The use of models in the estimation of disease epidemiology

Michelle E. Kruijshaar,[1, 2] Jan J. Barendregt,[1] & Nancy Hoeymans[2]

Objective To explore the usefulness of incidence–prevalence–mortality (IPM) models in improving estimates of disease epidemiology.
Methods Two artificial and four empirical data sets (for breast, prostate, colorectal, and stomach cancer) were employed in IPM models.
Findings The internally consistent artificial data sets could be reproduced virtually identically by the models. Our estimates often differed considerably from the empirical data sets, especially for breast and prostate cancer and for older ages. Only for stomach cancer did the estimates approximate to the data, except at older ages.
Conclusion There is evidence that the discrepancies between model estimates and observations are caused both by data inaccuracies and past trends in incidence or mortality. Because IPM models cannot distinguish these effects, their use in improving disease estimates becomes complicated. Expert opinion is indispensable in assessing whether the use of these models improves data quality or, inappropriately, removes the effect of trends.

Keywords Epidemiologic research design; Models, Theoretical; Incidence; Prevalence; Mortality; Reproducibility of results; Neoplasms/epidemiology; Breast neoplasms/epidemiology; Prostatic neoplasms/epidemiology; Colorectal neoplasms/epidemiology; Stomach neoplasms/epidemiology; Netherlands (*source: MeSH, NLM*).

Mots clés Projet d'étude épidémiologique; Modèle théorique; Incidence; Prévalence; Mortalité; Reproductibilité des résultats; Tumeurs/épidémiologie; Tumeur sein/épidémiologie; Tumeur prostate/épidémiologie; Tumeur colorectale/épidémiologie; Tumeur estomac/épidémiologie; Pays-Bas (*source: MeSH, INSERM*).

Palabras clave Diseño de investigaciones epidemiológicas; Modelos teóricos; Incidencia; Prevalencia; Mortalidad; Reproducibilidad de resultados; Neoplasmas/epidemiología; Neoplasmas de la mama/epidemiología; Neoplasmas de la próstata/epidemiología; Neoplasmas colorrectales/epidemiología; Neoplasmas gástricos/epidemiología; Países Bajos (*fuente: DeCS, BIREME*).

Bulletin of the World Health Organization 2002;80:622-628.

Voir page 626 le résumé en français. En la página 626 figura un resumen en español.

Introduction

Quantitative descriptions of disease epidemiology, such as incidence, prevalence and mortality, by age and sex, are essential inputs for burden of disease studies and cost-effectiveness analyses of interventions. Such studies serve as an important source of information for policy-making, planning, and research prioritization in health care. Empirical observation is obviously the gold standard for obtaining epidemiological information, but empirical data are often incomplete or of dubious validity. In addition, the validity of estimates tends to vary even for an individual disease. For example, in instances where incidence is more difficult to observe than mortality, more incident cases than deaths are likely to be missed. In this case, therefore, data on incidence are less complete than those on mortality, making these two parameters internally inconsistent.

One way to circumvent these data limitations is to exploit the causal structure of the disease process: incidence has to precede prevalence, and cause-specific mortality can only follow being diseased. Incorporating the causal structure into a mathematical model makes it possible to calculate data that are missing from the observational set and to check for the internal consistency of observations. An example of the first of these procedures is the back-calculation of (unobserved) human immunodeficiency virus (HIV) infection from data on the incidence of acquired immunodeficiency syndrome (AIDS) (*1*).

The Global Burden of Disease 1990 study, and many subsequent national burden of disease studies, made extensive use of DisMod, a generic mathematical disease model, which was specially designed to supplement observational data and check for internal consistency (*2*). Previously we have employed a conceptually similar disease model to calculate unobserved incidence data (*3*). The present article explores the usefulness of such generic disease models that describe the relation between incidence, prevalence and mortality (IPM models) for improving estimates of disease epidemiology. We consider whether these models calculate the correct results and, if so, how useful these results are. Our approach is to apply two IPM models (DisMod and our own model) to two artificial data sets known to be complete and consistent and to four high-quality empirical data sets for cancers, drawn from Dutch registries. The ability of the IPM models to describe adequately the data sets serves as an indicator of their usefulness.

