High Estimate of N	ACRSS reported by year	Report in years 1 or 2	2 year non-report rate using low estimate of N	2 year non-report rate using high estimate of N
1979	-	-	1	0		
1980	6	6	6	0	1.000	1.000
1981	9	9	9	4	0.556	0.556
1982	25	25	24	18	0.240	0.240
1983	64	65	62	49	0.203	0.200
1984	162	166	160	137	0.142	0.139
1985	364	372	360	301	0.162	0.159
1986	638	663	613	488	0.196	0.189
1987	953	992	897	741	0.164	0.157
1988	1162	1195	1084	896	0.162	0.157
1989	1387	1452	1271	1078	0.139	0.133
1990	1516	1603	1305	1054	0.166	0.157

Table 26 - 3 Year Non-report Rates

Year of Dx	Low Estimate of N	High Estimate of N	ACRSS reported by year	Report in years 1, 2, or 3 years	3 year non-report rate using low estimate of N	3 year non-report rate using high estimate of N
1979	-	-	1	0		
1980	6	6	6	2	0.667	0.667
1981	9	9	9	5	0.444	0.444
1982	25	25	24	19	0.200	0.200
1983	64	65	62	50	0.188	0.185
1984	162	166	160	141	0.117	0.114
1985	364	372	360	316	0.121	0.118
1986	638	663	613	520	0.146	0.140
1987	953	992	897	809	0.092	0.089
1988	1162	1195	1084	979	0.090	0.088
1989	1387	1452	1271	1165	0.076	0.073

Table 27 - 5 Year Non-report Rates

Year of Dx	Low Estimate of N	High Estimate of N	ACRSS reported by year	Report in years 1 to 5	5 year non-report rate using low estimate of N	5 year non-report rate using high estimate of N
1979	-	-	1	0		
1980	6	6	6	2	0.667	0.667
1981	9	9	9	6	0.333	0.333
1982	25	25	24	21	0.120	0.120
1983	64	65	62	54	0.125	0.123
1984	162	166	160	148	0.074	0.072
1985	364	372	360	336	0.066	0.065
1986	638	663	613	575	0.060	0.057
1987	953	992	897	862	0.037	0.035

6. DISCUSSION

6.1. Representativeness and Reporting Completeness

As information on risk group and age group at diagnosis is not present on death certificates, reporting completeness could not be calculated for these variables. However, the linked file was found to be representative of the overall ACRSS file in terms of risk and age group.

Sources of variation considered were those due to record linkage, capture-recapture, and sampling error. Sensitivity analysis was used to account for the possible error due to record linkage. Error due to capture-recapture would be small due to the large numbers involved. Sampling error was not believed to be a major source of error due to the large proportion of the population captured by both ACRSS and CMDB.

Date of onset of AIDS was earlier in the linked file compared to the ACRSS file by nine to twelve months. Survival time was shorter by less than 3 weeks. These findings are in keeping with a "death bias". Those AIDS cases which are linked to a death certificate are more likely to have developed AIDS earlier in the epidemic. The earlier onset of disease is expected as those with recent onset have not died yet. The difference in survival time is not substantial but this may be due to a bias to those with more severe disease (who therefore have died).[13] It is more likely that the longer survival time in ACRSS is due to the presence of

a slightly later cohort which leads to both improved treatment and lead-time bias through earlier diagnosis. A shorter survival time found in those reported after death compared to survival in the overall file has also been noted in the UK[43]. Some possible causes for this may be physicians attempting to minimize the reporting delay of the case by systematically choosing more proximal dates of onset, underestimation of time of survival through failure to review the medical chart, or truly more severe disease such that the physician is more concerned with the medical care of the patient than taking care of an administrative matter.

Reporting completeness overall was between 84.1 and 86.5%. This variation is due to "possible" links being included as definitely linked cases or DCOs. This is surprisingly good for a passive surveillance system. This estimate of reporting completeness is more likely to be an underestimate than an overestimate as we tried to err on the conservative side in our matching.

Reporting completeness by gender, province of residence, year of death, and size of community are discussed below. Results quoted are for reporting completeness with the "possibles" included as links since the variation due to "possibles" is minor for the purposes of this discussion.

6.1.1. Gender: The proportion of females was slightly higher in the ACRSS file at 5.9% compared to 5.2% in the linked file as females made up roughly 10% of the

DCO file. The lower reporting completeness in females at 77% versus 87% in males appears to be due to two causes. It is clearly lower in Canada outside Quebec at 86.2% versus 93.3% for males. This has been noted by Calzavara in Ontario[23]. The lack of initials in Quebec since November 1990 limited our ability to link. Therefore, this finding may partially represent failure to link rather than failure to report AIDS. As Quebec has 49.7% of all female cases of AIDS but only 29.7% of total cases[21], low reporting completeness in Quebec will have a disproportionate effect on the results for all of Canada. The cause of low reporting completeness in females is unknown. One possibility is that males with AIDS tend to belong to a distinct social group who seek care from a select group of physicians. These physicians have a lot of experience with AIDS patients and hence may be more familiar with the AIDS reporting mechanism. Females with AIDS may seek care from physicians throughout the community with less experience with AIDS.

6.1.2. Province of Residence: Reporting completeness by province varied from a high of 94% in Ontario to a low of 71% in PEI. Although the results were high in Ontario, there may be a lack of independence of the two data sources as death certificates have been used to trigger the reporting of AIDS cases. This result can be contrasted with the 75% found by Johnson et al in Ontario between 1985 and 1987 at which time death certificates were not being used to remind physicians to file AIDS case reports.[18] New Brunswick and Quebec have

reporting completeness of 76 and 73% respectively. The low results in Quebec may be partially due to confounding by failure to link due to the lack of initials. It is clear from the Quebec reporting completeness results that the problem with low reporting occurred even before 1990 when the problem with lack of initials intensified. There is a problem with failure to link with Quebec data, but there is also a failure to report. There is no other comparison data for Quebec other than the abstract by Remis et al which estimated "relatively complete" AIDS case reporting in Quebec based on aggregate comparison of the number of death certificates with projected AIDS mortality in 1987 and 1988.[19] The result of 92% in B.C. is close to that found by Le et al of 88% in the linkage done between the BC AIDS Registry and members of the Vancouver Lymphadenopathy-AIDS Study[22].

6.1.3. Year of Death: Reporting completeness by year of death shows a decline in the 1990s but these values of 83 to 85% do not take reporting delay into account. A number of DCO cases will be reported assuming the median reporting delay remains at 9 months or longer. This decline alone cannot be taken as evidence of decreased reporting completeness.

6.1.4. Community Size at Death: Reporting completeness by community size at death showed a decrease as the size of the community decreased. This differs from the findings of Jones et al in South Carolina where those physicians

practicing in large urban centres were less likely to report.[44] This is unlikely to be due to failure to link as record pairs with geographic locations with small population size were easier to match. The decreased reporting in our study may be due to unfamiliarity of physicians in smaller centres about the reporting mechanism and possibly the increased concern of the social impact of contact tracing in small communities. As the reporting of AIDS cases occurs through the same authorities that are responsible for contact tracing, this could provide a disincentive to reporting.

6.2. Change in Reporting Completeness with Time

As noted earlier, by comparing reporting completeness at equivalent times after the year of death certification, trends in reporting delay and underreporting can be seen. There is a trend to improved reporting completeness within the first year after death certification. This may be due to a decrease in reporting delay or an increase in survival time (increased survival will allow more time for reporting between the time of diagnosis and death). There is also a trend to decreased reporting completeness 6 to 7 years after death certification. It is possible that increasingly long reporting delays may be responsible for this change but it seems unlikely as survival to 5 years is very uncommon with AIDS. If death does not trigger a case report, it is improbable that the case will be reported unless a death certificate review brings it to the attention of provincial health authorities. It would appear that if a case of AIDS is not reported 2 to 4

years after onset, it is becoming less likely that it will ever be reported. Increased survival time should lead to higher reporting completeness as the total time between diagnosis and a set period of time after death should increase so this should not be a factor. Therefore, it appears that underreporting (ie reporting incompleteness when corrected for reporting delay) at 5 years is increasing by 0.4 to 0.7% a year. The reasons behind this deterioration in underreporting can only be postulated. AIDS is no longer a novel condition such that some "reporting fatigue" is probably occurring. Patients with AIDS are often cared for by a small group of family physicians with previous experience with this disease. This group may be overwhelmed with the workload such that reporting is forgotten. It seems clear that qualitative studies are needed to assess the causes of reporting delay and underreporting, if they are ever to be improved.

6.3. Non-reporting Rates

The advantage of calculating non-reporting rates is that the timeliness of reporting is emphasized and comparison can be made with results elsewhere (at least in the UK at the one year mark). Regardless of whether the case is eventually reported, a time-based non-reporting rate illustrates that the information was not available within a useful period of time. The concept of reporting completeness could also be used if a time period was specified as is done with cumulative incidence. Non-report rates as reported in Tables 24 to 27 are listed by year of diagnosis. These are estimated as the date of onset for

DCO cases had to be estimated based on median survival in ACRSS and the linked file. The results for 1990 to 1991 are likely underestimates as some of these deaths would not have occurred by December 31, 1992. However, the one year non-report rates of the late 1980s of 27 to 30% compare favourably to the 28% found by Williams et al for 1989-1990 in a central London district.[7] Note that the UK does not have compulsory AIDS reporting. One could argue that Canada also does not have compulsory reporting as action is never taken against individuals who fail to report despite AIDS being a reportable disease.

6.4. Limitations to the Study

There are two sources of possible bias in this study, those due to assumptions required for the use of capture-recapture methods and bias due to the use of death certificates as the secondary data source.

6.4.1. Capture-recapture

One of the required assumptions is that the population is a closed one. While there would be no losses from either of the two databases, new entries were made constantly to ACRSS up to September 1994 and to CMDB up to December 31, 1992. The results of this study do not extend beyond 1992 so effectively the population was right censored as of the end of 1992. The addition of new entries to ACRSS with time made possible the calculation of underreporting by following cumulative death cohorts over time. The assumption of independence

of databases is not considered to be a problem when the capture rate is high as it was in this case for both databases.[34,35] As noted earlier, there is conflicting evidence from the literature showing both positive and negative dependence between death certification and AIDS notification in the UK and US.[30,31] The lack of independence in Ontario specifically was considered in the interpretation of the results. As noted in the introduction, there is no other data source to allow a true test of independence by means of log-linear modelling or other techniques. Finally, one must assume that the capture history is accurate, that individual identifiers will not be lost or overlooked. This was a problem due to the paucity of overlapping variables. The use of "possible" matches introduced this uncertainty into the results. We tried to be conservative in our matching otherwise such that the overall reporting completeness is likely to be an underestimate.

6.4.2. Error Due to the Use of Mortality Records

These sources of bias were discussed in the introduction. They will be considered in turn:

1. *Not all patients die from AIDS.* As noted previously, a U.S. study showed that less than 3% of all deaths in people reported with AIDS are due to other causes. This would be less than 6% in Canada as some of these cases were diagnosed on the basis of CD4 counts which are not in the Canadian surveillance case definition. This was partially compensated by

the use of a linkage program with independent weights from a number of variables. Although the cause of death weighed highly, there were 142 high quality matches with weights over 200 which were certified as due to other causes (AIDSFLG 3 or 5). Deaths due to causes such as motor vehicle accidents were successfully linked in provinces other than Quebec. Matched pairs with less favourable composite weights were further examined by pulling the full death certificate so that antecedent and associated conditions could be checked.

2. *Not all deaths from AIDS are coded as AIDS.* The use of the AIDSFLG coding allowed the consideration of opportunistic infections and neoplasms as well as less common AIDS-associated conditions to be considered. Ideally, one needs all the diagnoses on death certificates to be in electronic format. Statistics Canada is in the process of taking this action but this will not be available for some time.

3. *Physicians may intentionally complete certificates inaccurately.* This was a potential problem which could not be assessed as this would require access to clinical records to quantify. A local survey was considered to be of no benefit as Ottawa-Carleton has an atypical record of AIDS reporting.

4. *Physicians may complete death certificates inaccurately due to a lack of understanding of the form (eg. certification of death as due to "cardiac arrest").* This was seen to be a problem in links which were matched on

the basis of other overlapping fields. It is likely that other links were missed due to this problem when there were errors in other fields such that the composite weight did not lead to verification of death diagnosis with the full death certificate.

5. *The lack of independence between ACRSS and the CMDB in Ontario was a problem which was accounted for in the discussion of results.* Apparently ACRSS cases which were produced in response to a death certificate were flagged in Toronto but it is unclear how many of these cases would have been reported in the absence of intervention. There is insufficient information on a death certificate alone to allow completion of an ACRSS report form.

6. *Some Canadians with AIDS die in other countries.* We had 23 deaths from outside Canada, 18 were DCO and 5 linked. Nearly all of these were from the U.S. Most American states forward death certificates back to Canada, those include New York and California. This would not be expected to be a large source of error.

7. *The results clearly show a selection bias or "death bias" as linked cases were diagnosed earlier than those on ACRSS overall.* Survival was minimally shorter and probably of no importance. If there had been a dramatic change in the pattern of AIDS in Canada as occurred in the U.S. (increase in IDU), this bias may have been significant. In the absence of a profound change in AIDS epidemiology[2], the results of this linkage are

valid although the cases were diagnosed earlier than most in ACRSS.

8. *There may be a concern that the measurement is of underreporting of death due to AIDS and not underreporting of AIDS.* It is clear that underreporting of death to ACRSS is not unimportant - 10.5% of the linked file were listed as alive on ACRSS. However, this situation was known although not quantified at the time of the linkage. Vital status and date of death were only two of the fields compared. A good match on the other fields generally led to a successful link.

9. *Insufficient data (eg lack of full names) on the two databases made linkage difficult, especially in the province of Quebec.* There is no doubt that the low reporting completeness calculated for that province is partially due to failure to link rather than failure to report. Unfortunately, this situation will only become worse with the passage of time. This raises the question of where one sets the balance in protecting the individual while attempting to protect the general public by monitoring changes in AIDS epidemiology.

10. *AIDS on the death certificate may not meet the case definition required by the surveillance system.* This may have been a problem with the earlier case definitions but would be less of a problem at present. Without doing a chart review, this source of bias cannot be quantified. It seems unlikely that physicians would certify AIDS as the cause of death in the absence of good clinical evidence due to the social stigma associated with this

disease.

