TRANSITION INTENSITIES FOR A MODEL FOR PERMANENT HEALTH INSURANCE[1]

BY

ISABEL MARIA FERRAZ CORDEIRO[2]

ABSTRACT

The purpose of this paper is to obtain approximations to the transition intensities defined for a multiple state model for Permanent Health Insurance (PHI) which enables us to analyse PHI claims by cause of disability.

The approximations to the transition intensities are obtained using a set of PHI data classified by 18 sickness categories and the graduations of the transition intensities defined for a simpler model proposed in Continuous Mortality Investigation Reports, 12 (1991).

In order to derive the approximations to the recovery and mortality of the sick intensities for our model, we carry out tests of hypotheses based on the distributions of average sickness durations. The approximations to the sickness intensities are obtained by estimating a statistical model for the number of claim inceptions, which can be formulated as a generalized linear model.

KEYWORDS

Permanent Health Insurance, Multiple State Models, Transition Intensities, Analysis by Cause of Disability, Tests of Hypotheses, Generalized Linear Models

1. PRESENTATION OF THE PROBLEM

Permanent Health Insurance (PHI for brevity) is a class of long-term sickness insurance which provides cover against the risk of loss of income due to disability. In general terms, a PHI policy entitles the policyholder to an income during periods of disability longer than the deferred period of the policy.

[1] This research was supported by FCT – Fundação para a Ciência e Tecnologia, Portugal under program PRAXIS XXI.

[2] Escola de Economia e Gestão, Universidade do Minho, Portugal and CEMAPRE/ISEG, Universidade Técnica de Lisboa, Portugal.

ASTIN BULLETIN, Vol. 32, No. 2, 2002, pp. 319-346

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Each PHI policy has a deferred period, which is chosen by the policyholder when the policy is effected. Benefits only start to be paid after the end of the deferred period.

There are several types of PHI policy. However, in this paper we are only interested in individual conventional policies with level benefits. For a precise description of this type of policy see Cordeiro (1998).

We will assume throughout this paper that a policy expires when the policyholder reaches age 65 or dies, whichever occurs first. We will also assume that each policy has one of the following deferred periods: 1 week, 4 weeks, 13 weeks or 26 weeks (for brevity, throughout the paper we will refer to these deferred periods as D1, D4, D13 and D26, respectively).

Cordeiro (1998) has introduced a new multiple state model for PHI which enables us to analyse claims by cause of disability. This model is very useful in the underwriting and claims control stages of PHI business since it allows the calculation of quantities such as the average duration of a claim and claim inception rates by cause of disability.

This new model, which can be described intuitively by the diagram in Figure 1, has $(n+2)$ states $(n>1)$: Healthy (denoted by H), Dead (denoted by D), Sick with a Sickness from Class 1 (denoted by S_1), Sick with a Sickness from Class 2 (denoted by S_2), ..., Sick with a Sickness from Class n (denoted by S_n). Each state S_i represents a different class of causes of disability. These n states, considered together, group all possible causes of disability.

The important quantities for the model are the transition intensities ($\sigma(i)_x$, $\rho(i)_{x,z}$, $\nu(i)_{x,z}$ $(i = 1, 2, ..., n)$ and μ_x), since their action governs the movements of a policyholder between the $(n+2)$ states. The movements which the model assumes to be possible are represented by arrows in the diagram.

The transition intensities $\sigma(i)_x$ (for a fixed i) and μ_x, which can be designated as sickness intensity for class i and mortality of the healthy intensity, respectively, depend only on x, the policyholder's attained age. The transition intensities $\rho(i)_{x,z}$ and $\nu(i)_{x,z}$ (for a fixed i), which can be designated as recovery intensity and mortality of the sick intensity for class i, respectively, depend on x and on z, the duration of the policyholder's current sickness. The model assumes that all the transition intensities are continuous functions of either x or (x, z). As a consequence of this assumption, the transition intensities are bounded on any bounded set of values of x or (x, z).

The model mentioned in the previous paragraphs can be considered as a generalization of the multiple state model for PHI proposed in Continuous Mortality Investigation Reports, 12 (1991) (for brevity we will refer to this Report as CMIR 12 (1991) throughout the paper). In fact, it is easy to see this, if we compare the diagram in Figure 1 with the corresponding diagram for the latter model (see Figure A1 in CMIR 12 (1991)). The two diagrams are similar, the only difference being that the model in CMIR 12 (1991) has only three states: Healthy, Sick and Dead, and, therefore, in the latter diagram the n boxes, which represent the different classes of causes of disability, are replaced by only one box, which represents all possible causes of disability considered together. Consequently, in the model proposed in CMIR 12 (1991) only 4 transition intensities are defined: σ_x, μ_x, $\rho_{x,z}$ and $\nu_{x,z}$.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

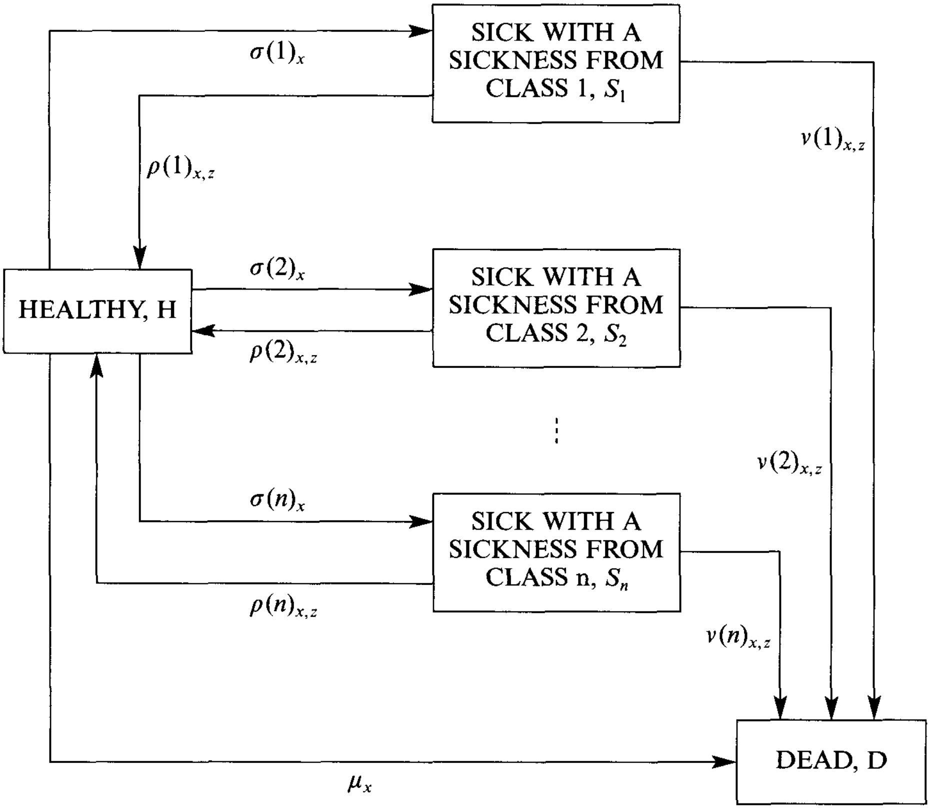

FIGURE 1: A multiple state model for the analysis of PHI claims by cause of disability.

All the important theoretical aspects of this new model have already been presented exhaustively elsewhere and, therefore, here we limit ourselves to mention those aspects which are concerned directly with the work in this paper. Cordeiro (1998, 2002) has: presented the mathematical basis of the model and defined the basic probabilities which are required for the calculation of more complex quantities concerning PHI business; presented formulae for the basic probabilities; derived numerical algorithms which make possible an efficient evaluation of some of the basic probabilities; and talked about the importance of the model for PHI business.

In order to make this new model operational, in the sense that it can be used to calculate quantities relevant to PHI business, we need to estimate the transition intensities. The purpose of this paper is to estimate the $\rho(i)_{x,z}$, the $v(i)_{x,z}$ and the $\sigma(i)_x$, i.e. the recovery intensities, the mortality of the sick intensities and the sickness intensities, respectively. We should note that the estimation of μ_x is not carried out in this paper. We will return to this matter in a later section.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

In CMIR 12 (1991, Parts A, B and C) the transition intensities σ_x, $\rho_{x,z}$ and $\nu_{x,z}$, defined for the simpler model mentioned above, were estimated and, subsequently, graduated by mathematical formulae, using a set of data from UK insurance companies, concerning individual PHI policies: the Standard Male Experience, 1975-78. In order to obtain graduations of the $\rho(i)_{x,z}$, $\nu(i)_{x,z}$ and $\sigma(i)_x$ the ideal situation would be to have available a set of data similar to the one used in CMIR 12 (1991), but classified into classes of causes of disability. Unfortunately, such a detailed set of data is not available (see Cordeiro (1998) for more details).