Methods

Artificial data sets

In order to demonstrate that the models can calculate the correct results, we first used them with internally consistent

[1] Department of Public Health, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands (email: kruijshaar@mgz.fgg.eur.nl). Correspondence should be sent to this author.

[2] Department for Public Health Forecasting, National Institute of Public Health and the Environment, Bilthoven, Netherlands.

Ref. No. **00-0904**

data (formal validity). Data sets for breast and colorectal cancer were generated by MISCAN, a microsimulation model for the evaluation of screening programmes (*4*, *5*). MISCAN creates a cohort of hypothetical individuals, each of whom has a risk of developing cancer, and, once the disease is present, a survival drawn from a lognormal distribution. Incidence, prevalence and mortality data generated by this model are, by definition, complete and internally consistent.

Empirical data

We applied the IPM models to national incidence and mortality data for breast cancer (ICD-9 code 174), prostate cancer (ICD-9, code 185), colorectal cancer (ICD-9 codes 153 and 154) and stomach cancer (ICD-9 code 151). Data averaged for 1991–95, specified by sex and 5-year age group (up to ⩾95 years), were used. Statistics Netherlands (CBS) collects mortality data by cause-of-death on a continuous basis using information from death certificates. Incidence data are collected continuously by the Dutch Cancer Registry (NKR), which receives its information from nine regional cancer centres. These data are based on pathology reports, complemented by national hospital admission data; death certificates are not used as an additional source.

The cancer registries do not estimate prevalence data on a regular basis. The Regional Cancer Centre South (IKZ) determined the prevalence of the specific cancers for which incidence has been determined for the eastern part of the coverage area on 1 January 1993: for all incident cases registered in the region from 1970 until 1992 the population registry was checked to determine whether the persons concerned were still alive. For the same region we obtained the regional mortality and incidence rates from IKZ, averaged for 1991–95. Mortality data for this region originated from the CBS database (region: COROP 36 and 37). Regional data were specified by sex and 5-year age group (up to ⩾85 years).

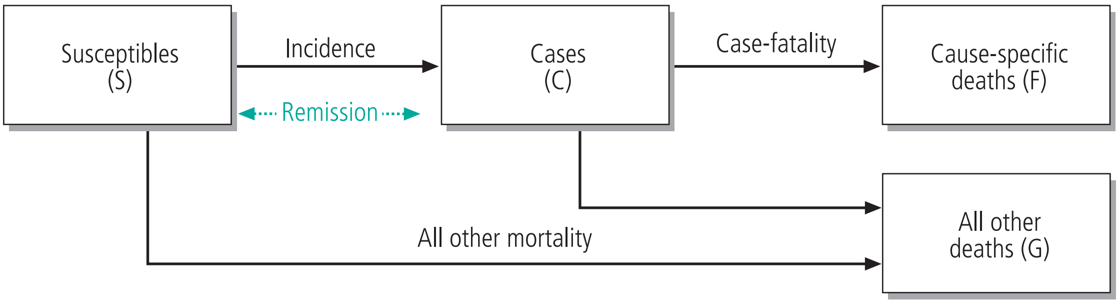
IPM models

Both DisMod and our model are based on the conceptual disease model depicted in Fig. 1. The population is described as being in different states, while transition hazards determine how people move from one state to another. Within a population, individuals can be either susceptibles or cases. Cases may die from their disease, while both cases and susceptibles are at risk of dying from other causes. There are consequently three transition hazards: incidence, case-fatality, and all other mortality. DisMod also includes remission as a fourth transition hazard, but for our analysis we set this hazard to zero since cure is not taken into account in the registered cancer prevalence. The framework in Fig. 1 shows that the number of cases can be calculated by following an initially disease-free cohort over time and applying the transition hazards. Under the important assumption made in the IPM models that there are no trends in the transition hazards, time is equal to a patient's age. The models thus permit calculation of prevalence at a certain age from the prevalence at the previous age and the mortality and incidence in the age interval.