In summary, there are numerous sources of potential bias in this study. These were minimized as much as possible through careful linkage. The error induced due to these forms of bias cannot be quantified, however the emphasis on conservative matching would suggest that the reporting completeness estimates presented in this paper would be underestimates rather than overestimates.

7. CONCLUSIONS

Reporting completeness to the AIDS Case Reporting Surveillance System is about 85% which compares well with that seen in other countries. Reporting completeness is lower for females than males. Reporting completeness varies by province with Quebec, PEI, and New Brunswick having the lowest rates of reporting. Alberta and Ontario have the highest rates of reporting. Those who die in smaller communities are less likely to be reported. Reporting completeness decreases slightly during the 1980s. The most important finding is the trend to increased underreporting with time which is seen when the effect of reporting delay is controlled for in the analysis. One year non-reporting rates are comparable with those seen in the U.K. Future validations of reporting completeness will become more difficult in Quebec due to the lack of initials on their dataset.

The causes of incomplete reporting will only be elucidated with some qualitative investigation.

8. REFERENCES

1.Remis RS, Sutherland WD. The epidemiology of HIV and AIDS in Canada: current perspectives and future needs. CJPH 1993;84(Supp 1):S34-S38.

2. Quarterly Surveillance Update: AIDS in Canada. April 1995. LCDC, Health Canada.

3. Calzavara LM, Coates RA, Craib KJP, et al. Underreporting of AIDS cases in Canada: a record linkage study. CMAJ 1990;142:36-39.

4. Revision of the surveillance case definition for AIDS in Canada. CCDR 1993;19(15):116-7.

5. Soskolne CL. Proportion of new AIDS case reports attributable to the 1987 CDC revised surveillance case definition: Canada-United States differences. Can J Public Health 1989;80:380-381.

6. Evans BG. Estimating underreporting of AIDS: straightforward in theory - difficult in practice. AIDS 1991;5:1261-1262.

7. Williams IG, Shergold BE, Beecham M, et al. Auditing AIDS reporting. Experience from a central London district 1982-1991. Genitourin Med 1992;68:390-393.

8. Hickman M, Aldous J, Gazzard B, et al. AIDS surveillance: a direct assessment of under-reporting. AIDS 1993;7:1661-1665.

9. Undercount of Cases and Lack of Key Data Weaken Existing Estimates: Report to Congressional Requesters. Washington,DC: US General Accounting Office;1989.

10. Chu SY, Hanson DL, Fanzo KM,et al. Capturing HIV-related mortality: effect of the 1993 expanded AIDS surveillance case definition. Ninth International Conference on AIDS, Berlin 1993. Abstract no. WS-C01-5.

11. Rushworth RL, Bell SM, Rubin GL, et al. Improving surveillance of infectious disease in New South Wales. Med J Aust 1991;154;828-831.

12. Fife D, MacGregor RR, McAnaney J. Limitations of AIDS reporting under favorable circumstances. Am J Prev Med 1993;9:317-20.

13. Gertig DM, Marion SA, Schechter MT. Estimating the extent of underreporting in AIDS surveillance. AIDS 1991;5:1157-1164.

14. Quarterly Report No. 23-30 Sept 1989. European Centre for the Epidemiological Monitoring of AIDS.

15. Rosenblum L, Buehler JW, Morgan MW, et al. The completeness of AIDS case reporting, 1988:a multisite collaborative surveillance project. Am J Public Health 1992;82(11):1495-1499.

16. Losos J, Wells G, Elmslie K, et al. Acquired immune deficiency in Canada: the first 5 years of surveillance. CMAJ 1988;139:383-388.

17. Wells GA, Tostowaryk W, Rylett DT. AIDS: A perspective for Canadians. Background papers, Royal Society of Canada. 1988:81-110.

18. Johnson RJ, Montano BL, Wallace EM. Using death certificates to estimate the completeness of AIDS case reporting in Ontario in 1985-87. CMAJ 1989;141:537-540.

19. Remis RS, Palmer RWH. Modelling AIDS mortality from survival analysis to evaluate completeness of reporting. Abstract WC97, VII International Conference on AIDS, June 1991.

20. Ricketts MN. Validation of AIDS Data in Canada. Assessment by Division of HIV/AIDS Epidemiology, LCDC. International Society for Epidemiology, England, 1992.

21. Quarterly surveillance update: AIDS in Canada. January 1995. LCDC, Health Canada.

22. Le LN, Strathdee SA, Craib KJ, et al. Underreporting of AIDS cases in British Columbia. International Conference on AIDS 1994;10(1):325 (Abstract no. PC0230).

23. Calzavara L, Strathdee S, Ricketts M, et al. Reported and unreported female AIDS cases in Ontario: how many and who. Fifth Canadian Conference on HIV/AIDS Research. Winnipeg, Man., June 1995, Abstract 305.

24. Munoz A, Wang M, Bass S, et al. Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men. Am J Epidemiol 1989;130:530-9.

25. Quarterly Surveillance Update: AIDS in Canada. October 1994. LCDC, Health Canada.

26. Modesitt SK, Smyser M, Hopkins SG, et al. Effects of increasing outpatient diagnosis on AIDS surveillance. J AIDS 1993;6:91-4.

27. Marzuk PM, Tierney H, Tardiff K, et al. Increased risk of suicide in persons with AIDS. JAMA 1988;259:1333-7.

28. Chu SY, Buehler JW, Lieb L, et al. Causes of death among persons reported with AIDS. Am J Public Health 1993;83(10):1429-32.

29. Personal communication, Pat Wood, nosologist, Statistics Canada, 1996.

30. Hayes T, Altman R, Akili-Obika A, et al. HIV-related deaths from selected infectious diseases among persons without AIDS in New Jersey. J AIDS 1994;7(10):1074-8.

31. Bobby JJ, Spencer PD, Wyatt JC, et al. AIDS deaths in the UK - how complete are the figures? Public Health 1988;102:519-524.

32. Personal communication, Helen Bangura, Ontario Ministry of Health, 1996.

33. McCarty DJ, Tull ES, Moy CS. Ascertainment corrected rates: applications of capture-recapture methods. Int J Epid 1993;22:559-565.

34. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing case ascertainment when using multiple information sources. J Chron Dis 1974;27:25-36.

35. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. Epid 1995;6(1):42-48.

36. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration. Amer Stat Ass J 1949;44:100-115.

37. Hubert B, Desenclos JC. Evaluation of the exhaustivity and representativeness of a surveillance system by the capture-recapture method. Rev Epidem et Sante Publ 1993;41:241-49.

38. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. Am J Epidemiol 1995;142:1047-58.

39. Newcombe HB. Handbook for record linkage. Methods for health and statistical studies, administration, and business. Oxford, England: Oxford University Press,1988.

40. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Stat Med 1995;14:499-509.

41. Jaro MA. Probabilistic linkage of large public health data files.Stat Med 1995;14:491-498.

42. Centre for Disease Control. Human immunodeficiency virus (HIV) infection classification. MMWR 1987;36(Supp 7):1-20.

43. Whitmore-Overton SE, Tillett HE, Evans BG, et al. Improved survival from diagnosis of AIDS in adult cases in the United Kingdom and bias due to reporting delays. AIDS 1993;7(3):415-20.

44. Jones JL, Meyer P, Garrison C, et al. Physician and infection control practitioner HIV/AIDS reporting characteristics. Am J Public Health 1992;82:889-891.

APPENDIX A

Figure 1 -Bimodal Distribution of Composite Weights

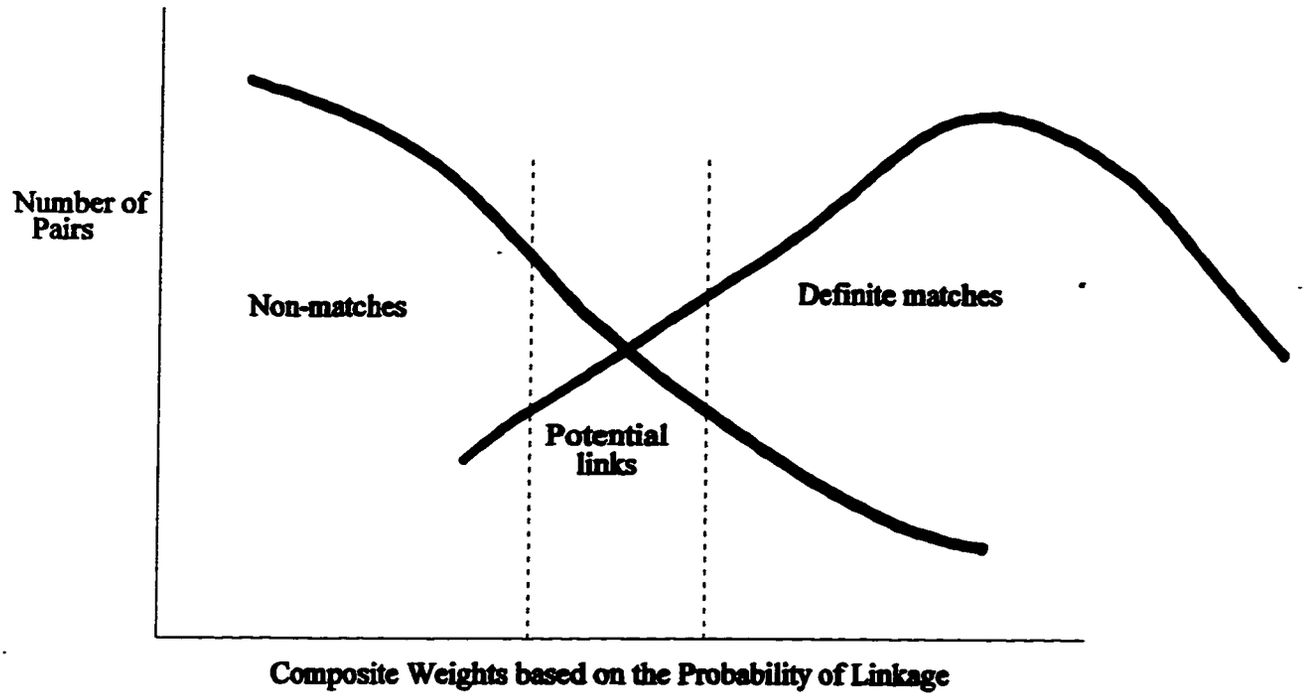
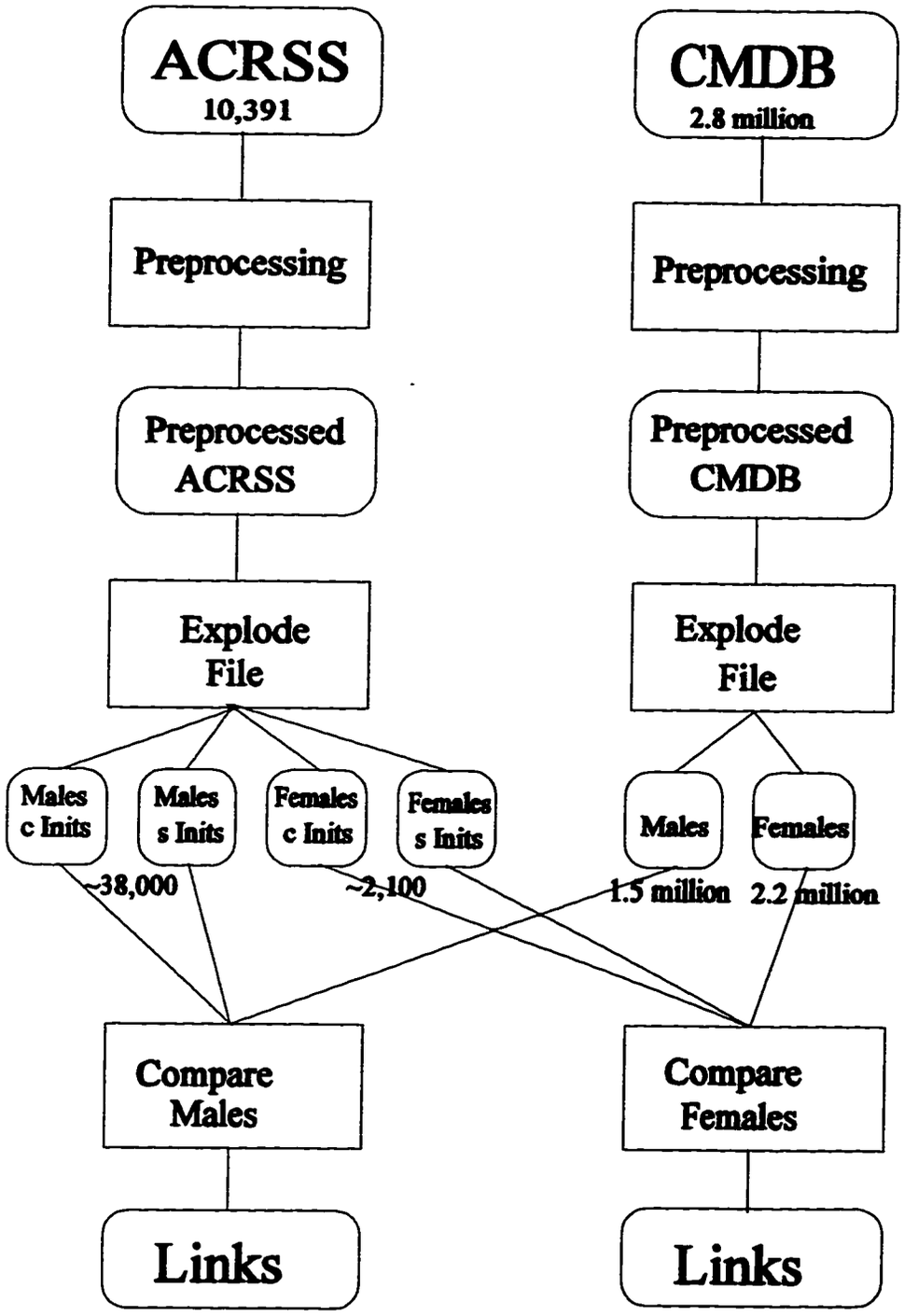