However, we have found a way of deriving continuous functions, which can be taken as approximations to the $\rho(i)_{x,z}$, $\nu(i)_{x,z}$ and $\sigma(i)_x$, from a set of PHI data, which is much less detailed than the one just mentioned, and the graduations of $\rho_{x,z}$, $\nu_{x,z}$ and σ_x obtained in CMIR 12 (1991). This set of data is presented in the next section.

In order to explain the basic idea behind the process by which we are going to obtain the approximations to the $\rho(i)_{x,z}$ and $\nu(i)_{x,z}$, let us consider, as an example, the case of the $\rho(i)_{x,z}$. Although $\rho(i)_{x,z}$ for a given i ($i = 1, 2, ..., n$) is a recovery intensity associated with a particular class of causes of disability whereas $\rho_{x,z}$ is a recovery intensity associated with the different classes of causes of disability taken as a whole, it is possible that they have common features. Since they are both recovery intensities, it is reasonable to expect that, for at least some classes, $\rho(i)_{x,z}$ has roughly the same shape as $\rho_{x,z}$.

Based on this idea, we are going to define and test a set of statistical hypotheses, using both the set of PHI data and the graduation of $\rho_{x,z}$ mentioned above. The results of these tests will enable us to derive the required approximations to the $\rho(i)_{x,z}$ for the different classes of causes of disability.

For obtaining the approximation to each $\sigma(i)_x$ we are going to estimate a model for the number of claim inceptions which assumes that $\sigma(i)_x$ is a function of σ_x. This model can be formulated as a generalized linear model.

2. PHI Data by Cause of Disability

Almost all the data which will be used to estimate the $\rho(i)_{x,z}$, the $\nu(i)_{x,z}$ and the $\sigma(i)_x$ are part of the following set of PHI data from UK insurance companies: the Cause of Disability Experience, Individual Standard Experience, 1979-82. This set of data was produced by the Continuous Mortality Investigation Bureau of the Institute of Actuaries and the Faculty of Actuaries (CMIB) and its main feature is that the claims, from which the information is extracted, are classified according to cause of disability. From this set of data only part of the male experience for deferred periods D1, D4, D13 and D26 will be used in this paper.

Due to limitations of space, it is not possible to present here all the data which will be used in the next sections. The full set of data can be found in Cordeiro (1998). Table 1 shows only a summary of part of these data. These data, in most cases, are classified by sickness category and age group.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

TABLE 1

NUMBER OF CLAIMS WHICH ENDED IN RECOVERY AND IN DEATH

Sickness Category	***Number of Claims which Ended in Recovery***				***Number of Claims which Ended in Death***			
	D1	**D4**	**D13**	**D26**	**D1**	**D4**	**D13**	**D26**
1. Other Infective	631	68	16	2	1	3	0	0
2. Malignant Neoplasms	67	27	13	8	25	20	36	19
3. Benign Neoplasms	44	18	5	3	1	3	2	2
4. Endocrine and Metabolic	30	7	6	2	1	1	5	0
5. Mental Illness	218	108	50	24	4	4	4	4
6. Nervous Disease	122	37	17	10	4	6	4	6
7. Heart/Circulating System	224	54	29	8	5	1	4	4
8. Ischaemic Heart Disease	168	115	74	23	6	5	5	9
9. Cerebro Vascular Disease	21	14	8	3	5	0	2	2
10. Acute Respiratory	953	30	2	3	1	0	1	1
11. Bronchitis Respiratory	396	28	9	2	1	4	2	1
12. Digestive	481	202	62	6	8	6	2	5
13. Genito-Urinary	211	43	5	2	3	2	2	2
14. Arthritis/Spondylitis	97	24	16	10	2	0	1	0
15. Other Musculoskeletal	565	159	70	15	4	0	0	1
16. R.T.A. Injuries	128	53	33	14	1	1	0	1
17. Other Injuries	598	183	59	18	1	0	0	0
18. All Others	469	85	51	11	0	3	2	2
All Sickness Categories	5423	1255	525	164	73	59	72	59

Almost all the data used in this paper are classified into 18 sickness categories, each of which corresponds to a specific group of diseases and injuries (see Table 1). These 18 sickness categories were obtained by amalgamating the 70 'causes for tabulation of morbidity' which form the following classification of causes of disability: List C; Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death; Eighth Revision; World Health Organization; 1967 (this classification can be found in CMIR 8 (1986)).

Most data are also classified into 4 age groups: 18-39, 40-49, 50-59 and 60-64. For convenience, and because the sickness categories are also numbered from 1 to 18 (see Table 1), from now on we will designate these age groups: age group 1, age group 2, age group 3 and age group 4, respectively.

The data which will be used to estimate the $p(i)_{x,z}$ concern PHI claims which ended in recovery during the period of investigation (1979-82). Some of these claims were already in force at the beginning of the period of investigation.

For each combination of deferred period and sickness category considered in Table 1 we will use the following data: the number of claims for each age group, the total number of claims (i.e. the number of claims for all age groups) and the average duration (in weeks) of a sickness (i.e. the average duration of a claim plus the deferred period). Table 1 shows only the total number of claims for each combination of deferred period and sickness category.

The policyholder's age corresponding to each claim which ended in recovery, needed to make the classification by age group, has been calculated as the age nearest birthday at the date of falling sick.

The set of data which will be used to estimate the $\nu(i)_{x,z}$ is of the same type of the one just described but, in this case, for claims which ended in death during the period of investigation. In Table 1 we present also only the total number of claims for each combination of deferred period and sickness category.

From the set of data by cause of disability the data which will be used in the estimation of the $\sigma(i)_x$ concern claims which started during the period of investigation, i.e. claim inceptions during the period of investigation. Some of these claims did not end during the period of investigation. For each combination of deferred period and sickness category considered in Table 1 we will use the following data: the number of claim inceptions for each age group and the total number of claim inceptions (i.e. the number of claim inceptions for all age groups).

The claim inceptions are classified into the 4 age groups we consider by age nearest birthday at the 1st January immediately preceding the date when claim payments started (which is broadly equivalent to age last birthday when claim payments started). For D1 it was assumed that claim payments started at the beginning of the sickness rather than at the end of the deferred period as for the other deferred periods.

Claims arising from duplicate policies were removed from the set of data presented above.

Since almost all the data which will be used in this paper are classified into 18 sickness categories, we have decided that the number of states which represent classes of causes of disability to be defined in our model is $n = 18$. Therefore, from now on, we will designate by $\rho(i)_{x,z}$, $\nu(i)_{x,z}$ and $\sigma(i)_x$ the recovery intensity, the mortality of the sick intensity and the sickness intensity (respectively) for sickness category i, where $i = 1, 2, ..., 18$. For a given deferred period, we will derive approximations to $\rho(i)_{x,z}$, $\nu(i)_{x,z}$ and $\sigma(i)_x$ for each of these 18 sickness categories.

3. Obtaining Approximations to the Recovery and Mortality of the Sick Intensities

3.1. Modeling the Average Duration of a Sickness

In the following paragraphs we present the notation and assumptions necessary to define the quantities and random variables which describe the set of PHI data presented in Section 2.

All the quantities and random variables we define in this section depend on the deferred period we are considering: D1, D4, D13 or D26. We decided to omit the deferred period in the notation in order not to make it too cumbersome.

We will denote the deferred period by d when it appears in formulae and, in those situations, it will be measured in years.

The sickness categories and age groups to which we refer throughout this section (and the remaining sections of this paper) are those defined in Section 2.

The notation we introduce in the following paragraphs concerns claims which ended in recovery. This is the reason why we use the superscript r in this notation.

We define n_i^r $(i = 1, ..., 18)$ to be the total number of claims for sickness category i and n_{ij}^r $(i = 1, ..., 18; j = 1, 2, 3, 4)$ to be the number of claims for sickness category i and age group j. Thus, we can write $n_i^r = \sum_{j=1}^4 n_{ij}^r$.

Assuming we can number the n_{ij}^r claims for sickness category i and age group j, we denote by T_{ijk}^r $(i = 1, ..., 18; j = 1, 2, 3, 4; k = 1, ..., n_{ij}^r)$ the random variable that represents the duration of the sickness corresponding to claim k in this category and age group. This random variable only takes on values greater than or equal to d, since to make a claim a policyholder must stay sick for at least the deferred period of his policy. On the other hand, since it is unlikely that an insurance company will continue to record the duration of a sickness after the policy expires, we assume that the variable T_{ijk}^r only takes on values less than or equal to the difference between 65 and the policyholder's age at the beginning of the sickness.