Although they are based on this common conceptual model, the actual model calculations differ. DisMod uses a set of linear differential equations that describe the transitions between the states. The solution of the equations is approximated by using the finite differences method. Incidence and case-fatality hazards are required as input parameters, and we approximated them using rates. Case-fatality rates were calculated from mortality data and the prevalence calculated from our own model (see below and Annex). Since DisMod cannot calculate data for age groups over 90 years and can only handle a limited number of age groups, we specified 5-year age groups from 15 years to 89 years. The calculation is performed using a competing risk life table (*6*). General mortality data for the Netherlands for 1991–95 reported by CBS were specified.

Our model gives an exact solution based on an analytical solution of a continuous time Markov process (*7*). We refer to it here as the analytical model. Using a spreadsheet we implemented the formula for the calculation of prevalence from incidence and mortality probabilities (see Annex). Mortality and incidence rates per 5-year age group up to ⩾95 years served as input. We first interpolated these data to 1-year age groups up to ⩾95 years and then converted them to probabilities (see Annex for formulas and methods). Apart from the different calculation method (approximate versus exact), the analytical model thus also differs from DisMod in the way it treats mortality. Mortality probability relates to the total population, whereas case-fatality, used in DisMod, concerns only prevalent cases. In the event of inconsistent data, the mortality probability may exceed the predicted prevalence, resulting in negative prevalence estimates in the analytical model, which is not possible if case-fatality is used.

Fig. 1. **Schematic representation of a Markov model for cancers**



WHO 02.93

The models were assessed by comparing the calculated prevalence with the observed data. We extrapolated the observed prevalence data from ⩾85 years to ⩾95 years by applying the cubic-spline methodology and using life-table derived mean ages of 89.6 years and 90.8 years, for men and women aged ⩾85 years, respectively.

Results

Application of the internally consistent MISCAN data to the models resulted in prevalence estimates that were virtually identical to those generated by MISCAN and to one another. When the observed data were applied the results of the two models were also practically identical. The reproducibility of MISCAN data and the consistency of the results from the two models suggest that they calculated the correct results.

The prevalences calculated from national data by the analytical model are shown with the observed prevalences in Figs. 2–5. The model estimates increase with age, at first exponentially, but subsequently at a slower pace. At ages >80 years the estimates reach a maximum and then decline. The decline at older ages is most apparent for stomach cancer, the calculated prevalence decreasing to zero or even to negative values; the smallest decline is that for breast cancer.

The predicted prevalences are nearly always larger than the observed ones. However, stomach cancer is exceptional in this respect in that the estimate approximates to the observed value, except for ages >85 years. For prostate and breast cancer the discrepancy is large (depending on age, the model calculations are up to about two and three times larger, respectively), while for colorectal cancer it is intermediate (up to about 1.5 times larger).

Discussion

The results from the artificial data sets support the validity of the IPM models: despite the difference of a lognormally distributed survival in MISCAN and a piecewise exponentially distributed survival in the IPM models, the latter are able to reproduce the MISCAN prevalence very well. In addition the two IPM models produce virtually the same results, a further indication of validity. Nevertheless, when registered incidence and mortality data were used, the predicted prevalence differed considerably from that observed, and for stomach cancer impossible results were produced. Three possible reasons for these discrepancies are considered below.

Regional differences

The regionally observed prevalence data, to which the national estimates are compared, may not be representative of the national situation. Breast cancer screening was introduced in the IKZ region between 1993 and 1997, whereas in the rest of the Netherlands it was introduced around 1990. Cause-specific prostate cancer mortality is unequally distributed (*8*). We explored the influence of regional differences by applying regional incidence and mortality data to the model. A comparison of the results of these calculations (not shown) with the empirical data revealed that the differences between estimates and observations were similar, and, if anything, somewhat larger than the discrepancies we found on using national data in the models. The regional variation in disease epidemiology therefore could not explain the differences.