Figure 2 - Linkage Process



APPENDIX B

Table 1 - New Fields Created when Preprocessing ACRSS

Field Name	Modification
SEQNO	Sequence number assigned by STC
DUPFLAG	Flags possible duplicate records
SASDOB	SAS date of birth
SASDOD	SAS date of death
LTONSET	Date of diagnosis of last onset (YYMM)
APROV2	Attributed province (vital STC codes)
CPROV2	Province of residence (vital STC codes)
DPROV2	Province of diagnosis (vital STC codes)
BIRTHPL2	Country of birth using STC codes
CCITYA	City of residence by name
DCITYA	City of diagnosis by name
HCITYA	City of hospital by name
CSGC86	City of residence by SGC86 code
DSGC86	City of diagnosis by SGC86 code
HSGC86	City of hospital by SGC86 code
CSGC91	City of residence by SGC91 code
DSGC91	City of diagnosis by SGC91 code
HSGC91	City of hospital by SGC91 code
POSTAL3	FSA postal code (Ontario only)
SASLAST	SAS date for last onset
POCKET	Number of pocket value
POCKET1	Pocket (block) for the first pass: birth date (YYMMDD)
POCKET2	Pocket value for the second pass: death date (YYMMDD)
POCKET3	Pocket value for the third pass: last initial, first initial, birth year (LFYY)
POCKET4	Pocket value for the fourth pass: last initial, birth month and day (LMMDD)
POCKET5	Pocket value for the fifth pass: last, first, death year and month (LFYYMM)
FILLER	

Table 2 - Preprocessed ACRSS File

Name	Information	Name	Information
SEQNO	SEQUENCE # ASSIGNED BY STC	DUPFLAG	DUPFLAG = '1'
EPICNO	ID #, ASSIGNED BY LCDC	PROVID	ID#, ASSIGNED BY PROVINCE
LAST	LAST INITIAL OF NAME	FIRST	FIRST INITIAL OF NAME
SECOND	MIDDLE INITIAL OF NAME	SEX	SEX
FULLDOB	FULL DATE OF BIRTH (YYMMDD)	BIRTHYR	YEAR OF BIRTH
SASDOB	SAS DATE OF BIRTH	SASDOD	SAS DATE OF DEATH
FULLDOD	FULL DATE OF DEATH (YYMMDD)	VITSTAT	VITAL STATUS
APROV	ATTRIBUTED PROVINCE	CPROV	CURRENT PROVINCE OF RESIDENCE
DPROV	PROVINCE OF DIAGNOSIS	CCITY	CURRENT CITY OF RESIDENCE
DCITY	CITY OF DIAGNOSIS	HOSPCITY	CITY WHERE HOSPITAL LOCATED
COUNTRY	COUNTRY OF BIRTH	IMMDATE	DATE OF IMMIGRATION (YYMM)
FTONSET	DATE OF DIAG OF EARLIEST ONSET(YYMM)	LTONSET	DATE OF DIAG OF LAST ONSET (YYMM)
DIAG1	DIAGNOSIS 1	DIAG2	DIAGNOSIS 2
DIAG3	DIAGNOSIS 3	DIAG4	DIAGNOSIS 4
APROV2	ATTRIBUTED PROV (VITAL STC CODES)	CPROV2	PROV OF RESIDENCE (VITAL CODES)
DPROV2	PROV OF DIAGNOSIS (VITAL STC CODES)	BIRTHPL2	COUNTRY OF BIRTH (STC CODES)
CCITYA	CITY OF RESIDENCE - NAME	DCITYA	CITY OF DIAGNOSIS - NAME
HCITYA	CITY OF HOSPITAL - NAME	CSGC86	CITY OF RESIDENCE - SGC86 CODE
DSGC86	CITY OF DIAGNOSIS - SGC86 CODE	HSGC86	CITY OF HOSPITAL - SGC86 CODE
CSGC91	CITY OF RESIDENCE - SGC91 CODE	DSGC91	CITY OF DIAGNOSIS - SGC91 CODE
HSGC91	CITY OF HOSPITAL - SGC91 CODE	POSTAL3	FSA POSTAL CODE (ONTARIO ONLY)
YREC	YEAR RECEIVED	YRDIAG	YEAR DIAGNOSED
SASLAST	SAS DATE FOR LAST ONSET	POCKET	# ll POCKET VALUE (#=POCKET NUMBER)
POCKET1	POCKET1 VALUE: BIRTH DATE(YYMMDD)	POCKET2	POCKET2 VALUE: DEATH DATE (YYMMDD)
POCKET3	POCKET3 VALUE: LAST, FIRST, BYY	POCKET4	POCKET4 VALUE: LAST, BMMDD
POCKET5	POCKET5 VALUE: LAST, FIRST, DYYMM	FILLER	

Table 3 - Preprocessed CMDB File

Name	Information	Name	Information
CCD	CONTROL CODE DIGIT	DUPFLAG	DUPLICATE FLAG
SEQUENCE	SEQUENCE NUMBER	DTHYEAR	DEATH YEAR
DTHPROV	PROVINCE OF DEATH	DTHREGNO	REGISTRATION NUMBER
NYSIIS	NYSIIS CODE OF SURNAME	SURNAME	SURNAME ('LAST'=1ST LETTER)
GIVEN1	FIRST GIVEN NAME('FIRST'=1ST INITIAL)	GIVEN2	2ND GIVEN NAME ('SECOND'=2ND INIT)
BIRTHYR	BIRTH YEAR EG:1942 OR 0 IF N/A	BYEAR	BIRTH YEAR EG:942
BIRTHMN	BIRTH MONTH	BIRTHDY	BIRTH DAY
SEX	SEX CODE (1,2)	MARSTAT	MARITAL STATUS (1-5)
BIRTHPL	BIRTH PLACE CODE	DTHPLACE	PLACE OF DEATH (CITY)
OCCNTY	OCCURRENCE-COUNTY OR CENSUS DIV	OCCULOC	OCCURRENCE-LOCALITY
RESPROV	RESIDENCE - PROVINCE	RESCDIV	RESIDENCE - COUNTY OR CENSUS DIV
RESCSUB	RESIDENCE-LOCALITY OR CENSUS SUBD	FATHNAM	FATHER'S NAME
AIDSFLG	AIDS FLAG OF CAUSES (1-5)	CAUSE	CAUSE OF DEATH - ICD CODE
DTHDATE	DEATH DATE (YYMMDD)	SASDOD	SAS DATE OF DEATH
SASDOB	SAS DATE OF BIRTH	POCKET	PASS II POCKET VALUE (1,2,3,4,5)

Table 4 - Comparison Rules

Rule	Description
FILSIZE	Assigns a negative weight based on the size of the file being searched.
YLKA	Death date on CMDB is compared to the dates of first and last onset, year of immigration, and death year from ACRSS. It is expected that the death date of CMDB is greater or equal to the dates on ACRSS.
BIRTHYR	Checks for level of agreement on birth year in both files.
AGEB	Age at death is calculated using birth year and death year in CMDB, likelihood of death at each age is incorporated into weight.
BIRTHMN	Checks for level of agreement on birth month in both files.
BIRTHDY	Checks for level of agreement on birth day in both files.
XDATES	Checks for transposition of day and month of birth.
DTHDATE	Checks for agreement of date of death if present on both records.
NONAME	Identifies links with no names and having date conflicts via rule YLKA.
SURNAME	Checks for agreement of first initial of surname or father's surname.
INIT	Checks for agreement of initials and transposition of initials on both files.
RESID1	Checks for agreement of place of diagnosis, city of residence, and hospital location on ACRSS to place of death and place of residence on CMDB. It determines the best agreement.
RESID2	Same as RESID1 but determines second best agreement.
RESCITY	Checks for agreement of city of diagnosis, city of residence, and hospital city on ACRSS with city of death on CMDB.
BIRTHPL	Checks for agreement on country of birth.
AIDSFLG	Additional weight given for causes of death indicative of AIDS.
DIAG1	Checks for agreement of the first recorded diagnosis in ACRSS with cause of death on CMDB.
DIAG2	Checks for agreement of the second recorded diagnosis in ACRSS with cause of death on CMDB.
DIAG3	Checks for agreement of the third recorded diagnosis in ACRSS with cause of death on CMDB.
DIAG4	Checks for agreement of the fourth recorded diagnosis in ACRSS with cause of death on CMDB.
The following rules did not contribute weighting to the matches: VITSTAT, APROV, CANCEL, PASS, and REGNO.	

Table 5 - Results of Manual Review of 10% of Males with Initials Match (n=511)

Cumulative Weights	Percent of Total Links	"Probable" Matches (%)	"Possible" Matches (%)	Non Matches (%)	% "Probable" of Total
500-200	79.1	79.1			100%
199-180	3.7	3.7(?2DC)			100%
179-160	3.7	3.7(?4DC)			100%
159-140	2.5	2.5(?1DC)			100%
139-120	1.6	1.2	0.4		75%
119-110	0.6	0.6			100%
109-100	0.6	0.4	0.2		67%
99-90	0.6	0.2	0.4		33%
89-80	0.6	0.2	0.4		33%
79-70	0.4	0.4			100%
69-50	0.6	0.2	0.4		33%
49-30	0.6	0.2	0.2	0.2	33%
29-0	0.6	0.2	0.4		33%
-1-(-20)	0.6		0.6		0%
-21-(-40)	0.6			0.6	0%
-41-(-60)	1.0		0.2	0.6	0%
-61-(-80)	1.0			1.0	0%
-81-(-100)	1.8		0.2	1.6	0%
Total	100.0% of 511	92.8%	3.3%	3.9%	92.8%
DC indicates death certificates would need to be pulled to confirm "probable" status					

Table 6 - Grouping of "Possible" Matches (males with initials)

Category of Match	Number of Matches
DOB match DOD missing first initial match but not the second	2
DOB non-match DOD match initials match low prevalence city	1
DOB non-match DOD match initials match high prevalence city	8
DOB match DOD missing initials match AIDSFLG 5	9
DOB match DOD match initials non-match	3
DOB different but possible DOD missing initials match	2
DOB non-match DOD missing initials match born outside of Canada	1
DOB partially missing (unusual) DOD match initials match	1
DOB match DOD non-match initials match AIDSFLG 3	1

APPENDIX C

Table 1 - Reporting Delay by Years after Diagnosis by Year of Diagnosis

Year of Diagnosis	Years after Diagnosis						
	1	2	3	4	5	6 to 10	Other
1979	0	0	0	0	0	1	0
1980	0	0	2	0	0	2	2
1981	1	3	1	1	0	2	1
1982	14	4	1	1	1	3	0
1983	43	6	1	2	2	7	1
1984	124	13	4	3	4	12	0
1985	277	24	15	8	12	24	1
1986	430	58	32	37	18	38	4
1987	631	110	68	28	25	35	2
1988	730	166	83	49	35	21	4
1989	848	230	87	67	36	3	0
1990	814	240	160	75	16	0	1
1991	874	317	139	27	0	0	3
1992	1079	309	66	0	0	0	6
Total	5865	1480	659	298	149	148	25

Table 2 - Estimate of Total Numbers of AIDS Cases by Year of Diagnosis using Overall Survival Times Estimated for DCO Cases and Including "Possibles" with Linked File

Year of Dx	ACRSS total (n_2)	Linked total (n_{12})	DCO total using median survival	CMDS total (n_1) = n_{12} + DCO	$N=n_1n_2/n_{12}$
1979	1	0	0	0	
1980	6	2	0	2	6.0
1981	9	4	0	4	9.0
1982	24	20	1	21	25.2
1983	62	55	3	58	65.4
1984	160	136	4	140	164.7
1985	361	316	9	325	371.3
1986	617	555	34	589	654.8
1987	899	805	62	867	968.2
1988	1088	927	74	1001	1174.9
1989	1271	1009	108	1117	1407.0
1990	1306	858	145	1003	1526.7
1991	1360	699	222	921	1791.9
1992	1460	364	226	590	2366.5