We assume, for obvious reasons, that the variables T_{ijk}^r for different sickness categories are independent.

We assume also that, for a given sickness category i, the variables T_{ijk}^r for different age groups are independent. This is a reasonable assumption to make considering, as we have seen in Section 2, that claims arising from duplicate policies were removed from the set of data we are going to use.

We have seen in Section 2 that in the set of data we are going to work with we only have available, for each category i, the number of claims for each age group j. We do not know, for each claim in an age group, the policyholder's precise age at the beginning of the corresponding sickness. As the distribution of any T_{ijk}^r depends on this precise age, we assume that the variables T_{ij1}^r, T_{ij2}^r, ..., $T_{ijn_{ij}^r}^r$ for a given category i and a given age group j, are i.i.d. with the same distribution as the duration of a sickness in category i for a claimant aged x_j at the beginning of the sickness, where x_j is the midpoint of the age interval associated with age group j.

Considering the assumption just introduced, we can write the distribution function of any T_{ijk}^r, for sickness category i and age group j, in terms of $\rho(i)_{x,z}$ and $v(i)_{x,z}$:

$$F_{T_{ijk}^r}(t) = \begin{cases} 0 & \text{for } t \leq d \\ \dfrac{\int_d^t {}_s p_{x_j}^{\overline{S_i S_i}} \, \rho(i)_{x_j+s,\, s} \, ds}{\int_d^{65-x_j} {}_s p_{x_j}^{\overline{S_i S_i}} \, \rho(i)_{x_j+s,\, s} \, ds} & \text{for } d < t \leq 65 - x_j \\ 1 & \text{for } t > 65 - x_j \end{cases} \tag{1}$$

where

$${}_s p_{x_j}^{\overline{S_i S_i}} = \exp\left\{-\int_0^s \left(\rho(i)_{x_j+u,u} + \nu(i)_{x_j+u,u}\right) du\right\}$$

is the probability of a policyholder staying sick, with a sickness from category i, from age x_j to at least age $(x_j + s)$, given that he fell sick at age x_j. ${}_s p_{x_j}^{\overline{S_i S_i}}$ is a basic probability for our model and the derivation of its formula can be found in Cordeiro (1998, 2002). For a more detailed explanation of the expression of $F_{T_{ijk}^r}(t)$ see also Cordeiro (1998).

Denoting by $\overline{T}_i^r$ the average of the durations of the sicknesses corresponding to all the claims for category i, i.e. the claims for category i and all the age groups, we have:

$$\overline{T}_i^r = \frac{\sum_{j=1}^4 \sum_{k=1}^{n_{ij}^r} T_{ijk}^r}{n_i^r} \qquad i = 1, \ldots, 18 \qquad (2)$$

Recall from Section 2 that this random variable represents the average duration of a sickness for category i. From now on, we will use this simpler designation for $\overline{T}_i^r$. $\overline{T}_i^r$ only takes on values in the interval $[d, 65 - x_1]$ since $x_1 = \min\{x_j, j = 1, 2, 3, 4\}$.

Now, for each random variable introduced so far, let us define a similar one but for claims which ended in death. We use a similar notation to denote these new variables, the only difference being the superscript r, which now is replaced by the superscript m. We make also the same assumptions about these new variables as those we have made about the variables for claims which ended in recovery.

As far as the variables for claims which ended in death are concerned, the only significant difference we should note is the distribution function of T_{ijk}^m, the duration of the sickness (which ended in death) corresponding to claim k in category i and age group j, which is given by formula (1) with $\rho(i)_{x_j+s,s}$ replaced by $\nu(i)_{x_j+s,s}$.

As we have mentioned above, we assume that the variables T_{ijk}^r for a given sickness category i are independent (either when they are associated with claims in the same age group or with claims in different age groups). On the other hand, it is easy to see that the variances of these variables are finite. Despite the fact that these variables are only identically distributed within each age group, we can apply the central limit theorem to obtain an approximation to the distribution of their mean, i.e. the distribution of $\overline{T}_i^r$. Hence, considering (2), we can state that, provided n_i^r is large, the variable $\overline{T}_i^r$ has approximately the following normal distribution:

$$N\left(\frac{\sum_{j=1}^4 n_{ij}^r E T_{ij}^r}{n_i^r}, \frac{\sum_{j=1}^4 n_{ij}^r V T_{ij}^r}{(n_i^r)^2}\right) \qquad (3)$$

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

where $ET^r_{ij} = E(T^r_{ijk})$ and $VT^r_{ij} = V(T^r_{ijk})$ (the reason for using these notations is that $E(T^r_{ijk})$ and $V(T^r_{ijk})$ do not depend on k). For more details about this approximate distribution see Cordeiro (1998).

Similarly, an approximate distribution for $\overline{T}^m_i$ is the distribution (3) with the superscript r replaced by the superscript m.

3.2. Tests of Hypotheses to Decide on the Shapes and Levels of the Recovery Intensities

In the present section we propose tests of hypotheses to investigate, for each sickness category, whether the recovery intensities $\rho(i)_{x,z}$ for the 4 deferred periods we consider have the same shapes as the corresponding recovery intensities $\rho_{x,z}$.

More formally, for each sickness category i and each of the 4 deferred periods we consider, we want to test the null hypothesis

$$H_0 : \rho(i)_{x,z} = k_i \rho_{x,z} \tag{4}$$

against the alternative

$$H_a : \rho(i)_{x,z} \neq k_i \rho_{x,z} \quad \text{for any } k_i \tag{5}$$

where k_i is a positive constant factor which allows for the possibility of $\rho(i)_{x,z}$ having a different level than $\rho_{x,z}$. We should note that in (4) and (5) we assume that the factor k_i is the same for the 4 deferred periods we consider. This assumption has to do with the points in the following paragraphs.

In order to describe some of the features of the graduations of the $\rho_{x,z}$ which are relevant to this paper, it is convenient to consider them as functions of the policyholder's age at the date of falling sick, y, and of the duration of sickness, z, instead of functions of x and of z. We should point out that in CMIR 12 (1991) both $\rho_{x,z}$ and $\nu_{x,z}$ are regarded as $\rho_{y+z,z}$ and $\nu_{y+z,z}$ respectively. Note that the two different notations are consistent. From CMIR 12 (1991) we can also see that the graduation of $\rho_{y+z,z}$ is different according to the deferred period we consider. For a fixed y, the graduations of $\rho_{y+z,z}$ for D4, D13 and D26, when compared with the one for D1, have 4 week 'run-in' periods of lower recovery intensities, immediately after the end of their respective deferred periods, due to the fact that some sicknesses which do not last much longer than the deferred period are not reported. After the first 4 weeks of sickness that follow their respective deferred periods, the graduations for D4, D13 and D26 are equal to the graduation for D1. For more details about these features see CMIR 12 (1991) or Cordeiro (1998).

We assume that the approximations to the $\rho(i)_{y+z,z}$ have 'run-in' patterns similar to those in the graduations of the corresponding $\rho_{y+z,z}$. This assumption is a consequence of another assumption we make: the 'non-reported' claims are distributed more or less uniformly among the different sickness categories we consider.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Note that, since the approximate distribution of $\overline{T}_i^r$, given by (3), depends only on the n_{ij}^r and on $\rho(i)_{x,z}$ and $v(i)_{x,z}$, in the cases where n_i^r is large, this distribution can be used to define a two-tailed test to test H_0 against H_a.

However, if we analyse Table 1, we can see that, in many cases, n_i^r is not large. In fact, for D13 and D26, the n_i^r for the vast majority of the categories are less than 30. This also happens for a few categories in the cases of D1 and D4. Furthermore, as we will see below, we can conclude that the test proposed in the previous paragraph is not adequate even for some categories with n_i^r much higher than 30.

It is possible to obtain very close approximations to the distributions of the $\overline{T}_i^r$ using simulation. In fact, since we know, for a given category i and a given deferred period d, the distributions of the durations of individual sicknesses in the 4 age groups we consider (see Section 3.1), we can simulate a very large number of observations of the corresponding variable $\overline{T}_i^r$ and then use these simulated observations, as we would use actual observations, to estimate the distribution function or the density of $\overline{T}_i^r$ by some appropriate method. A full description of the process used to simulate an observation of a given $\overline{T}_i^r$ can be found in Cordeiro (1998).