Fig. 2. **Observed and estimated prevalence of female breast cancer, by age.** Empirical data (broken line) are from the IKZ region and were interpolated from 5-year to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid line) are calculated from national incidence and mortality data using the analytical model

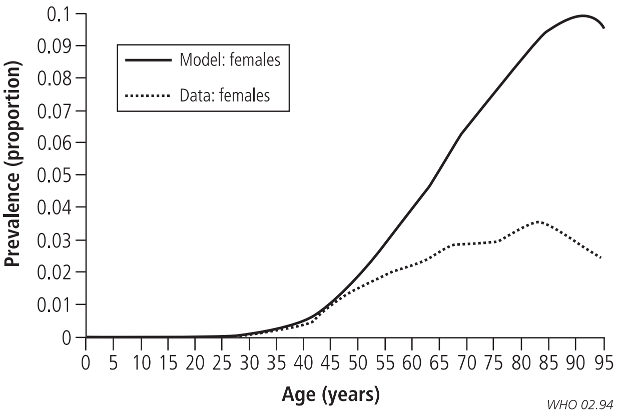
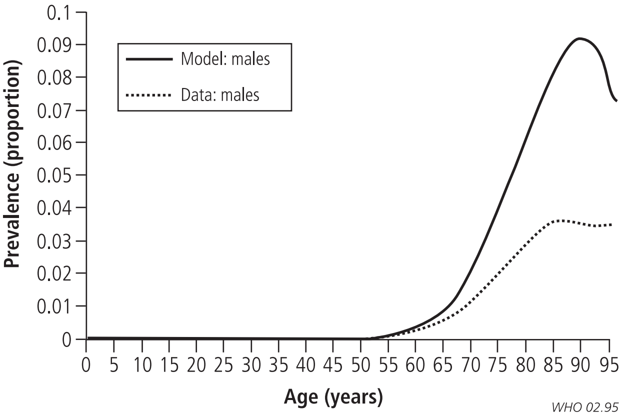


Fig. 3. **Observed and estimated prevalence of prostate cancer, by age.** Empirical data (broken line) are from the IKZ region and were interpolated from 5-year to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid line) are calculated from national incidence and mortality data using the analytical model



Past trends

Because both models are based on the assumption that incidence and mortality are in a steady state, the occurrence of trends in incidence or mortality would lead to discrepancies between the model estimates and the observations. Prevalence is a stock variable, comprising all past incident cases that are still alive. It is therefore dependent on incidence and case-fatality from the past as well as the present.

Cancer incidence has a tendency to rise for the tumours we studied, because of increased awareness and screening, and for other, unknown, reasons (*9*, *10*). The incidence of breast cancer, for example, is presumed to show a secular trend of 1% per year (*11*), on top of which an additional increase is imposed because of the introduction of breast cancer screening around 1990 in the Netherlands. A notable exception to this rising incidence is presented by stomach cancer, for which there has been a long-term secular decline (*9*, *10*).

Fig. 4. **Observed and estimated prevalence of colorectal cancer, by age and sex.** Empirical data (broken lines) are from the IKZ region and were interpolated from 5-year to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid lines) are computed on the basis of national incidence and mortality data using the analytical model

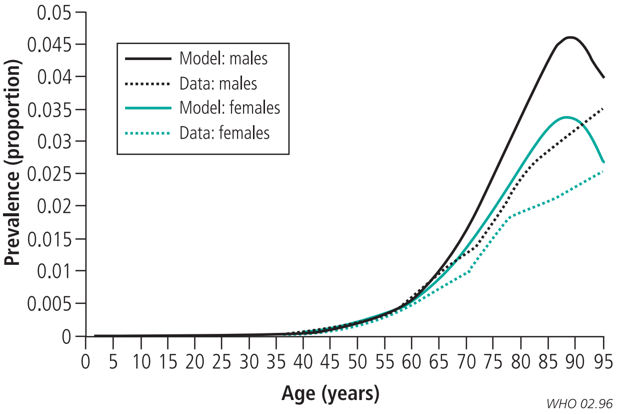
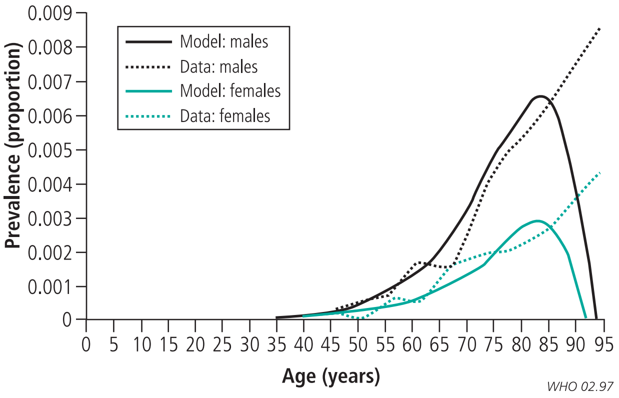