Chart 1: Reporting Completeness at end of Year of Death - "Possibles" as Linked

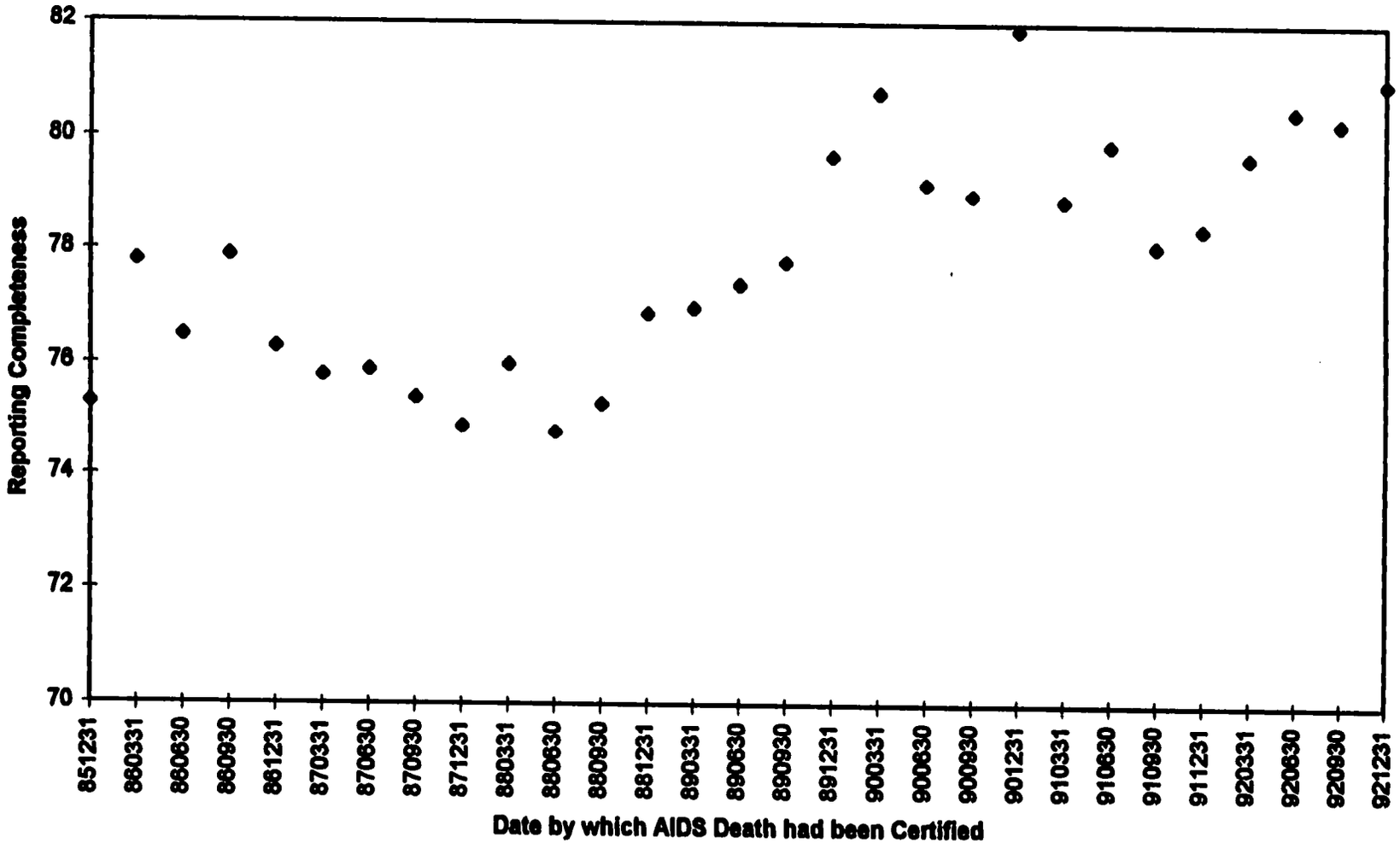


Chart 2: Reporting Completeness at Year 1 - "Possibles" as Linked

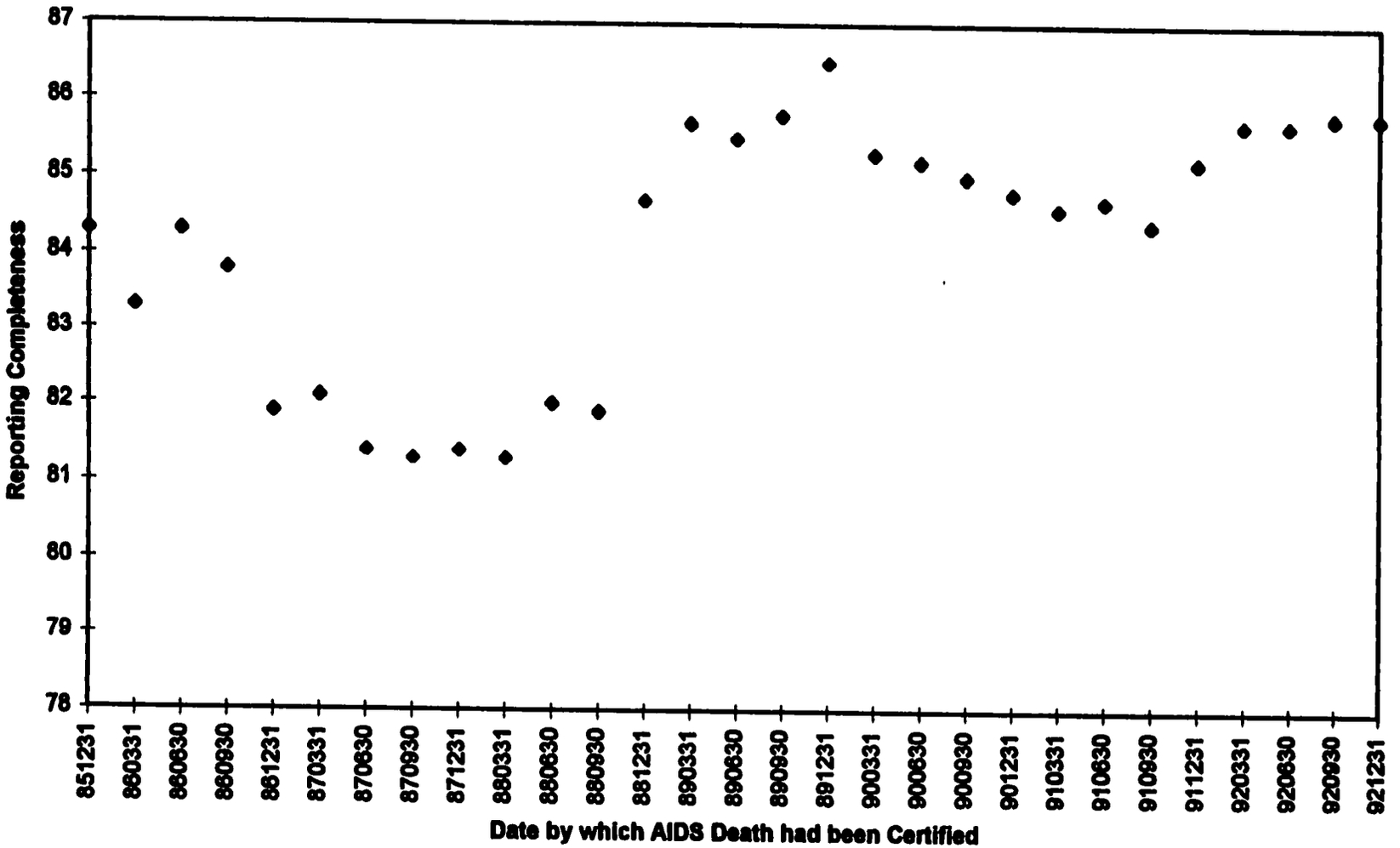


Chart 3: Reporting Completeness at Year 2 - "Possibles" as Linked

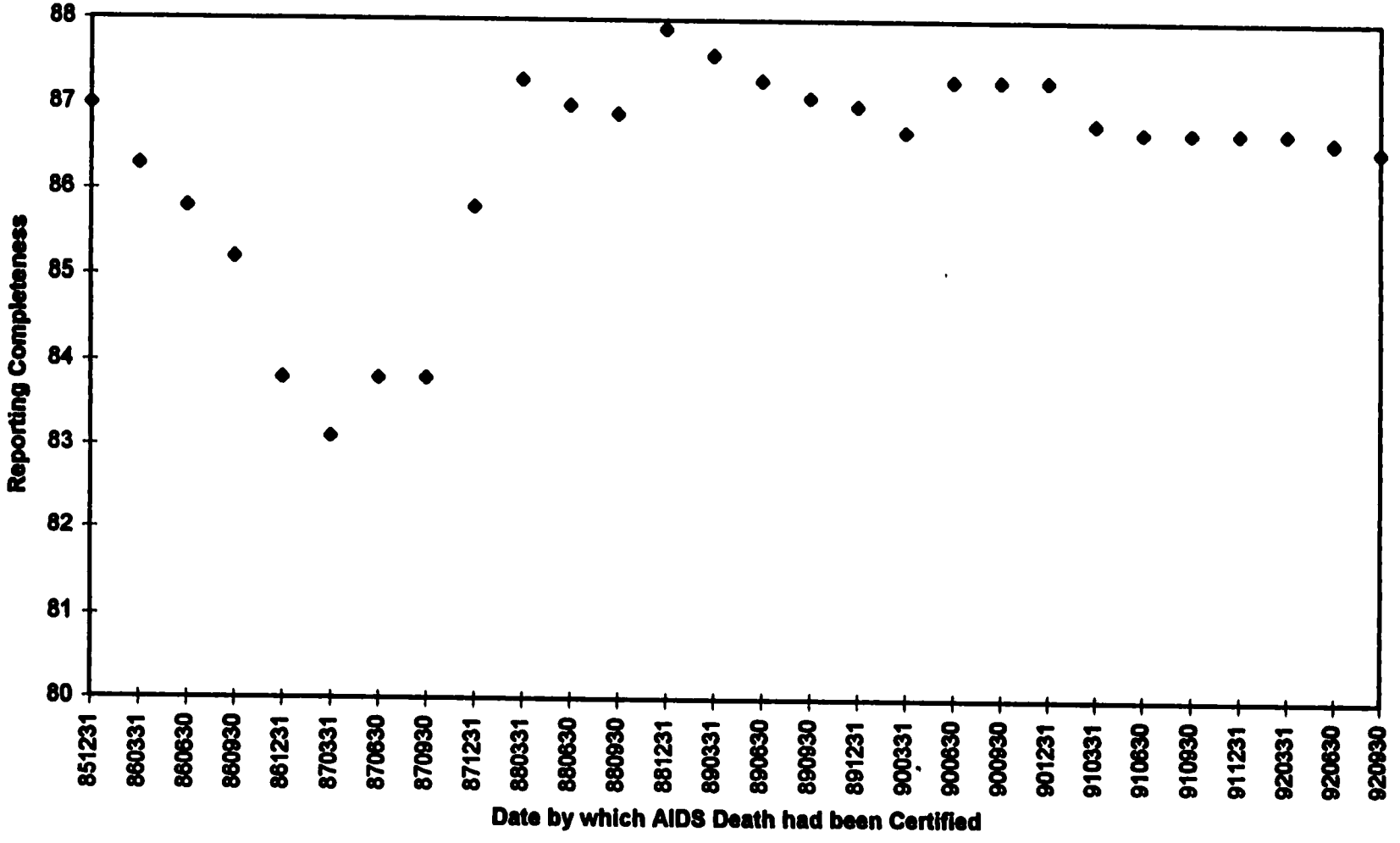


Chart 4: Reporting Completeness at Year 3 - "Possibles" as Linked

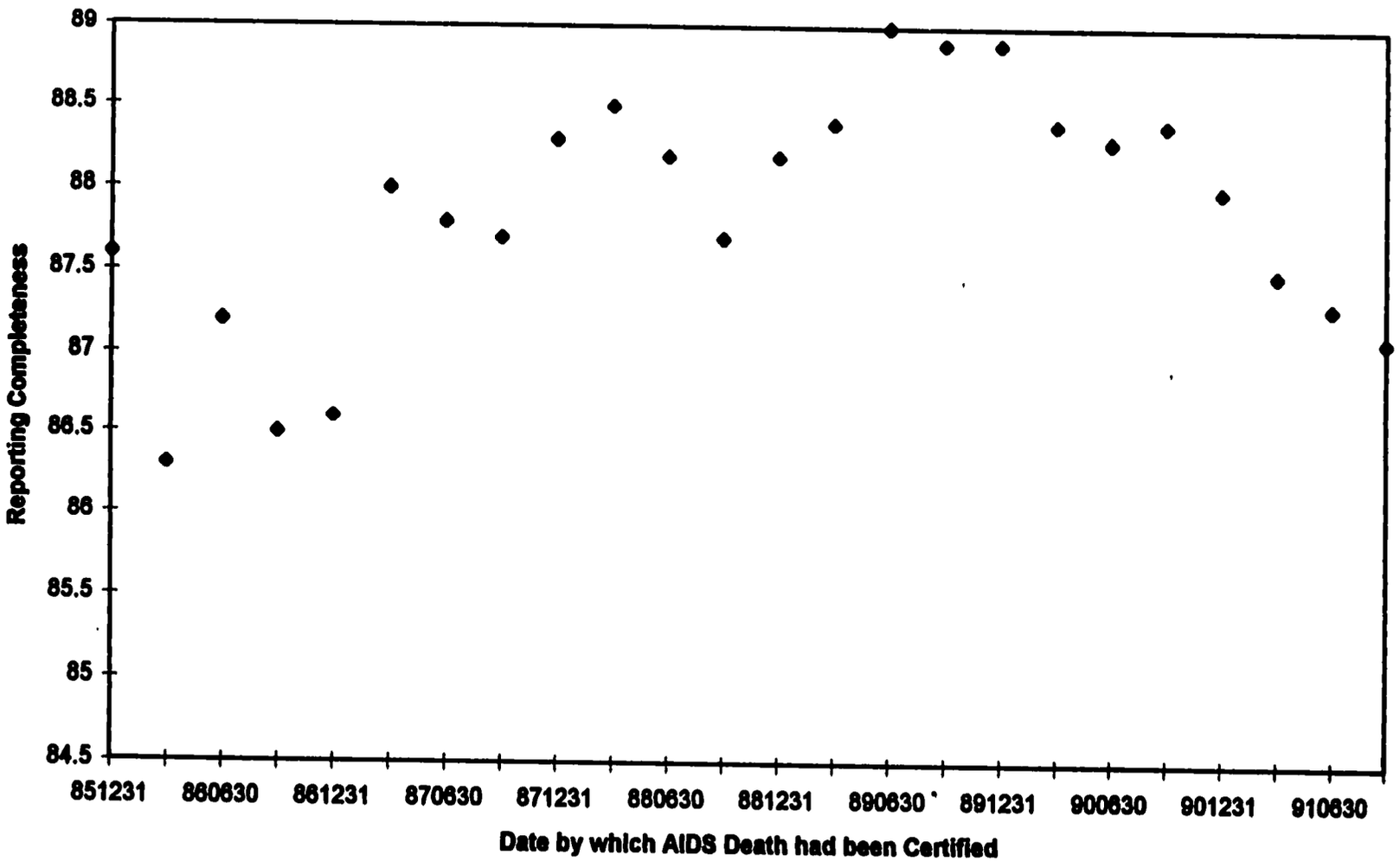


Chart 5: Reporting Completeness at Year 4 - "Possibles" as Linked

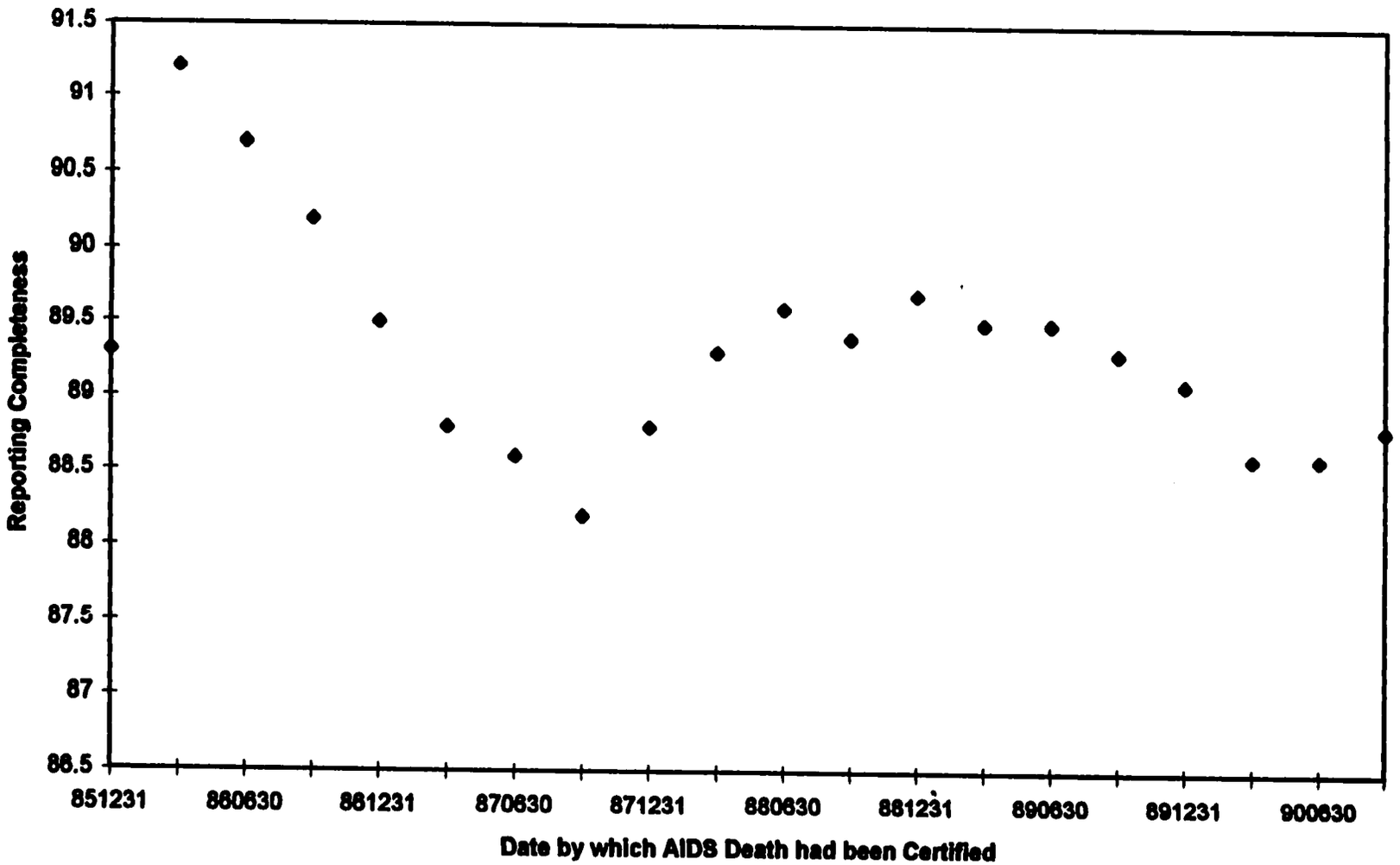


Chart 6: Reporting Completeness at Year 5 - "Possibles" as Linked

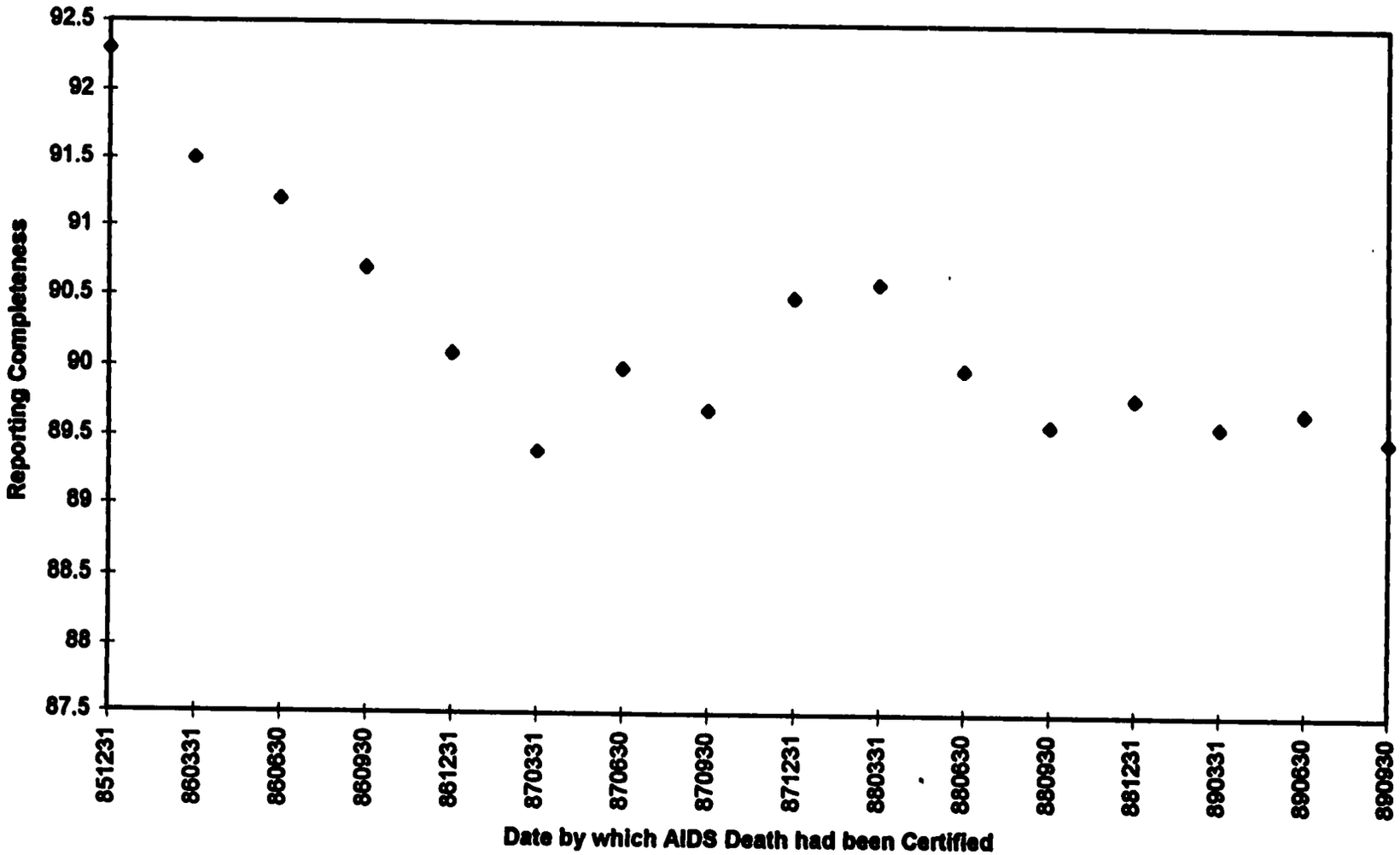


Chart 7: Reporting Completeness at Year 6 -"Possibles" as Linked

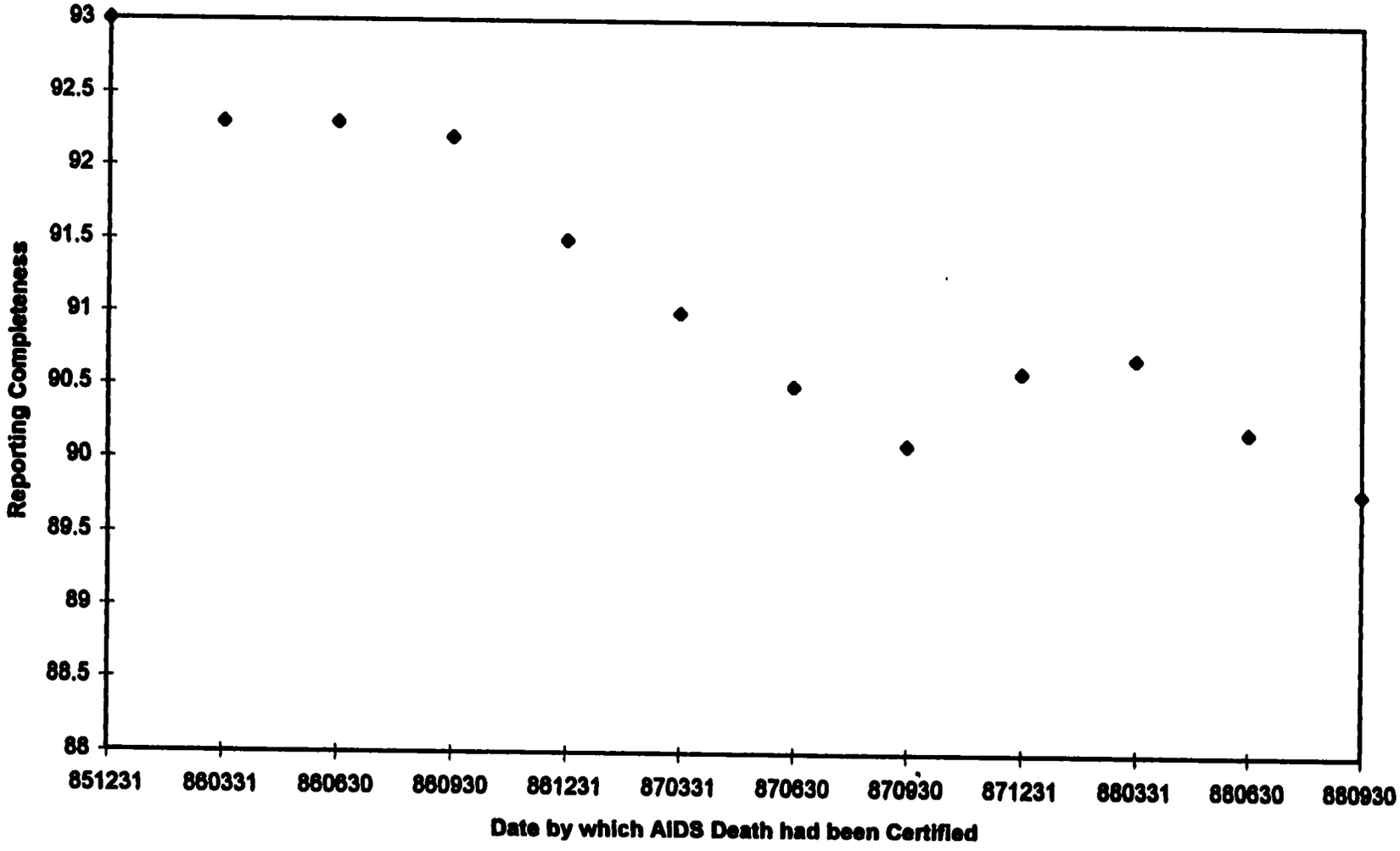


Chart 8: Reporting Completeness at Year 7 -"Possibles" as Linked

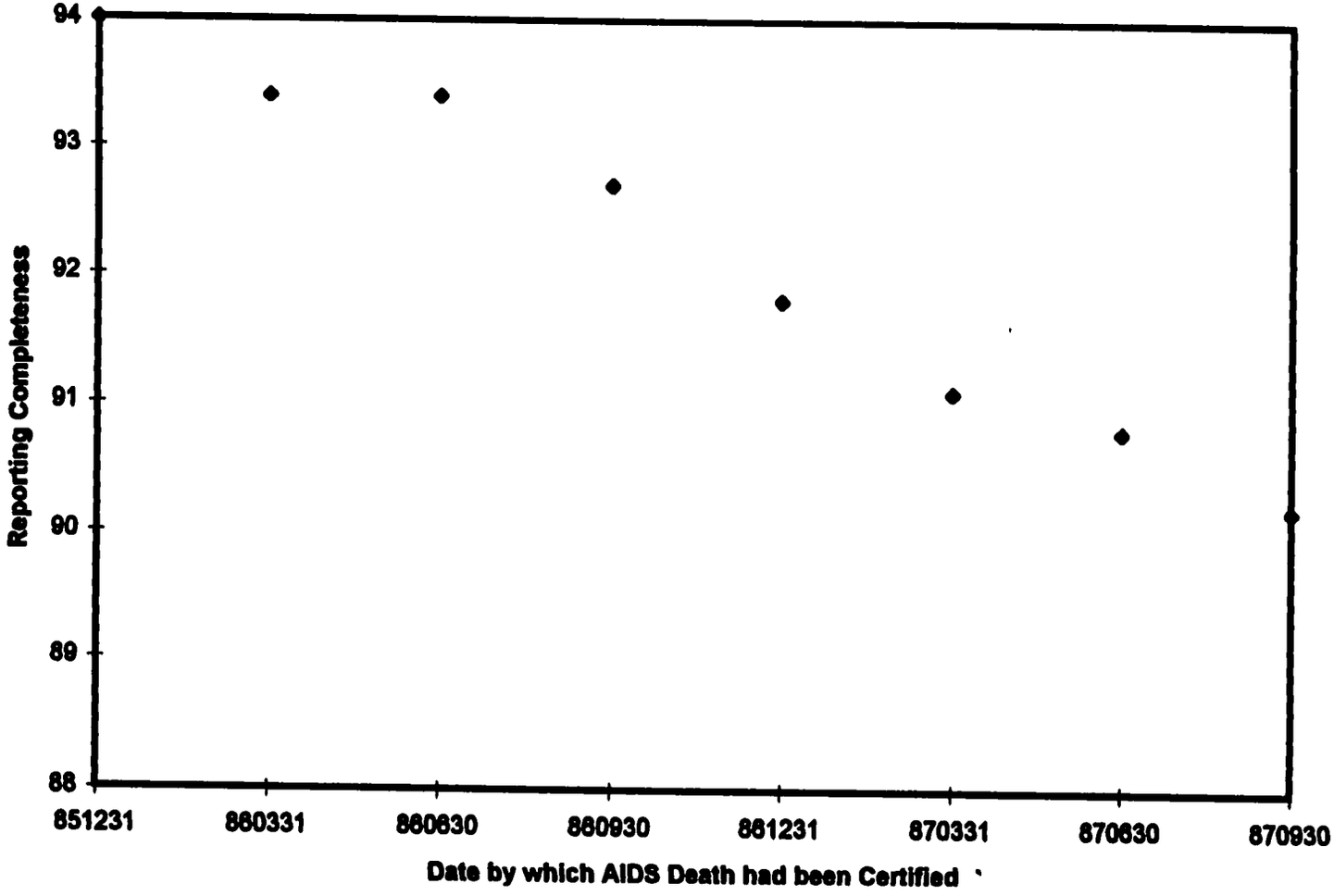


Chart 9: Reporting Completeness at Year 8 -"Possibles" as Linked

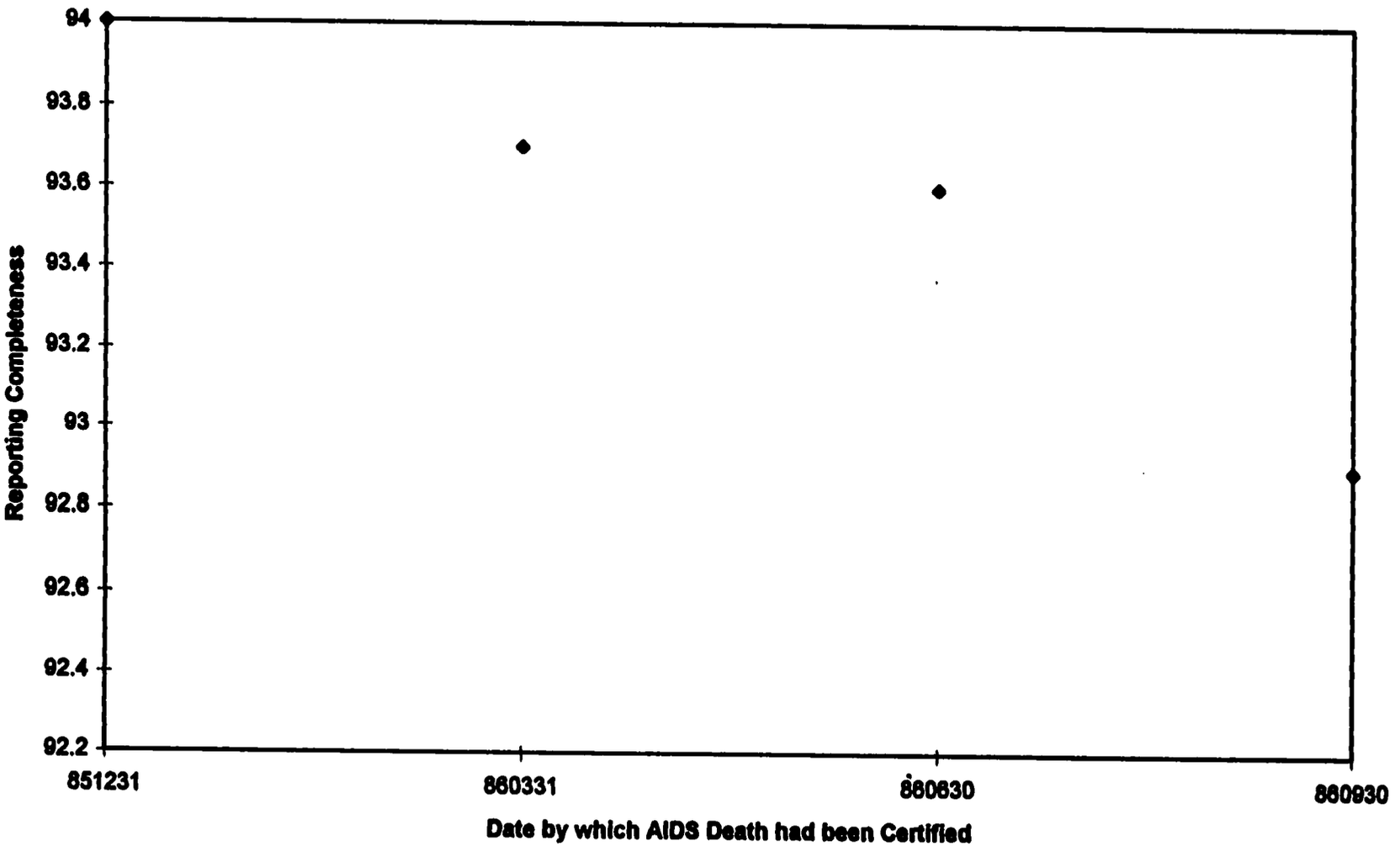