Using simulation, we have produced graphs showing the densities of some variables T_{ijk}^r and also of some $\overline{T}_i^r$ (both with $n_i^r < 30$ and with n_i^r much higher than 30). From these graphs we have concluded that: in general, the densities of the T_{ijk}^r are heavily skewed to the right; the distributions of the $\overline{T}_i^r$ for many categories with $n_i^r < 30$ are quite skewed and even some $\overline{T}_i^r$ with n_i^r much higher than 30 have distributions which are also quite skewed. Some of the graphs just mentioned are shown in Cordeiro (1998).

After concluding that we need to propose more adequate tests than those based on the central limit theorem, we are going to show below how these new tests can be defined using the distributions of the $\overline{T}_i^r$ obtained by simulation. More precisely, these tests are based on the empirical cumulative distribution functions (e.c.d.f.) of the simulated samples of the $\overline{T}_i^r$.

Assume we have simulated n observations (n being large) of the variable $\overline{T}_i^r$ for a given category i and a given deferred period d, when H_0 is true, and let us denote the e.c.d.f. of the simulated sample by $\hat{F}_{\overline{T}_i^r}(t)$. Assuming we are going to use a significance level $\alpha = 0.05$, as above, we think that also in this case the most adequate test is a two-tailed test, with 2.5% of the total probability located in each of the tails of the distribution. Then, we can propose the following test: we reject H_0 in favour of H_a at the significance level $\alpha = 0.05$ if the observed value of $\overline{T}_i^r$, denoted by $\bar{t}_i^r$, lies in the rejection region given by the interval:

$$R_i' = (-\infty, t_1) \cup (t_2, +\infty)$$

with t_1 and t_2 satisfying the equations:

$$\hat{F}_{\overline{T}_i^r}(t_1) = 0.025 \tag{6}$$

$$\hat{F}_{\overline{T}_i^r}(t_2) = 0.975 \tag{7}$$

respectively.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Let us denote the values of the n simulated observations of $\overline{T}_i^r$ arranged in increasing order by $\bar{t}_{i(1)}^r, \bar{t}_{i(2)}^r, ..., \bar{t}_{i(n)}^r$. Note that, as the estimate of the distribution function of $\overline{T}_i^r$, for a given t, obtained using the e.c.d.f. of the simulated sample can be defined by:

$$\hat{F}_{\overline{T}_i^r}(t) = \frac{\max\left\{l : \bar{t}_{i(l)}^r \leq t\right\}}{n}$$

it is obvious that neither t_1 nor t_2, satisfying equations (6) and (7) respectively, are unique. Therefore, assuming that we choose n such that $(0.025n)$ and $(0.975n)$ are integers, one of the rejection regions that can be proposed is:

$$R_i^{'} = \left(-\infty, \bar{t}_{i(0.025n)}^r\right) \cup \left(\bar{t}_{i(0.975n)}^r, +\infty\right) \tag{8}$$

Considering that, with only a few exceptions, the n_i^r for D1 are large (see Table 1) and also that the test proposed in the previous paragraph is very heavy in computational terms when the corresponding n_i^r is large, we have decided to use this test only for deferred periods D4, D13 and D26. For testing the hypotheses for D1 we will use the test based on distribution (3), mentioned above.

In order to test the hypotheses concerning the $\rho(i)_{x,z}$ for a given sickness category i, the value of k_i has to be chosen in some way. The most adequate value of k_i to test these hypotheses is not known in advance.

The method we propose to choose the value of k_i which best represents the level of the $\rho(i)_{x,z}$ is to choose, from among the values of k_i for which none of the null hypotheses for the 4 deferred periods is rejected in favour of the respective alternative, the one which maximizes the likelihood function for the value $\bar{t}_i^r$ observed for D1. This likelihood function, which we denote by $L^*(k_i)$, is defined as the density of $\overline{T}_i^r$ for D1, where it is assumed that H_0 is true, the density is evaluated at the corresponding $\bar{t}_i^r$ and the parameter k_i is considered as a variable. The reasons for proposing this method can be found in Cordeiro (1998).

The reason for not proposing a method similar to this one but which uses the likelihood function for the values $\bar{t}_i^r$ observed for the 4 deferred periods we consider is that, since the exact density functions of the $\overline{T}_i^r$ are not known and the corresponding approximate distributions (3) are not valid for many sickness categories in the cases of deferred periods D4, D13 and D26 (as we have seen above), it is not possible to obtain a valid likelihood function for the majority of the categories.

We apply the method proposed above in the following way: for each category i, firstly we find the value of k_i which maximizes $L^*(k_i)$ (we are referring here to a maximization without any constraints) and, then, for that value of k_i, we test H_0 against H_a for the 4 deferred periods we consider. In the case of none of the 4 null hypotheses being rejected, we can propose the functions $(k_i\ \rho_{x,z})$, with that value of k_i, as the approximations to the $\rho(i)_{x,z}$. Otherwise, we should search for other values of k_i for which none of the 4 null hypotheses is rejected and, in the case of their existence, we then have to apply the method exactly as it is described above.

After carrying out some calculations, we found out that, in general, a good approximation to the value of k_i which maximizes $L^*(k_i)$ (without any constraints) is the value of k_i which satisfies the following equation for D1:

$$\bar{t}_i^r = \frac{\sum_{j=1}^{4} n_{ij}^r ET_{ij}^{r*}}{n_i^r} \tag{9}$$

where the right-hand side of the equation is the expected value of both the actual distribution and the approximate normal distribution of $\overline{T}_i^r$, when H_0 is true (ET_{ij}^{r*} is ET_{ij}^r with $\rho(i)_{x,z}$ replaced by (k_i $\rho_{x,z}$). Note that this approximation is the method-of-moments estimate of k_i. We have decided to use this approximation in our work because it implies simpler calculations.

3.3. Results of the Tests for the Recovery Intensities

From Section 3.2 we can see that, apart from the values of the k_i, the only elements missing to carry out the tests concerning the $\rho(i)_{x,z}$ are the graduations of the $\nu(i)_{x,z}$. As we have seen in Section 1, in order to obtain the approximations to the $\nu(i)_{x,z}$, we are also going to carry out tests, using the data concerning claims which ended in death, presented in Section 2, and the graduation of $\nu_{x,z}$ proposed in CMIR 12 (1991). From now on, when we refer to the graduation of $\nu_{x,z}$, we mean this particular graduation. This graduation is the same for all 4 deferred periods we consider.

Since to carry out the tests to obtain the approximations to the $\nu(i)_{x,z}$, we need to have the graduations of the $\rho(i)_{x,z}$, as we will see in the next section, we are facing here a vicious circle. Therefore, at this stage, to carry out the tests concerning the $\rho(i)_{x,z}$, the only solution is to propose temporary approximations to the $\nu(i)_{x,z}$.

We have decided to propose the graduation of $\nu_{x,z}$ as the temporary approximation to $\nu(i)_{x,z}$ required for all the deferred periods and sickness categories we consider. The reasons for this decision and other details concerning the tests for the $\rho(i)_{x,z}$, which are not given here, can be found in Cordeiro (1998).

As far as the tests for D1 are concerned it is important to note that, for any category i and the value of k_i which satisfies equation (9), H_0 for D1 is automatically not rejected in favour of H_a (see the rejection region associated with the test based on distribution (3)). Therefore, in the cases where we use the value of k_i just mentioned, we do not need to present the result of the test for D1.

Considering rejection region (8) and that, in order to obtain the distribution of a given $\overline{T}_i^r$, we simulate 10000 observations of this variable, when testing the corresponding H_0, if we find that $\bar{t}_i^r$ is located among the 250 smallest values of the simulated observations or among the 250 greatest values of the simulated observations, we should reject this hypothesis in favour of the corresponding H_a.

After having carried out all the tests concerning the $\rho(i)_{x,z}$, we have concluded that, except for sickness categories 12 and 17, for each category i we

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

consider and the value of k_i which satisfies equation (9) for D1, none of the null hypotheses for the 4 deferred periods is rejected in favour of the respective alternative. Since, for each of the categories 12 and 17 and the value of k_i just mentioned, H_0 for D4 is rejected in favour of H_a, in these cases we had to search for other values of k_i for which none of the null hypotheses for the 4 deferred periods is rejected. In both cases these values of k_i exist and, therefore, for each of the two categories, we had to choose, from among these values, the one at which the corresponding $L^*(k_i)$ takes on the highest value.

The results of the tests for D4, D13 and D26 are displayed in Table 2. This table shows, for each of the 18 categories we consider, (a close approximation to) the value of k_i which maximizes $L^*(k_i)$ from among those for which none of the null hypotheses for the 4 deferred periods is rejected and the position of $\bar{t}_i^r$ among the values of the 10000 simulated observations of $\overline{T}_i^r$ arranged in increasing order for D4, D13 and D26. In this table, when we say that the position of $\bar{t}_i^r$ is s, we mean that $\bar{t}_i^r$ lies between the (s)th and the $(s+1)$th smallest values of the simulated observations.