Fig. 5. **Observed and estimated prevalence of stomach cancer, by age and sex.** Empirical data (broken lines) are from the IKZ region and were interpolated from 5-year to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid lines) are computed on the basis of national incidence and mortality data using the analytical model



Cancer mortality, meanwhile, remains relatively stable (*9*, *10*, *12–14*), except for stomach cancer and female colorectal cancer, for which it has declined. With increasing incidence and constant mortality, prevalence increases over time but does so less rapidly than incidence, since it also includes persons who became incident in the past. Consequently, applying current incidence and mortality to the model produces estimates that are higher than the observations. The largest deviations were seen concordantly for breast cancer and prostate cancer, the two cancers for which the rise in incidence has been most apparent, mainly because of screening.

In an additional analysis old incidence data for breast cancer were used in the model in order to check whether trends could explain the discrepancies. The risk of mortality for breast cancer patients remains elevated for more than 20 years after diagnosis. We therefore used regional incidence data for 1968–72 (*12*), obtaining an estimated prevalence close to the observed value for 1993. This shows that the trend in incidence may indeed cause a difference. Nevertheless it cannot explain the discrepancy entirely: the average incidence that cases prevalent on 1 January 1993 were exposed to lies somewhere between the 1991–95 and the 1968–72 incidences. Although we believe the effect of the trends in incidence is considerable, other factors evidently also play a role.

Data inaccuracies

Inaccuracies in the epidemiological estimates are the third possible reason for the differences between observations and model predictions. Statistics on mortality by cause-of-death in the Netherlands are assumed to be reliable, although no studies are known in which the completeness of the death registry has been investigated in absolute terms. Compared with other European countries, in the Netherlands the detection fraction for cancer as a cause of death is high (*15*). Furthermore, it has been argued that deaths from cancer in general are not likely to be missed (*16*), although misclassification between cancers may occur for older age groups. Thus it is unlikely that underregistration of cancer deaths is a causative factor in our generally higher prevalence estimates, especially with regard to young and middle-aged people. Nevertheless, the underestimation of mortality remains a possible explanation for discrepancies. Since we did not include excess mortality from other diseases in our model we implicitly assumed it to be zero. However, cancer patients also suffer from an increased risk of dying from diseases other than cancer (*17*). We believe that, in addition to the effect of trends, the impact of ignoring this factor makes an important contribution to the discrepancies.

At older ages, where multiple medical problems are frequent and pathological examinations are performed relatively infrequently, misclassification of cancer deaths may lead to the overregistration of deaths for the more frequent types of cancer. This would cause prevalence estimates to be too low and could contribute to the decline of our prevalence estimates at older ages.

Cancer incidence data in the Netherlands are reliable. Nevertheless, because they are based on pathology and hospitalization data, those incident cancer cases that did not undergo a pathological examination and were not hospitalized would be systematically excluded from registration. It has been estimated that this would lead to an underregistration of 1.3–1.6% (*18*, *19*). Moreover, some cases that are included in pathology or hospitalization registries are missed. This non-systematic exclusion has been estimated to occur in 2.2% of cases (*19*). A completeness of approximately 96.2% is thus achieved, which is comparable to the level of completeness in several other national cancer registries (*19*). This incompleteness seems to be concentrated in the highest age groups; one study suggested that missed incident cases mostly related to elderly persons with cancer of the digestive tract (*18*), although this was not confirmed in another study (*19*). The underregistration of incidence may help to explain the impossible negative prevalence calculated for stomach cancer and the decline in the estimates for colorectal cancer at older ages (Fig. 4 and Fig. 5).

Underregistration cannot, however, explain the finding that the prevalence estimates are generally higher than the observations for the other age groups. Multiple malignancies can contribute to this. The incidence registry counts the number of malignancies, whereas the prevalence data are based on the number of persons with a malignancy. Consequently, a person with multiple malignancies in the same organ is counted more than once in the incidence data but only once in the prevalence data. For breast and colorectal cancer such multiple malignancies may be present, and can account for up to 10% and 15%, respectively, of the incident cases (J.-W. Coebergh, personal communication, 2000). This would make our prevalence estimate too high and would explain part of the differences, but not more than the 10% or 15% by which the incidence is overestimated.