Chart 10: Reporting Completeness at end of Year of Death - "Possibles" as DCO

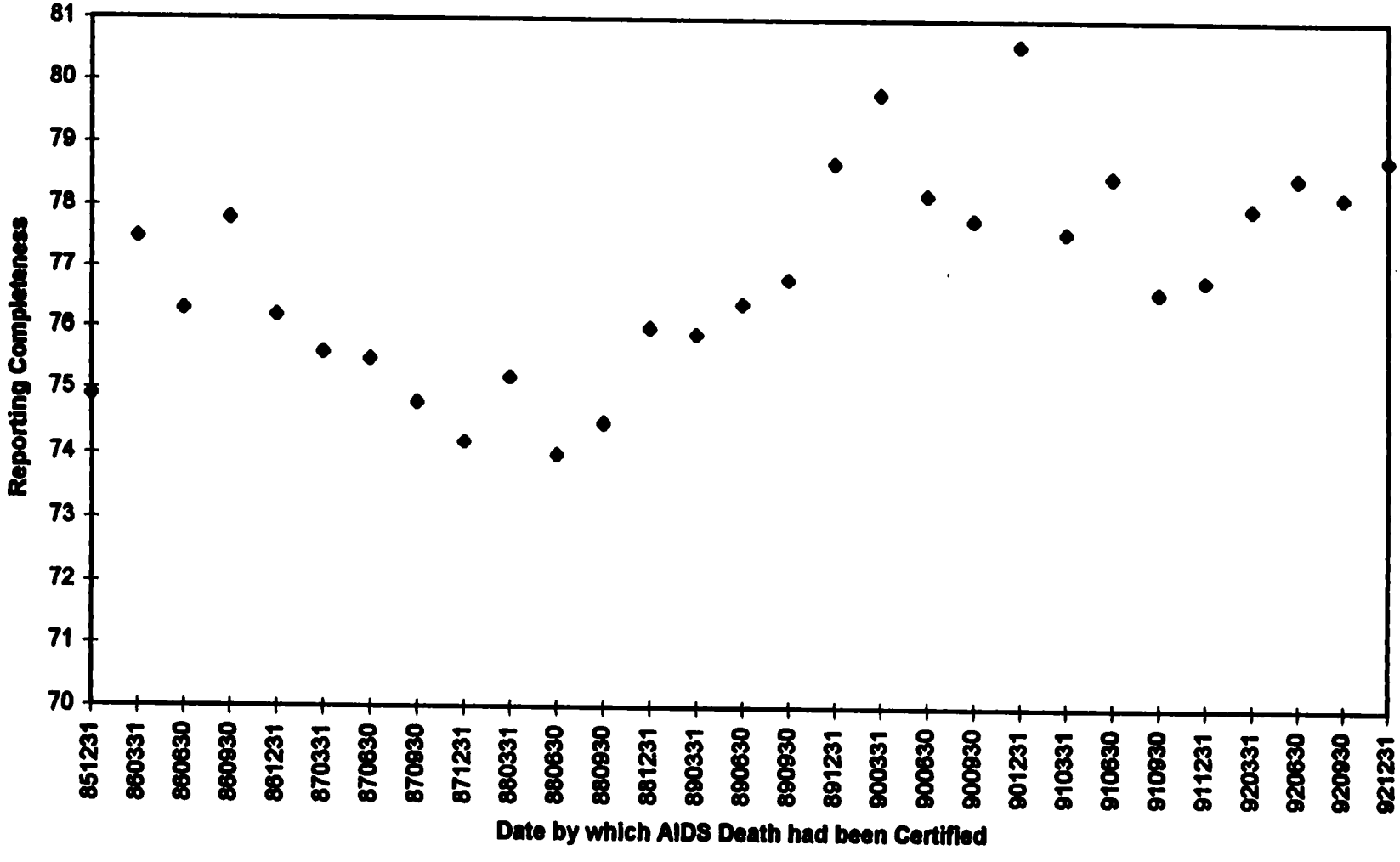


Chart 11: Reporting Completeness at Year 1 - "Possibles" as DCO

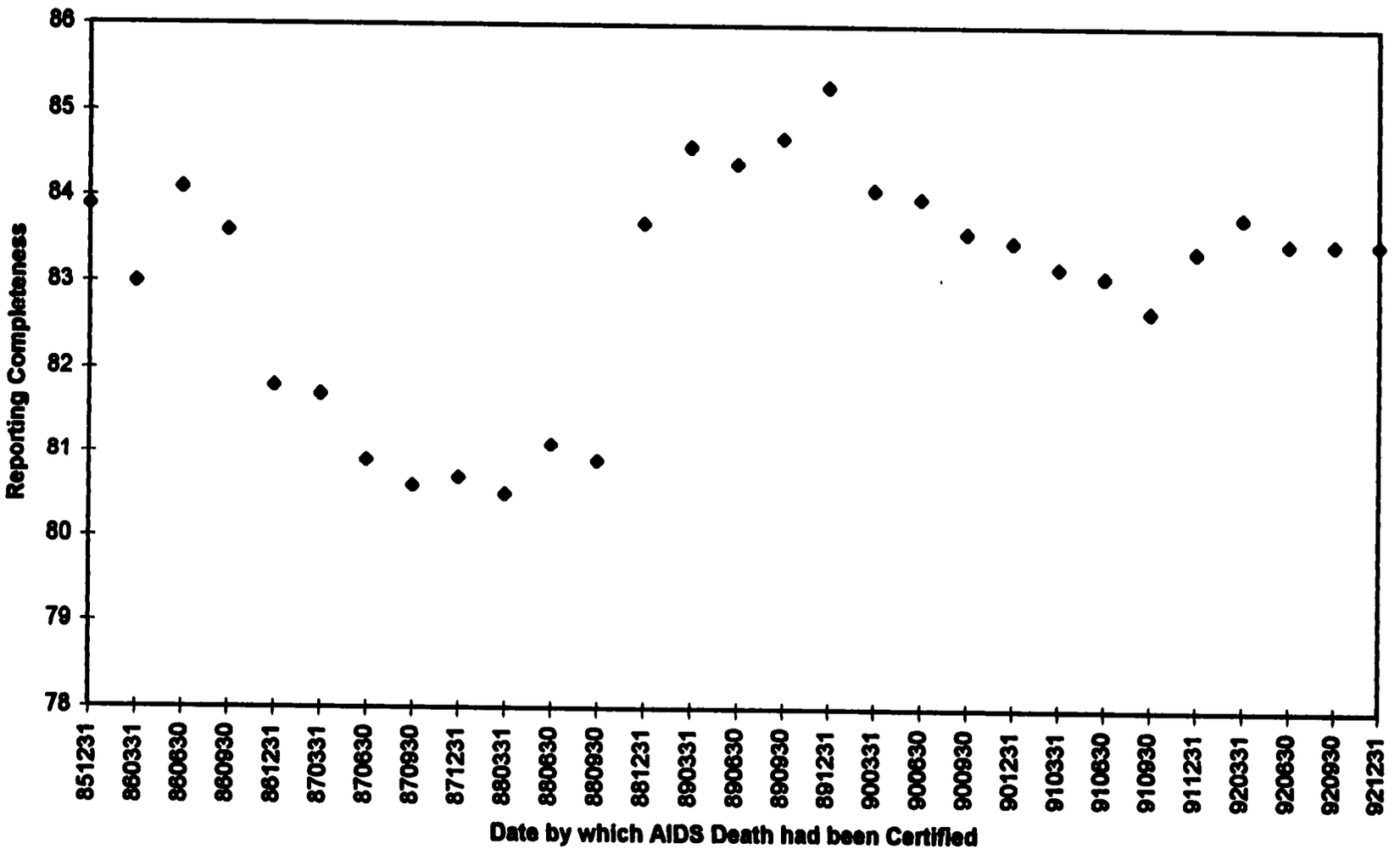


Chart 12: Reporting Completeness at Year 2 - "Possibles" as DCO

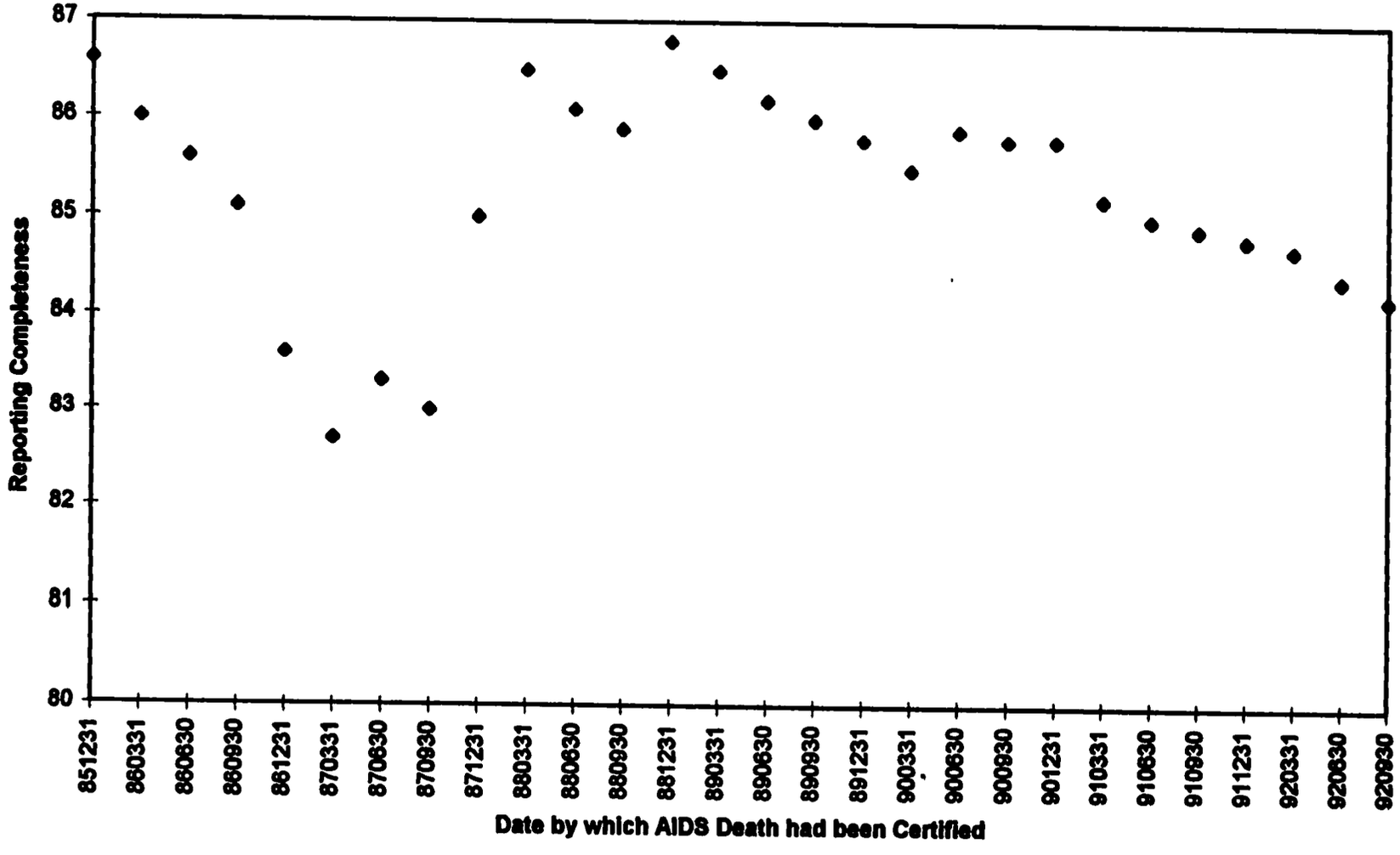


Chart 13: Reporting Completeness at Year 3 - "Possibles" as DCO

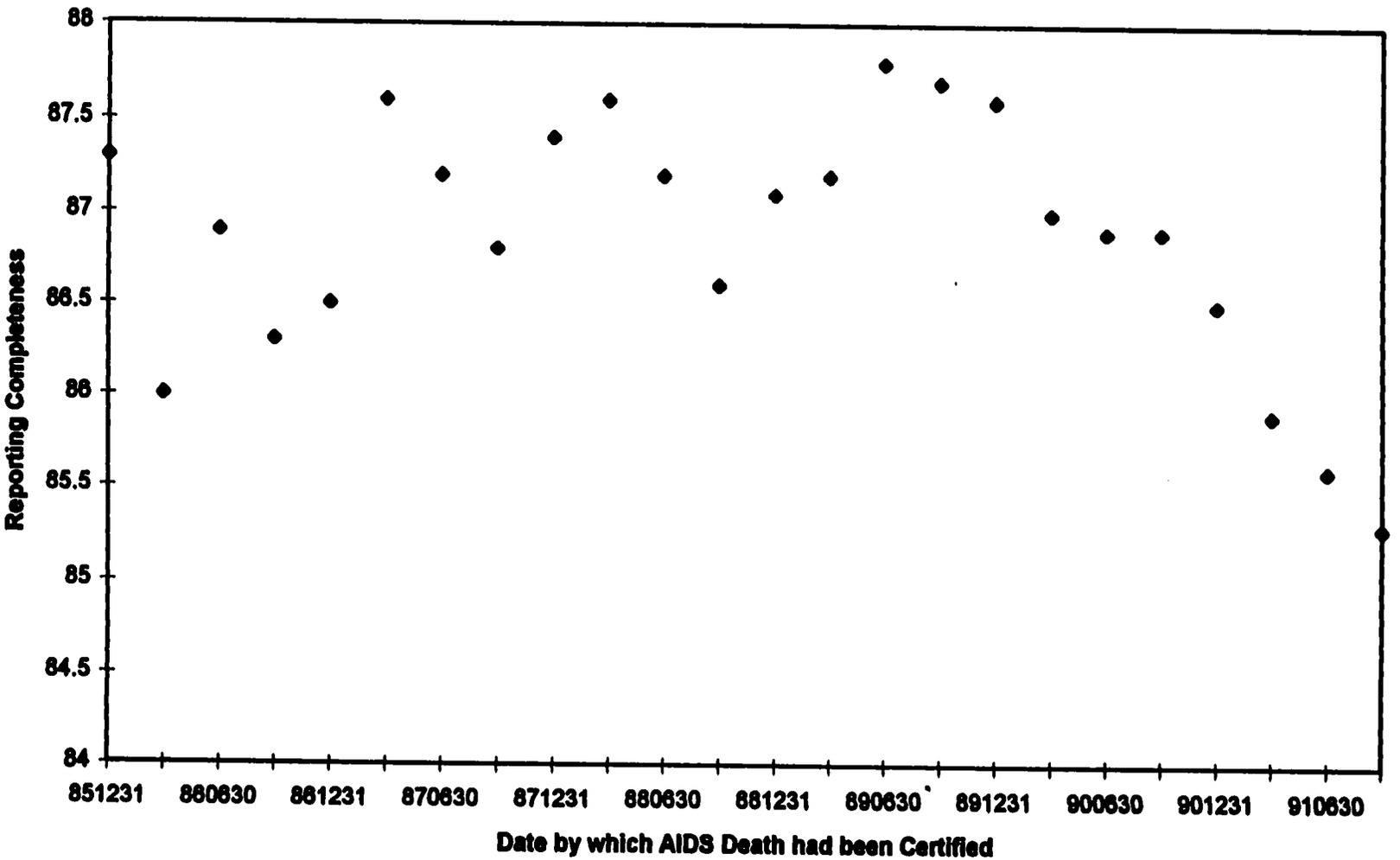


Chart 14: Reporting Completeness at Year 4 - "Possibles" as DCO

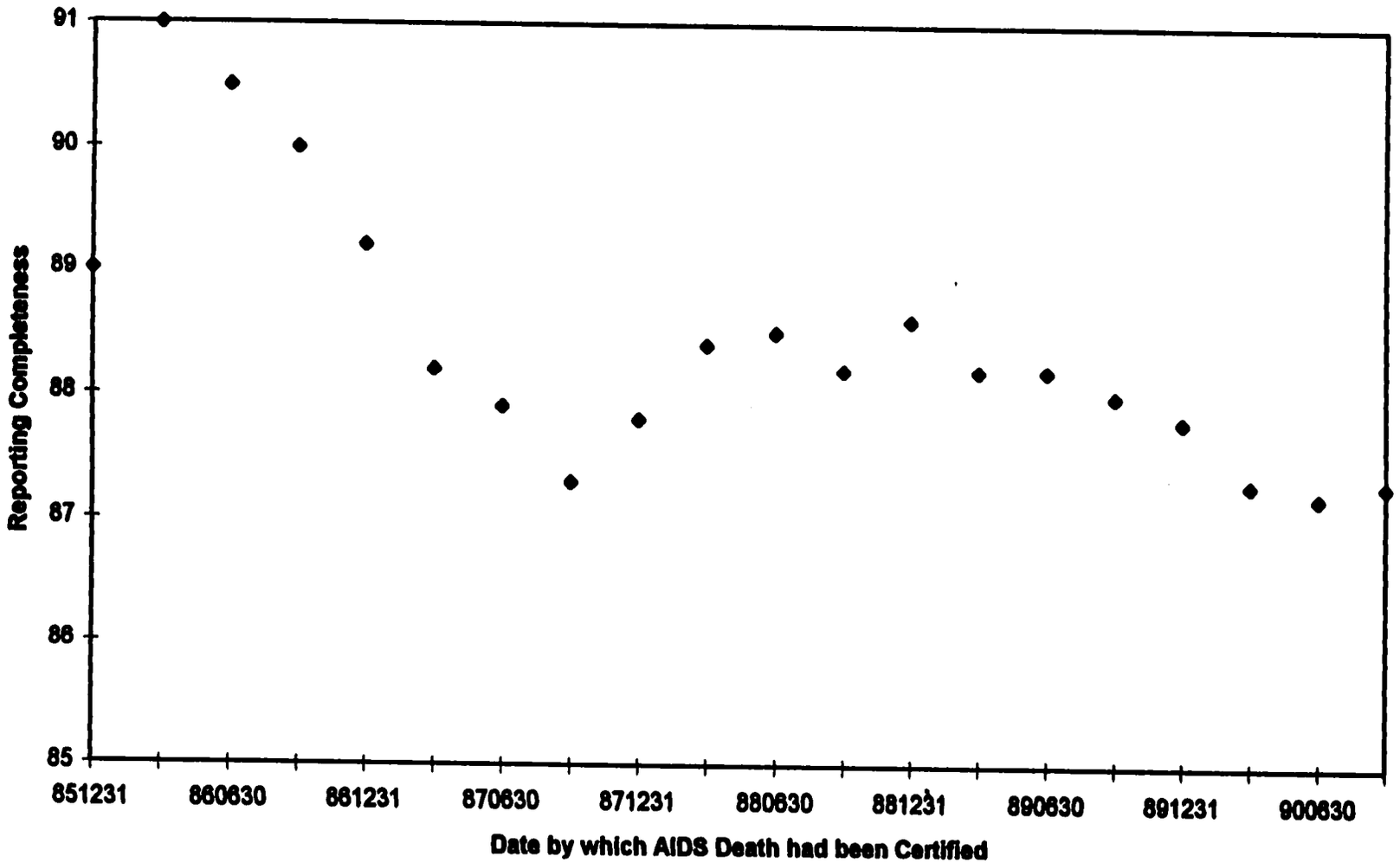


Chart 15: Reporting Completeness at Year 5 - "Possibles" as DCO

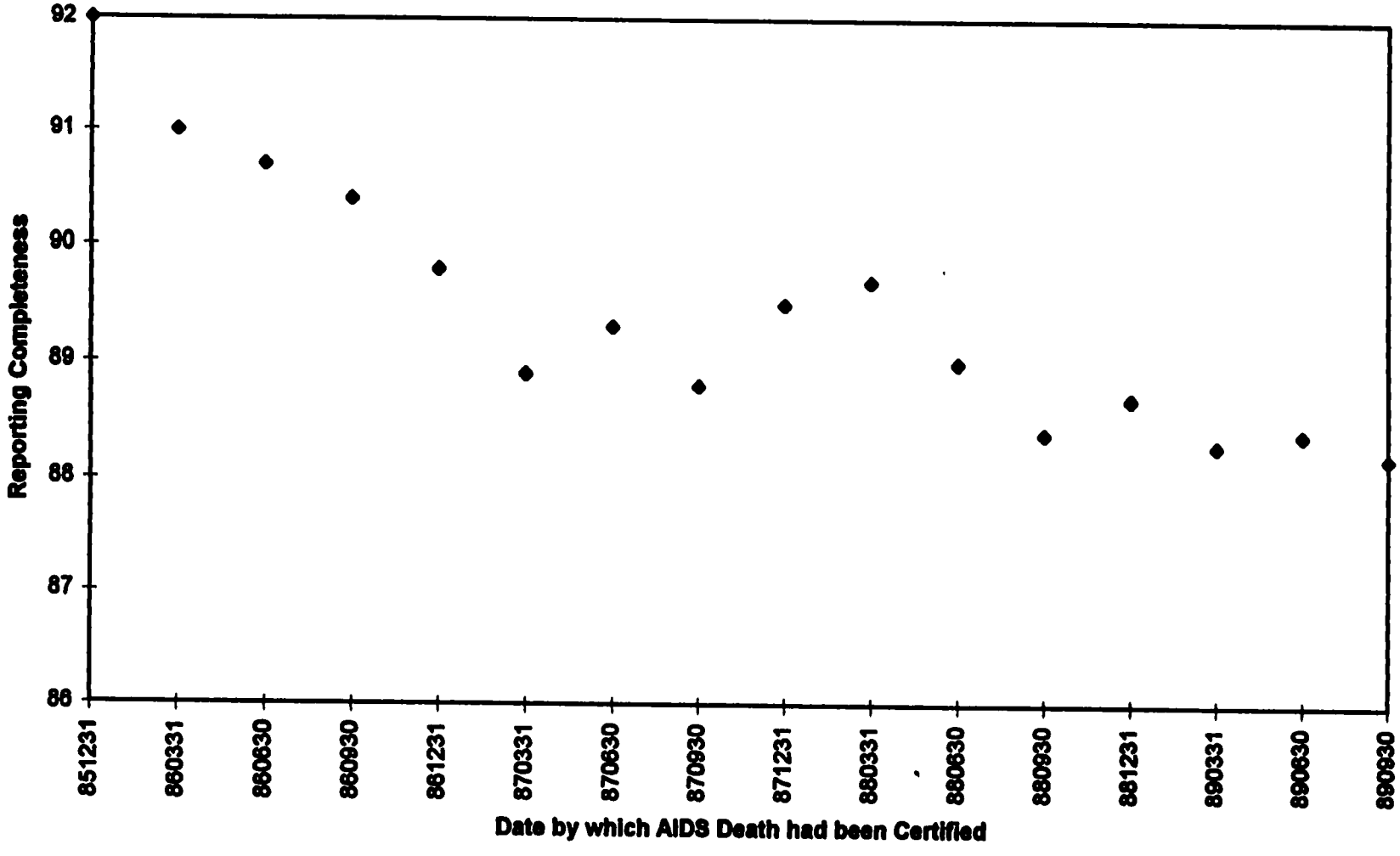


Chart 16: Reporting Completeness at Year 6 - "Possibles" as DCO

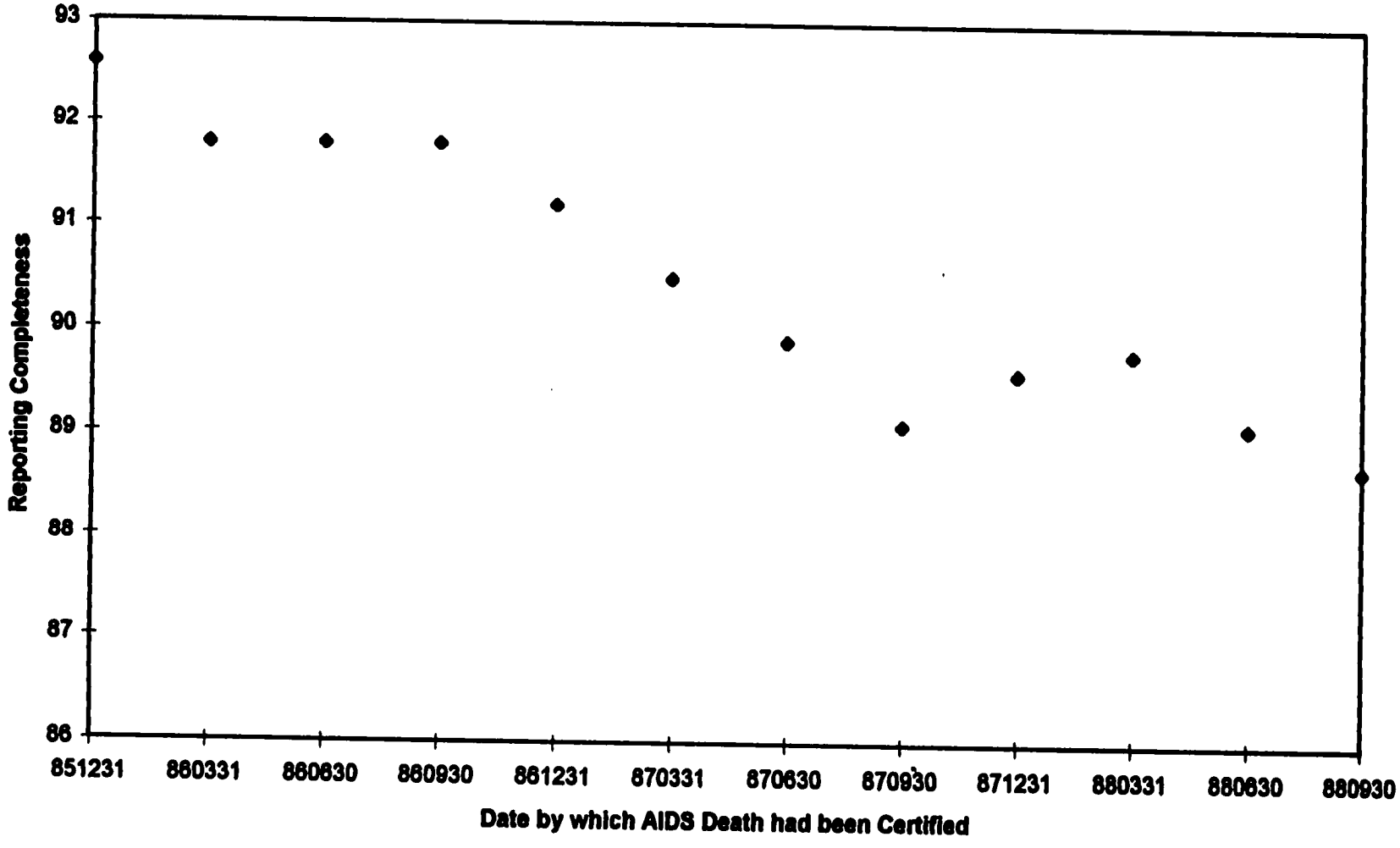


Chart 17: Reporting Completeness at Year 7 - "Possibles" as DCO

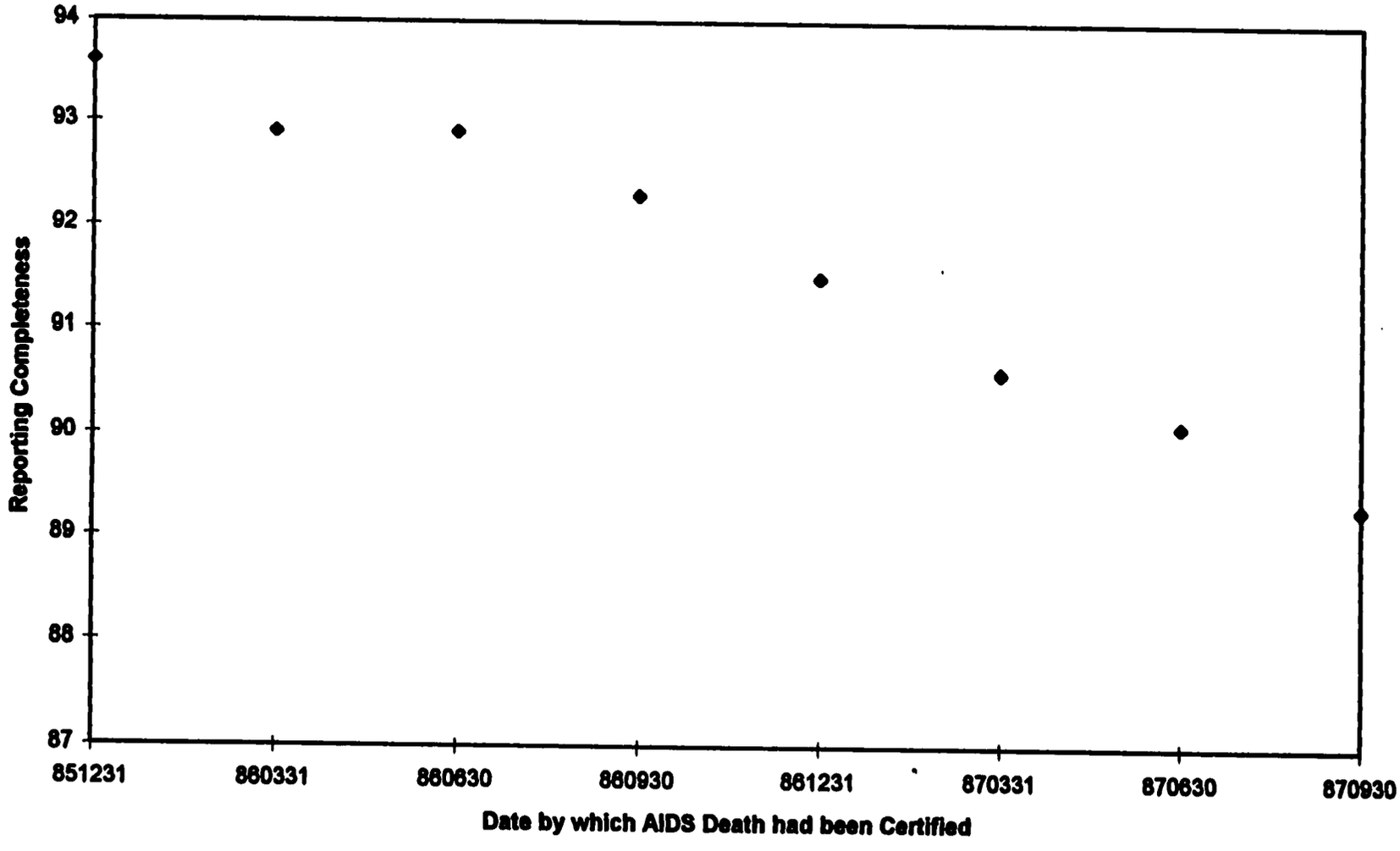
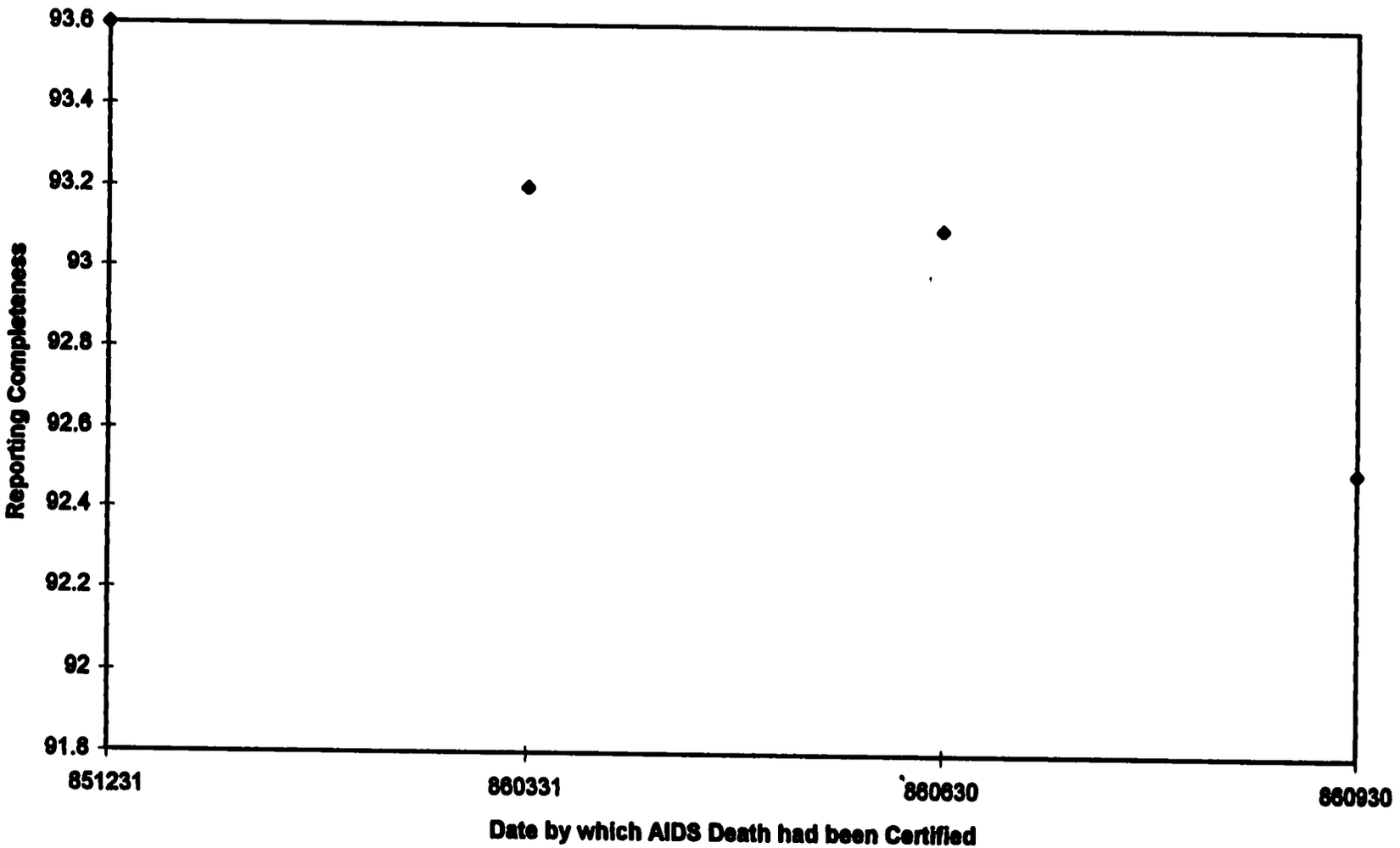


Chart 18: Reporting Completeness at Year 8 - "Possibles" as DCO



APPENDIX D