Analysing Table 2, we can see that, in fact, for D4, D13 and D26, all the $\bar{t}_i^r$ are located between the 250th and the 9750th smallest values of the simulated observations of the corresponding $\overline{T}_i^r$.

TABLE 2

RESULTS OF THE TESTS CONCERNING THE $\rho(i)_{x,z}$ FOR D4, D13 AND D26. $\rho(i)_{x,z} = k_i\,\rho_{x,z}$. $\nu(i)_{x,z} = \nu_{x,z}$

Sickness Category	k_i	***Position of $\bar{t}_i^r$ Among the 10000 Simulated Observations***		
		D4	**D13**	**D26**
1. Other Infective	1.45	5309	5822	3779
2. Malignant Neoplasms	0.6	5108	2252	312
3. Benign Neoplasms	1.45	7280	470	9455
4. Endocrine and Metabolic	0.4	1463	4489	7638
5. Mental Illness	0.35	1274	2359	1155
6. Nervous Disease	0.9	5461	7068	6227
7. Heart/Circulating System	0.85	4345	3711	7149
8. Ischaemic Heart Disease	0.4	1551	670	819
9. Cerebro Vascular Disease	0.65	8691	8622	7446
10. Acute Respiratory	2.25	7936	5001	9672
11. Bronchitis Respiratory	1.65	8603	8872	4770
12. Digestive	1.2	312	2823	2621
13. Genito-Urinary	1.35	5499	5104	7102
14. Arthritis/Spondylitis	0.5	476	5306	3677
15. Other Musculoskeletal	1.0	274	4043	3022
16. R.T.A. Injuries	0.8	5746	5707	1496
17. Other Injuries	1.05	339	1252	2334
18. All Others	1.0	1415	3949	4705

TABLE 3

RESULTS OF THE TESTS CONCERNING $\rho(12)_{x,z}$ AND $\rho(17)_{x,z}$ FOR D1. $\rho(i)_{x,z} = k_i\, \rho_{x,z}$. $v(i)_{x,z} = v_{x,z}$

Sickness Category	k_i	$\bar{t}_i^r$	$E(\bar{T}_i^r)$	$\sqrt{V(\bar{T}_i^r)}$	p-value
12. Digestive	1.2	5.5	4.546	0.635	0.134
17. Other Injuries	1.05	4.9	4.573	0.745	0.66

Since, for each of the categories 12 and 17, the value of k_i chosen is not the one which satisfies equation (9) for D1, in both cases we should also present the result of the test (based on the central limit theorem) for D1. The results of these tests are presented in Table 3. In this table $\bar{t}_i^r$, $E(\bar{T}_i^r)$ and $\sqrt{V(\bar{T}_i^r)}$ are expressed in weeks.

From the results of the tests concerning the $\rho(i)_{x,z}$, we can conclude that, for each category i ($i = 1, \ldots, 18$), we can propose the functions ($k_i\, \rho_{x,z}$) for the 4 deferred periods we consider as the required approximations to the corresponding $\rho(i)_{x,z}$, where the values of the k_i which specify these approximations are presented in Table 2. However, we will only be able to propose these functions as the definitive approximations to the $\rho(i)_{x,z}$, after carrying out the necessary tests and concluding that the graduation of $v_{x,z}$ is a reasonable approximation to the $v(i)_{x,z}$ for the different categories.

4. CHECKING THE APPROXIMATIONS TO THE MORTALITY OF THE SICK INTENSITIES

In the present section we check if we can propose definitively the graduation of $v_{x,z}$ as the approximation to $v(i)_{x,z}$ for all the deferred periods and sickness categories we consider.

We want to test, for each sickness category i, the null hypothesis

$$H_0 : v(i)_{x,z} = v_{x,z}$$

against the alternative

$$H_a : v(i)_{x,z} \neq v_{x,z}$$

for each of the 4 deferred periods we consider. Note that, since the graduation of $v_{x,z}$ is the same for all 4 deferred periods we consider, for a given category i, H_0 and H_a are also the same for the 4 deferred periods.

Although, in general, the distributions of the T_{ijk}^m (the durations of individual sicknesses which ended in death) are less heavily skewed to the right than those of the T_{ijk}^r, since most of the n_i^m (all, except those for sickness category 2) are much smaller than 30 (see Table 1), the distributions of most of the $\bar{T}_i^m$ are still quite skewed (see Cordeiro (1998) for more details about this matter).

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Taking into consideration the points discussed in the previous paragraph, we are going to obtain the distributions of the $\bar{T}_i^m$ using simulation and to base the tests, to test the hypotheses concerning the $v(i)_{x,z}$, on these distributions. The rejection region which specifies the test concerning each $v(i)_{x,z}$ is (8) with the superscript r replaced by the superscript m and $n = 10000$.

We present the results of the tests concerning the $v(i)_{x,z}$ for all the deferred periods and sickness categories we consider in Table 4, which is similar to Table 2.

Analysing Table 4, we can see that, for sickness category 2, the null hypotheses associated with the 4 deferred periods are all rejected and that the results strongly indicate that $v(2)_{x,z}$ (regardless of its shape) has a higher level than $v_{x,z}$. We were expecting this, since, as we have mentioned above, the n_i^m for category 2 are much higher than the n_i^m for the other categories (see Table 1). Therefore, we cannot consider the graduation of $v_{x,z}$ as the definitive approximation to $v(2)_{x,z}$. We will return to this matter at the end of this section.

From this table we can also see that there are three more cases where H_0 is rejected. For D1, the cases of categories 1 and 13. For D13, the case of category 5. There is also a case where, despite H_0 not being rejected, $\bar{t}_i^m$ is very close to the rejection region: the case of D1 and category 8.

TABLE 4

RESULTS OF THE TESTS CONCERNING THE $v(i)_{x,z}$ FOR ALL SICKNESS CATEGORIES WITH $n_i^m > 0$ AND ALL DEFERRED PERIODS. $\rho(i)_{x,z} = k_i\,\rho_{x,z}$ (THE VALUES OF THE k_i ARE SHOWN IN TABLE 2). $v(i)_{x,z} = v_{x,z}$

Sickness Category	*Position of $\bar{t}_i^m$ Among the 10000 Simulated Observations*			
	D1	D4	D13	D26
1. Other Infective	9888	2249	–	–
2. Malignant Neoplasms	$\bar{t}_2^m < \bar{t}_{2(1)}^m$	1	1	$\bar{t}_2^m < \bar{t}_{2(1)}^m$
3. Benign Neoplasms	8298	5888	3789	7761
4. Endocrine and Metabolic	7809	4579	2097	–
5. Mental Illness	7136	2990	9958	5668
6. Nervous Disease	6623	8033	3289	1541
7. Heart/Circulating System	4074	6846	1249	3300
8. Ischaemic Heart Disease	252	1633	4076	2961
9. Cerebro Vascular Disease	8852	–	3112	3246
10. Acute Respiratory	4847	–	8312	4525
11. Bronchitis Respiratory	8933	7717	8749	8310
12. Digestive	1840	5954	6214	2809
13. Genito-Urinary	9840	4069	557	5488
14. Arthritis/Spondylitis	3475	–	2487	–
15. Other Musculoskeletal	5281	–	–	7896
16. R.T.A. Injuries	8516	1992	–	5695
17. Other Injuries	6380	–	–	–
18. All Others	–	1547	3656	4455

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Despite these results, we decided that it is reasonable to still propose the graduation of $\nu_{x,z}$ as the definitive approximation to $\nu(1)_{x,z}$, $\nu(5)_{x,z}$ and $\nu(13)_{x,z}$. The reasons for our decision are twofold. Firstly, we should consider the fact that the three null hypotheses rejected are associated with three different sickness categories. If two (or all) of the null hypotheses rejected were associated with the same category we should have not considered the graduation of $\nu_{x,z}$ as an adequate approximation to $\nu(i)_{x,z}$ for this category. Secondly, we should bear in mind that, as we are using a significance level $\alpha = 0.05$, it is possible that we are rejecting H_0, when it is true, in 5% of the cases. This means that, since we have carried out 54 independent tests (without considering the tests for category 2), it is quite reasonable to expect having approximately three null hypotheses rejected, despite their being true.

In conclusion, we propose the graduation of $\nu_{x,z}$ as the definitive approximation to the $\nu(i)_{x,z}$ for all sickness categories, except for category 2.

As far as sickness category 2 is concerned, the results of the tests indicate that we should try to find an approximation to $\nu(2)_{x,z}$ with the same shape as the graduation of $\nu_{x,z}$ but a higher level or one with a different shape from the graduation of $\nu_{x,z}$ that satisfies: $\nu(2)_{x,z} > \nu_{x,z}$ for all (x, z).