The incompleteness of prevalence data could also be a factor contributing to the higher estimates. Although based on regional incidence data, prevalence data may be somewhat less complete because cancer registration was less complete in its early years than more recently (*19*). Since only old cases are underestimated in this way, prevalence is only affected if the survival time is long. Furthermore, this underestimation might be diminished by the opposite phenomenon: overestimation of prevalence resulting from incomplete ascertainment of survival status. The latter incompleteness would be very small, however, since deaths are unlikely to be missed, although problems may arise when persons have moved out of the country. We believe that the underestimation of prevalence data is not large and that it is unlikely to explain a large part of the differences.

Conclusion

The test with the artificial data supports the formal validity of IPM models. However, the confrontation with the four empirical data sets of presumed high quality shows that, in practice, there may be large discrepancies between measurements and calculations. The discrepancies are likely to be attributable in considerable measure to past trends in incidence but also to data inaccuracies, the most important source of which seems to be underestimation of mortality as a result of ignoring excess mortality from other causes.

The model cannot distinguish between the effects of trends and the effects of data inaccuracies. Separating these effects would require a dynamic model that describes the disease processes over time, and could incorporate the effects of past trends. Unfortunately, such a model would be much more complex. Moreover, since the trends would have to be quantified, more input data would be required, and these have proved difficult to obtain. Consequently, a dynamic analysis is often not feasible.

In practice use of IPM models such as DisMod occurs particularly when data are incomplete and/or of low quality. In such circumstances it is impossible to distinguish between the apparent inconsistencies that represent real data problems and those that are attributable to past trends. This complicates the use of such models in improving estimates of disease epidemiology. Considerable judgement has to be exercised when the disadvantage of forcing data to comply with the assumption of a steady state is weighed against the goal of reducing the unreliability of the data. Expert knowledge on disease epidemiology and registries remains indispensable for guiding this process. ■

Acknowledgements

We thank Dr Jan-Willem Coebergh, Dr Marie-Louise Essink-Bot, Dr Pieter G.N. Kramers, Professor Paul J. van der Maas and Professor Daan Kromhout for their constructive suggestions and comments on this manuscript. The present work was sponsored by the Netherlands Institute of Health Sciences.

Conflicts of interest: none declared.

Résumé

L'utilisation de modèles dans les estimations en matière d'épidémiologie

Objectif Etudier l'utilité des modèles d'incidence, de prévalence et de mortalité pour améliorer les estimations en matière d'épidémiologie.

Méthodes On a appliqué les modèles d'incidence, de prévalence et de mortalité à deux séries de données artificielles et quatre séries de données empiriques (pour les cancers du sein, de la prostate, de l'estomac et du cancer colo-rectal).

Résultats Les séries de données artificielles ayant une cohérence interne ont pu être reproduites virtuellement de manière identique par les modèles. Souvent, nos estimations différaient sensiblement des données empiriques, particulièrement pour les cancers du sein et de la prostate et pour les personnes les plus âgées. Pour le cancer de l'estomac seulement, les estimations se rapprochaient des données, sauf lorsqu'il s'agissait des personnes les plus âgées.

Conclusion Il semble que les écarts entre les estimations fournies par les modèles et les observations soient causés à la fois par l'inexactitude des données et par les tendances passées en matière d'incidence ou de mortalité. Les modèles d'incidence, de prévalence et de mortalité ne permettant pas de distinguer entre les effets tenant à l'inexactitude des données et ceux dus aux tendances passées, il devient difficile de les utiliser pour améliorer les estimations. L'avis de spécialistes est donc indispensable pour évaluer si l'utilisation de ces modèles améliore la qualité des données ou si, malencontreusement, elle fait disparaître l'influence des tendances.

Resumen

Uso de modelos para estimar la epidemiología de enfermedades

Objetivo Determinar la utilidad de los modelos basados en la incidencia, la prevalencia y la mortalidad para mejorar las estimaciones epidemiológicas.