Sekar-Deming Test for Independence

The correlation coefficient r is calculated as follows:

$$r=S_{12}/S_1S_2=(\sum N_i[p_1^{(i)}-p_1][p_2^{(i)}-p_2])/S_1S_2\sum N_i$$

N_i=number of cases in that stratum
$p_1^{(i)}$=probability of CMDB detecting a case in that stratum
p_1=probability that CMDB detects a case
$=\sum N_i p_1^{(i)}/\sum N_i$
$p_2^{(i)}$=probability of ACRSS detecting a case in that stratum

p_2=probability that ACRSS detects a case
$=\sum N_i p_2^{(i)}/\sum N_i$
S_1=square root of $(\sum N_i[p_1^{(i)}-p_1]^2)/\sum N_i$
S_2=square root of $(\sum N_i[p_2^{(i)}-p_2]^2)/\sum N_i$

Both data sources and the linked file were stratified by year of death. Calculation of N_i is shown below:

Table 1 - Reporting Completeness of CMDB By Year of Death

Year of Death	Linked (n_{12})	ACRSS Only	n_2 (AO + n_{12})	DCO	n_1 (DCO + n_{12})	$N=n_1n_2/n_{12}$
82	13	1	14	1	14	15.1
83	26	2	28	3	29	31.2
84	73	5	78	4	77	82.3
85	169	6	175	10	179	185.4
86	324	20	344	36	360	382.2
87	508	21	529	60	568	591.5