We have chosen to obtain the former approximation to $\nu(2)_{x,z}$ just mentioned. Only in the case of this approximation being rejected, would we then try to obtain the latter approximation. This approximation and new approximations to the $\rho(2)_{x,z}$ for the 4 deferred periods we consider have been obtained by an iterative process of hypotheses testing. The reason for having to carry out this process is the fact that, as we have seen in Section 3.3, to carry out the tests concerning the $\nu(i)_{x,z}$, we need to have the graduations of the $\rho(i)_{x,z}$ and, conversely, to carry out the tests concerning the $\rho(i)_{x,z}$, we need to have the graduations of the $\nu(i)_{x,z}$.

Again, due to limitations of space, it is not possible to present here the details and the intermediate results of the iterative process just mentioned. See Cordeiro (1998) for a fuller description of this process. At the end of the process we have obtained the following definitive approximations for each of the 4 deferred periods we consider: the function ($0.01\ \rho_{x,z}$) as the approximation to $\rho(2)_{x,z}$ and the function ($13.55\ \nu_{x,z}$) as the approximation to $\nu(2)_{x,z}$.

5. Obtaining Approximations to the Sickness Intensities

5.1. Modeling the Number of Claim Inceptions

In this section we present the statistical model for the number of claim inceptions which is going to be used in a later section to obtain the approximations to the $\sigma(i)_x$.

Although all the new random variables and other quantities which are introduced in this section depend also on the deferred period, we omit it in the corresponding notation.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Before presenting the model mentioned above we should explain how we can obtain, for a given deferred period, the probability that a sickness from a given category leads to a claim.

Recall from Section 3.2 that the approximations to the $\rho(i)_{y+z,z}$ for the 4 deferred periods we consider and each category i have the same shapes as the graduations of the corresponding $\rho_{y+z,z}$. This means that the approximations to the $\rho(i)_{y+z,z}$ for D4, D13 and D26, when compared with the approximation to $\rho(i)_{y+z,z}$ for D1, have 4 week 'run-in' periods of lower recovery intensites, immediately after the end of their respective deferred periods, due to a phenomenon of 'non-reported claims'.

As we have seen in Section 3.1, for a policyholder aged x at the beginning of a sickness from category i, the probability that the sickness lasts for at least the deferred period, d, is ${}_{d}p_{x}^{\overline{S_i S_i}}$. For D1, where all potential claims are assumed to be reported, this is the probability that the sickness results in a claim. For D4, D13 and D26, where not all potential claims are reported, the probability that the sickness leads to a claim is

$${}_{d}p_{x}^{\overline{S_i S_i}}\, r(i)_x$$

where $r(i)_x$ is the probability that a sickness from category i, beginning at age x and lasting to at least the end of deferred period, d, is reported and hence becomes a claim. From CMIR 13 (1993), where a probability similar to $r(i)_x$ has originally been presented, we can deduce that (see also Cordeiro (1998) for this result)

$$\begin{aligned} r(i)_x &= \frac{{}^{\mathrm{D1}}_{(4/52.18)}p_{x+d,d}^{\overline{S_i S_i}}}{{}^{\mathrm{Ds}}_{(4/52.18)}p_{x+d,d}^{\overline{S_i S_i}}} \\ &= \exp\left\{\int_{d}^{d+\frac{4}{52.18}}\left(\rho(i)_{x+z,z}^{\mathrm{Ds}}-\rho(i)_{x+z,z}^{\mathrm{D1}}\right)dz\right\} \end{aligned}$$

where

$${}_{t}p_{x,z}^{\overline{S_i S_i}} = \exp\left\{-\int_{0}^{t}\left(\rho(i)_{x+s,z+s}+\nu(i)_{x+s,z+s}\right)ds\right\}$$

is the probability of a policyholder remaining sick until at least age $(x+t)$ given that he is sick at age x with a sickness from category i and with duration of sickness z (note that ${}_{t}p_{x}^{\overline{S_i S_i}}$ is the particular case of this basic probability where $z=0$), Ds is the notation we use in the text to denote deferred period d, ${}^{\mathrm{Ds}}_{(4/52.18)}p_{x+d,d}^{\overline{S_i S_i}}$ is the probability ${}_{(4/52.18)}p_{x+d,d}^{\overline{S_i S_i}}$ calculated with $\rho(i)_{x,z}$ for Ds and $\rho(i)_{x+z,z}^{\mathrm{Ds}}$ is $\rho(i)_{x+z,z}$ for Ds (Ds = D1, D4, D13, D26). Note that $r(i)_x$ can also be defined for D1 but, in this case, $r(i)_x = 1$.

For a given observation period and a given deferred period d, let us denote by

$$I_{i(x)} \quad i = 1, \ldots, 18;\ x = 18, \ldots, 64$$

the number of sicknesses from category i which start in the observation period, for which the policyholder is aged between x and $(x + 1)$ at the start of the sickness, which last beyond the deferred period and become claims.

Using a theorem presented in Hoem (1987), we propose the following statistical model for $I_{i(x)}$:

$$I_{i(x)} \overset{a}{\sim} Po\left(\int_x^{x+1} E(y)\sigma(i)_y \, {}_dp_y^{\overline{S_iS_i}} r(i)_y dy\right) \tag{10}$$

(i.e. asymptotically, $I_{i(x)}$ has a Poisson distribution with the parameter given within parentheses), where $E(y)$ is the total time spent in the observation period by policyholders who are healthy and aged y (this quantity is more commonly designated by exposure at age y). $I_{i(x)}$ has not an exact Poisson distribution due to the fact that $E(y)$ $(x \le y < x + 1)$ is a random variable and not pre-determined. For more details concerning the distribution of $I_{i(x)}$ see CMIR 12 (1991), Hoem (1987), Macdonald (1996) and Sverdrup (1965).

In CMIR 12 (1991) the transition intensities σ_x were estimated using a model for the number of claim inceptions similar to the one proposed in the previous paragraph. The main difference is that the former model assumes that σ_x and the probability corresponding to ${}_dp_x^{\overline{S_iS_i}}$ are piece-wise constant, i.e. that these functions are constant over a range of values of x with a certain length, and, therefore, in this model the Poisson parameter does not have to be stated as an integral. In our case we do not need to make this assumption since our purpose is not to obtain a sequence of point estimates of each $\sigma(i)_x$.

Thus, assuming that, for a given sickness category i, variables $I_{i(x)}$ for different integer ages are independent and considering the 4 age groups defined in Section 2, the number of claim inceptions for sickness category i concerning sicknesses for which the policyholder is in age group j at the beginning of the sickness, which we denote by I_{ij} $(i = 1, ..., 18;\ j = 1, 2, 3, 4)$, has the following distribution:

$$I_{ij} = \sum_{x=a_j}^{b_j} I_{i(x)} \overset{a}{\sim} Po\left(\int_{a_j}^{b_j+1} E(y)\sigma(i)_y \, {}_dp_y^{\overline{S_iS_i}} r(i)_y dy\right) \tag{11}$$

where $[a_j, b_j + 1)$ is the age interval associated with age group j (recall from Section 2 that $a_1 = 18$ and $b_1 = 39$, $a_2 = 40$ and $b_2 = 49$, $a_3 = 50$ and $b_3 = 59$ and $a_4 = 60$ and $b_4 = 64$).

Note that distribution (11) can also be written as follows:

$$I_{ij} \overset{a}{\sim} Po\left(E_j \, \sigma(i)_{x_j} \, {}_dp_{x_j}^{\overline{S_iS_i}} r(i)_{x_j}\right) \tag{12}$$

where x_j is a certain age in the interval $[a_j, b_j + 1)$ which is given by the mean value theorem for integrals and where $E_j = \int_{a_j}^{b_j+1} E(y)dy$. We are going to estimate the $\sigma(i)_x$ for the different sickness categories using this equivalent model for I_{ij}.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Note that if $E(I_{ij})$ is large we can use the normal approximation to the Poisson distribution and assume that

$$I_{ij} \overset{\circ}{\sim} N\left(E_j \, \sigma(i)_{x_j} \, {}_d p_{x_j}^{\overline{S_i S_i}} r(i)_{x_j}, E_j \, \sigma(i)_{x_j} \, {}_d p_{x_j}^{\overline{S_i S_i}} r(i)_{x_j}\right) \tag{13}$$

As we will see in a later section, we need this assumption for the purpose of hypothesis testing. Investigations carried out with this kind of model have suggested that a value for $E(I_{ij})$ greater than 10 is large enough for this assumption to hold (see Schou and Vaeth (1980)).