Métodos Los modelos de incidencia, prevalencia y mortalidad (IPM) se aplicaron a dos conjuntos de datos artificiales y cuatro empíricos (para el cáncer de mama, próstata, colon y recto, y estómago).

Resultados Los modelos consiguieron reproducir de forma casi idéntica los conjuntos de datos artificiales, internamente coherentes. En cambio, nuestras estimaciones difirieron a menudo considerablemente de los conjuntos de datos empíricos, sobre todo en el caso de los cánceres de mama y de próstata y en lo que respecta a las personas de más edad. Sólo en el caso del cáncer de estómago las estimaciones se aproximaron a los datos, exceptuando de nuevo las personas mayores.

Conclusión Hay indicios de que las discrepancias entre las estimaciones arrojadas por los modelos y las observaciones se deben tanto a inexactitudes de los datos como a las tendencias seguidas por la incidencia y la mortalidad en el pasado. Dado que los modelos IPM no permiten distinguir esos efectos, resulta complicado usarlos para mejorar las estimaciones. La opinión de los expertos se revela por tanto como indispensable para evaluar si el uso de esos modelos mejora la calidad de los datos o elimina de manera indebida el efecto de las tendencias.

References

1. Rosenberg PS, Biggar RJ, Goedert JJ, Gail MH. Back calculation of the number with human immunodeficiency virus infection in the United States. *American Journal of Epidemiology* 1991;133:276-85.
2. Murray CJ, Lopez AD. Quantifying disability: data, methods and results. *Bulletin of the World Health Organization* 1994;72:481-94.
3. Barendregt JJ, Baan CA, Bonneux L. An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus. *Epidemiology* 2000;11:274-9.
4. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. *Computer Methods and Programs in Biomedicine*, 1985;20:79-93.
5. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research*, 1999;32:13-33.
6. Manton KG, Stallard E. *Chronic disease modelling*. New York: Oxford University Press; 1988.
7. Barendregt JJ, van Oortmarssen GJ, van Hout BA, van den Bosch JM, Bonneux L. Coping with multiple morbidity in a life table. *Mathematical Population Studies* 1998;7:29-49.
8. *Atlas of cancer mortality in the Netherlands, 1979–1990*. The Hague: SDU/publishers, CBS publications; 1992.
9. Visser O, Coebergh JWH, Schouten LJ, van Dijck JAAM. *Incidence of cancer in the Netherlands 1995.* Utrecht: Vereniging van Integrale Kankercentra; 1998.
10. Maas IAM, Gijsen R, Lobbezo IE, Poos MJJC. [Public health status and forecast 1997. Part 1. The health status: an update].Bilthoven: Rijksinstituut voor Volksgezondheid en Milieu; 1997 (in Dutch).
11. Coebergh JW, Crommelin MA, Kluck HM, van Beek M, van der Horst F, Verhagen-Teulings MT. [Breast cancer in southeast North Brabant and in North Limburg; trends in incidence and earlier diagnosis in an unscreened female population, 1975–1986.] *Nederlands Tijdschrift voor Geneeskunde*, 1990;134:760-5 (in Dutch).
12. Coebergh JWW, van der Heijden LH, Janssen-Heijnen MLG. *Cancer incidence and survival in the southeast of the Netherlands 1955–1994: a report from the Eindhoven Cancer Registry.* Eindhoven: Integraal Kankercentrum Zuid; 1995.
13. Bonneux L, Barendregt JJ, Looman CW, van der Maas PJ. Diverging trends in colorectal cancer morbidity and mortality: earlier diagnosis comes at a price. *European Journal of Cancer* 1995;31A:1665-71.
14. Post PN, Kil PJ, Crommelin MA, Schapers RF, Coebergh JW. Trends in incidence and mortality rates for prostate cancer before and after prostate-specific antigen introduction. A registry-based study in southeastern Netherlands, 1971–1995. *European Journal of Cancer* 1998;34:705-9.
15. Mackenbach JP, Van Duyne WM, Kelson MC. Certification and coding of two underlying causes of death in the Netherlands and other countries of the European Community. *Journal of Epidemiology and Community Health*. 1987;41:156-60.
16. Smith DW. Cancer mortality at very old ages. *Cancer* 1996;77:1367-72.
17. Brown BW, Brauner C, Minnotte MC. Noncancer deaths in white adult cancer patients. *Journal of the National Cancer Institute* 1993;85:979-87.
18. Berkel J. General practitioners and completeness of cancer registry. *Journal of Epidemiology and Community Health* 1990;44:121-4.
19. Schouten LJ, Hoppener P, van den Brandt PA, Knottnerus JA, Jager JJ. Completeness of cancer registration in Limburg, the Netherlands. *International Journal of Epidemiology* 1993;22:369-76.