88	613	12	625	78	691	704.5
89	811	33	844	109	920	957.4
90	906	40	946	155	1061	1107.8
91	1064	51	1115	215	1279	1340.3
92	1245	38	1283	230	1475	1520.0
Total	5752	229	5981	901	6653	6917.9

Calculation of the numerator and denominator are shown on the next page.

Table 2 - Calculation of Numerator for Independence Test

Yr of Dth	p_1^i(CMDB)	p_2^i(ACRSS)	$p_1^i-p_1$	$p_2^i-p_2$	N_i	$N_i(p_1^i-p_1)(p_2^i-p_2)$
82	0.9286	0.9286	-0.0331	0.0640	15.08	-0.0320
83	0.9286	0.8966	-0.0331	0.0640	31.23	-0.0662
84	0.9359	0.9481	-0.0258	0.0713	82.27	-0.1515
85	0.9657	0.9441	0.0040	0.1011	185.36	0.0750
86	0.9419	0.9000	-0.0199	0.0773	382.22	-0.5864
87	0.9603	0.8944	-0.0014	0.0957	591.48	-0.0798
88	0.9808	0.8871	0.0191	0.1162	704.53	1.5632
89	0.9609	0.8815	-0.0008	0.0963	957.44	-0.0747
90	0.9577	0.8539	-0.0040	0.0932	1107.84	-0.4118
91	0.9543	0.8319	-0.0075	0.0897	1340.31	-0.8956
92	0.9704	0.8441	0.0087	0.1058	1520.02	1.3944
Sum					6917.77	0.7346

p_1=reporting completeness for CMDB=.9617 p_2=reporting completeness for ACRSS=.8645