5.2. Estimating the Sickness Intensities

Since the set of data we have available for estimating the $\sigma(i)_x$ is not as detailed as we would like (see Section 1), it is more appropriate to use the model (12) together with the assumption that, for a given deferred period, each $\sigma(i)_x$ is a function of σ_x.

Some preliminary investigations have indicated that in the cases of many sickness categories we should not assume that $\sigma(i)_x$ is a multiple of σ_x. Hence, based on the functional form used in CMIR 12 (1991, Part C) to obtain the graduation of σ_x, we make the following assumption for a given deferred period d and a given category i:

$$\sigma(i)_x = \exp\{\alpha_i + \beta_i x\} \sigma_x \quad i = 1, \dots, 18 \tag{14}$$

where α_i and β_i are unknown parameters which can vary according to the deferred period and the category being considered. Therefore, we have decided that all the approximations to the $\sigma(i)_x$ will have the functional form (14) with σ_x replaced by the corresponding graduation obtained in CMIR 12 (1991). This graduation is different according to the deferred period we consider. We consider possible the situation where $\beta_i = 0$, in which case the approximation to $\sigma(i)_x$ will be a multiple of the graduation of σ_x.

As we can see from the model (12), when an observation period is fixed, the data we need to estimate $\sigma(i)_x$ for a given deferred period d and a given sickness category i are the number of claim inceptions for this deferred period and category and for the 4 age groups we consider and also the exposure (i.e. the observed value of E_j) for this deferred period and the same 4 age groups.

As far as the claim inceptions are concerned and considering the observation period 1979-82, we are going to use the claim inceptions data described in Section 2, which are taken to be the observed values of the I_{ij}.

As far as the exposures are concerned the only data available are the exposures for single ages for the Standard Male Experience for 1979-82 (SME 79-82 for brevity). These exposures, which are available for each of the 4 deferred periods we consider, are also presented in Cordeiro (1998). In this set of data the exposure for age x is the total time spent in the period 1979-82 as healthy by policyholders aged x last birthday.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

For several reasons the exposures just mentioned are not appropriate for being used together with the claim inceptions data presented in Section 2. The main reason is the fact that the data presented in Section 2, which are classified by cause of disability, represent only part of the SME 79-82 to which the exposures refer. For the other reasons see Cordeiro (1998).

Since the number of claim inceptions by single ages concerning the SME 79-82 are available (they can also be found in Cordeiro (1998)), in order to overcome the problem mentioned in the previous paragraph, we have decided to obtain an approximation to the observed value of E_j for a given deferred period by assuming that the proportion of this value to the exposure for age group j concerning the SME 79-82 is the same as the proportion of the number of claim inceptions for age group j concerning the Cause of Disability Male Experience for 1979-82 to the number of claim inceptions for age group j concerning the SME 79-82.

As we have stated above, we are going to obtain the approximation to a given $\sigma(i)_x$ using the model (12) together with assumption (14). Thus, in this model we assume that

$$\log E(I_{ij}) = \log\left(E_j\, \sigma_{x_j}\, {}_d p_{x_j}^{\overline{S_i S_i}}\, r(i)_{x_j}\right) + \alpha_i + \beta_i x_j \tag{15}$$

From (15) we conclude that this model can be formulated as a generalized linear model (GLM) with a response variable I_{ij} which has a Poisson distribution, a log link function, a linear preditor $\eta_{ij} = \alpha_i + \beta_i x_j$ and an offset term $\left[\log\left(E_j\, \sigma_{x_j}\, {}_d p_{x_j}^{\overline{S_i S_i}}\, r(i)_{x_j}\right)\right]$. For an extended exposition of the GLMs theory see Dobson (1990) and McCullagh and Nelder (1989).

In practice we are going to estimate the parameters α_i and β_i in the GLM just presented by maximum likelihood using the statistical package GLIM (for the details about the estimation of GLMs using GLIM see Francis et al. (1993)).

In order to estimate the parameters α_i and β_i for a given deferred period d and a given sickness category i, we have to evaluate the functions σ_x, ${}_d p_x^{\overline{S_i S_i}}$, $r(i)_x$ and the preditor η_{ij} at some appropriate age in each of the intervals $[a_j, b_j + 1)$ associated with the 4 age groups we consider. We have decided that this age should be $(\bar{x}_j + 1/2)$ in the case of deferred period D1 and $(\bar{x}_j + 1/2 - d)$ in the cases of deferred periods D4, D13 and D26, where the age $\bar{x}_j$ is obtained as the following weighted average:

$$\bar{x}_j = \sum_{x=a_j}^{b_j} x \frac{E(x, x+1)}{E_j} \qquad j = 1, 2, 3, 4$$

where

$$E(x, x+1) = \int_x^{x+1} E(y)\,dy \qquad x = 18, \ldots, 64$$

is the exposure for (integer) age x. The reasons for this decision can be found in Cordeiro (1998).

In the estimation process for obtaining the approximations to the $\sigma(i)_x$ we are going to evaluate the functions ${}_dp_x^{\overline{S_iS_i}}$ and $r(i)_x$ using the approximations to the $\rho(i)_{x,z}$ and $\nu(i)_{x,z}$ obtained in Sections 3.3 and 4.

We are aware that the quality of the approximations to the $\sigma(i)_x$ we are going to obtain is probably not very good. However, we also believe that these approximations are the best that can be obtained with the data we have available.

The main reasons for not expecting to obtain approximations to the $\sigma(i)_x$ of good quality are the following: for a given deferred period d and a given category i, we are going to estimate 2 parameters (α_i and β_i) with only 4 observations; the data we have available is not enough to ensure that the estimated value of each $E(I_{ij})$ is greater than or equal to 10 and, therefore, there are combinations of deferred period, sickness category and age group for which the model (13) might not be valid. A fuller account of the limitations of the data available and of their possible consequences can be found in Cordeiro (1998).

5.3. Analysis of the Results

The purpose of this section is to present and analyse the results of the estimation process for obtaining the approximations to the $\sigma(i)_x$.

From the outputs of the GLIM programs we have run, we found out that, for a given deferred period d and a given category i, $\sum_{j=1}^{4} I_{ij} = \sum_{j=1}^{4} \widehat{E(I_{ij})}$, where $\widehat{E(I_{ij})}$ is the estimated value of $E(I_{ij})$. It can be easily shown that this is a mathematical consequence of having used the model (12) and the assumption (14).

The main purpose of the GLIM program we have run for each combination of deferred period d and sickness category i was to estimate the parameters α_i and β_i. From the output of this program we can compute the value

$$\chi_i^2 = \sum_{j=1}^{4} \frac{\left(I_{ij} - \widehat{E(I_{ij})}\right)^2}{\widehat{E(I_{ij})}} \qquad i=1,\ldots,18$$

which, taking into account assumption (13), can be regarded as the observed value of a chi-square goodness of fit statistic with a $\chi^2(2)$ distribution (a chi-square distribution with 2 degrees of freedom) and, therefore, we can carry out a goodness of fit test. The adequacy of the functional form (14) can be checked by comparing the p-value associated with the test with the significance level $\alpha = 0.05$.

As explained in Section 5.2, we consider the possibility of having $\beta_i = 0$ in (14) for some cases. We have decided to set $\beta_i = 0$ in (14) and use this new assumption to estimate again the $\sigma(i)_x$ in the cases where the estimate of β_i is not significantly different from zero. We consider that the estimate of a parameter β_i is not significantly different from zero when the absolute value of this estimate is less than twice its standard error.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

The results of all the GLIM programs we have run are summarized in Tables 5 to 8. Each of these tables shows the results for a given deferred period. Each table shows the following results for each category i: the estimates of α_i and β_i (or only the estimate of α_i, when we assume $\beta_i = 0$), the standard error of the estimate of β_i (or the standard error of the estimate of α_i, when we assume $\beta_i = 0$), the p-value associated with the corresponding goodness of fit test and the number of $\widehat{E(I_{ij})}$ greater than or equal to 10. When we assume $\beta_i \neq 0$, the table does not show the standard error of the estimate of α_i because the parameterisation used by GLIM makes this standard error irrelevant.

Analysing Tables 5 to 8, we can see that, for any of the 4 deferred periods we consider, there are categories for which the approximations to the $\sigma(i)_x$ are multiples of the graduations of the corresponding σ_x. In all there are 22 of these cases.

From Tables 5 to 8 we can also see that, for deferred periods D1, D4 and D13, there are categories for which the corresponding p-value is smaller than 0.05. The total number of these cases is 14 and the deferred period for which there are most cases is D1 (there are 7 cases for D1).