Annex

The analytical model

For the disease process described in Fig. 1, the prevalence proportion at exact age n can be calculated from the prevalence at the previous age and mortality and incidence probability.

With respect to Fig. 1, $p_n = C_n/(C_n + S_n)$ is the prevalence at exact age n, and $m_n = (F_{n+1} - F_n)/(C_n + S_n)$ is the mortality probability of age group n, where S = susceptibles, C = cases and F = cause-specific deaths. Because of competing risks, the incidence probability of susceptibles at age n (i_n) does not have a straightforward expression in terms of the model compartments depicted in Fig. 1. Given p_n, m_n and i_n, we can calculate prevalence at age $n + 1$ using the following expression (7):

$$p_{n+1} = [p_n - m_n + i_n * (1 - p_n)]/(1 - m_n) \qquad \text{eq.}\{1\}$$

For this expression to be valid, a steady state situation and independence of all other causes of death must be assumed. Furthermore, the hazards are assumed to be constant within a given age interval. In order to minimize deviations from this assumption, we used 1-year age intervals. Therefore, we first interpolated the rates specified per 5-year interval to 1-year age groups, using the cubic-spline method.

The interpolated rates were then converted into the appropriate input formats, using the following expressions:

$$\begin{aligned} IRs_{n+1} &= IRp_{n+1}/(1 - p_n) \\ i_n &= 1 - \exp(- IRs_n) \qquad \text{eq.}\{2\} \\ m_n &= 1 - \exp(- MR_n), \text{ if } p_n > 0, \text{ otherwise } 0, \end{aligned}$$

where IRs_n is the incidence rate among susceptibles, IRp_n is the incidence rate in the population, and MR_n is the cause-specific mortality rate in the population, all for the age group n (the other parameters are mentioned above).

Calculation of the incidence among susceptibles from population data requires information on prevalence. Since prevalence itself has to be calculated in the model, we used the results from the previous age to calculate incidence among susceptibles. This produced a small deviation.

From these 1-year data the prevalence at exact age n was calculated using eq.{1}. We transformed the outcome of the model to a mean prevalence for age n by averaging the results of two successive ages. This can be given in either 1-year or 5-year age groups.

The DisMod model

The DisMod model can be downloaded from the Internet (at URL: http://www.hsph.harvard.edu/organizations/bdu/dismod/index.html). (Two versions of DisMod are available at this URL – we used DisMod I in our analyses.) For this model the same assumptions of steady state, constant hazards in the age interval and independence of other-cause mortality are required.

We approximated input hazards by rates. The appropriate input formats were calculated using the following expressions:

$$CFR_n = LN[1 - \{(m_n - r_n * (1 - p_n))/p_n\}], \text{ if } m_n < p_n, \text{ otherwise } 0 \ \text{eq.}\{3\}$$

$$r_n = [IR_n * (1 - \exp(- CFR_{n-1})) - CFR_{n-1} * (1 - \exp(- IR_n))]/[IR_n - CFR_{n-1}],$$

where CFR_n is the case-fatality rate for age group n and r_n is the probability of making two transitions in age group n (the other parameters are mentioned above).

Conversions to the appropriate input formats were made by using the interpolated data (1-year age intervals). We used the prevalence calculated by the previous model to compute case-fatality. Because the formula for calculating r_n requires the CFR_n as input data, and vice versa, we used the case-fatality rate of the previous age (CFR_{n-1}) in this calculation. This produced a slight deviation.

We then back-transformed the input data to 5-year age groups. DisMod was then used to calculate annual incidence and mortality rates and mean prevalence per 5-year age group.