Table 3 - Calculation of Denominator for Independence Test

Yr of Dth	p_1^i (CMDB)	p_2^i (ACRSS)	$(p_1^i-p_1)^2$ =x	N_i	$(p_2^i-p_2)^2$ =y	N_i*x	N_i*y
82	0.9286	0.9286	0.0011	15.08	0.0041	0.0166	0.0618
83	0.9286	0.8966	0.0011	31.23	0.0010	0.0343	0.0319
84	0.9359	0.9481	0.0007	82.27	0.0070	0.0549	0.5733
85	0.9657	0.9441	0.0000	185.36	0.0063	0.0030	1.1732
86	0.9419	0.9000	0.0004	382.22	0.0013	0.1506	0.4797
87	0.9603	0.8944	0.0000	591.48	0.0009	0.0012	0.5252
88	0.9808	0.8871	0.0004	704.53	0.0005	0.2567	0.3582
89	0.9609	0.8815	0.0000	957.44	0.0003	0.0006	0.2750
90	0.9577	0.8539	0.0000	1107.84	0.0001	0.0177	0.1259
91	0.9543	0.8319	0.0001	1340.31	0.0011	0.0744	1.4307
92	0.9704	0.8441	0.0001	1520.02	0.0004	0.1142	0.6389
Sum				6917.77		0.7241	5.6739

p_1=reporting completeness for CMDB=.9617 p_2=reporting completeness for ACRSS=.8645

S_1=sq root(.7241/6917.8)=.01023
S_2=sq root(5.674/6917.8)=.02864

$r=.7346/S_1S_2\sum N_i=.7346/(.01023*.02864*6917.8)=.36$