Fortunately, for 6 of these cases the p-value is not much smaller than 0.05 and, therefore, it is not unreasonable to still consider the functional form (14) as adequate in these cases. These cases are: for D1, categories 4, 10 and 11; for D4, category 4; and for D13, categories 5 and 8.

In the cases where the p-value is much smaller than 0.05 we know that the functional form (14) is not adequate with a high probability and, therefore, that we should use a new functional form to obtain the approximations to the corresponding $\sigma(i)_x$. However, the new functional form we would propose for these cases is similar to (14) but with a polynomial of a higher degree as the power of the exponential, which would imply the estimation of 3 or more parameters for each case. Under the circumstances, this is not advisable or even possible (see Section 5.2).

Considering the points in the previous paragraph, we have decided to propose as the approximations to the $\sigma(i)_x$ the functions whose estimated parameters are presented in Tables 5 to 8, although we are aware that for a small number of cases these approximations are not adequate. These cases are: for D1, categories 2, 7, 8 and 15; for D4, categories 8 and 15; and for D13, categories 12 and 14.

Finally, from Tables 5 to 8 we can also confirm the existence of cases where $\widehat{E(I_{ij})}$ is less than 10 (see Section 5.2). In fact, for any of the 4 deferred periods we consider, there are combinations of sickness category and age group for which $\widehat{E(I_{ij})}$ is less than 10. For deferred periods D13 and D26 there are even more cases where $\widehat{E(I_{ij})}$ is less than 10 than cases where the reverse happens.

Since the estimates in Tables 5 to 8 give only a very vague idea of the relative and absolute importance of each sickness category at each attained age as far as the sickness intensity is concerned, we have decided to present graphs of the approximations to the $\sigma(i)_x$ for the 18 categories we consider. Due to limitations of space we only present these graphs for deferred period D1.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

TABLE 5

RESULTS CONCERNING THE ESTIMATION OF THE $\sigma(i)_x$ FOR D1

Sickness Category	$\hat{\alpha}_i$	$\hat{\beta}_i$	**se** $\hat{\beta}_i$ **(se** $\hat{\alpha}_i$**)**	p**-value**	$\#\{\widehat{E(I_{ij})} \geq 10\}$
1. Other Infective	0.4062	-0.05716	0.00422	0.928	4
2. Malignant Neoplasms	-9.033	0.08126	0.0122	0.004	4
3. Benign Neoplasms	-4.729	–	0.1529	0.156	3
4. Endocrine and Metabolic	-7.593	0.03609	0.0184	0.041	1
5. Mental Illness	-3.963	–	0.0704	0.348	4
6. Nervous Disease	-5.549	0.03307	0.00849	0.324	4
7. Heart/Circulating System	-6.111	0.05402	0.00695	$\simeq 0$	4
8. Ischaemic Heart Disease	-9.067	0.1004	0.00932	$\simeq 0$	4
9. Cerebro Vascular Disease	-9.795	0.08003	0.023	0.068	1
10. Acute Respiratory	1.428	-0.05711	0.0033	0.038	4
11. Bronchitis Respiratory	-1.789	-0.0131	0.0048	0.035	4
12. Digestive	-3.457	0.01997	0.00445	0.117	4
13. Genito-Urinary	-4.716	0.03108	0.00689	0.5	4
14. Arthritis/Spondylitis	-8.504	0.08079	0.0112	0.203	4
15. Other Musculoskeletal	-2.492	–	0.0427	$\simeq 0$	4
16. R.T.A. Injuries	-4.135	–	0.0903	0.225	4
17. Other Injuries	-0.9051	-0.03509	0.0042	0.403	4
18. All Others	-1.957	-0.01662	0.00452	0.549	4

TABLE 6

RESULTS CONCERNING THE ESTIMATION OF THE $\sigma(i)_x$ FOR D4

Sickness Category	$\hat{\alpha}_i$	$\hat{\beta}_i$	**se** $\hat{\beta}_i$ **(se** $\hat{\alpha}_i$**)**	p**-value**	$\#\{\widehat{E(I_{ij})} \geq 10\}$
1. Other Infective	0.958	-0.07193	0.0133	0.652	3
2. Malignant Neoplasms	-9.632	0.08141	0.0164	0.175	2
3. Benign Neoplasms	-0.8802	-0.05857	0.0244	0.215	0
4. Endocrine and Metabolic	-10.106	0.07474	0.0365	0.035	0
5. Mental Illness	-4.306	–	0.0976	0.084	3
6. Nervous Disease	-3.887	–	0.1507	0.924	3
7. Heart/Circulating System	-3.794	–	0.1359	0.234	3
8. Ischaemic Heart Disease	-8.309	0.09	0.0106	0.003	4
9. Cerebro Vascular Disease	-9.947	0.09477	0.0286	0.6	0
10. Acute Respiratory	2.735	-0.08656	0.0185	0.543	1
11. Bronchitis Respiratory	-2.703	–	0.1856	0.527	1
12. Digestive	-1.723	–	0.0711	0.287	4
13. Genito-Urinary	-2.884	–	0.149	0.315	3
14. Arthritis/Spondylitis	-11.433	0.1239	0.0262	0.686	1
15. Other Musculoskeletal	-1.231	-0.02556	0.00824	$\simeq 0$	4
16. R.T.A. Injuries	-3.932	–	0.1374	0.174	3
17. Other Injuries	-0.1392	-0.04544	0.00801	0.387	3
18. All Others	-2.983	–	0.1072	3.802	3

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

TABLE 7

RESULTS CONCERNING THE ESTIMATION OF THE $\sigma(i)_x$ FOR D13

Sickness Category	$\hat{\alpha}_i$	$\hat{\beta}_i$	se $\hat{\beta}_i$ (se $\hat{\alpha}_i$)	p-value	$\#\{\widehat{E(I_{ij})} \geq 10\}$
1. Other Infective	2.78	–0.1163	0.0298	0.333	0
2. Malignant Neoplasms	–11.625	0.1249	0.0155	0.87	2
3. Benign Neoplasms	–2.946	–	0.3148	0.213	0
4. Endocrine and Metabolic	–6.648	–	0.3338	0.266	0
5. Mental Illness	–6.005	0.02827	0.0129	0.023	3
6. Nervous Disease	–3.247	–	0.1489	0.623	3
7. Heart/Circulating System	–6.054	0.05624	0.0164	0.649	2
8. Ischaemic Heart Disease	–9.426	0.1083	0.0124	0.035	3
9. Cerebro Vascular Disease	–9.165	0.08748	0.0269	0.917	0
10. Acute Respiratory	9.028	–0.2502	0.118	0.892	0
11. Bronchitis Respiratory	–1.745	–	0.2338	0.601	0
12. Digestive	–2.013	–	0.1326	0.002	3
13. Genito-Urinary	–3.179	–	0.3015	0.123	0
14. Arthritis/Spondylitis	–13.481	0.1631	0.0308	0.001	1
15. Other Musculoskeletal	0.07251	–0.05264	0.0122	0.27	3
16. R.T.A. Injuries	0.2045	–0.09588	0.0228	0.747	1
17. Other Injuries	0.2256	–0.05638	0.0134	0.079	3
18. All Others	–1.383	–0.03234	0.0157	0.377	3

TABLE 8

RESULTS CONCERNING THE ESTIMATION OF THE $\sigma(i)_x$ FOR D26

Sickness Category	$\hat{\alpha}_i$	$\hat{\beta}_i$	se $\hat{\beta}_i$ (se $\hat{\alpha}_i$)	p-value	$\#\{\widehat{E(I_{ij})} \geq 10\}$
1. Other Infective	5.842	–0.1953	0.0678	0.505	0
2. Malignant Neoplasms	–9.832	0.08095	0.022	0.828	0
3. Benign Neoplasms	3.358	–0.1297	0.05	0.253	0
4. Endocrine and Metabolic	–13.752	0.1355	0.0648	0.837	0
5. Mental Illness	–6.445	0.03715	0.0148	0.586	3
6. Nervous Disease	–2.721	–	0.1493	0.905	2
7. Heart/Circulating System	–7.626	0.07889	0.0281	0.449	1
8. Ischaemic Heart Disease	–11.537	0.1393	0.0195	0.44	2
9. Cerebro Vascular Disease	–8.648	0.08906	0.0248	0.476	2
10. Acute Respiratory	6.108	–0.1309	0.0651	0.153	0
11. Bronchitis Respiratory	–1.959	–	0.3782	0.491	0
12. Digestive	0.1318	–0.05621	0.0271	0.259	0
13. Genito-Urinary	5.127	–0.1903	0.0703	0.337	0
14. Arthritis/Spondylitis	–8.602	0.07581	0.0243	0.432	1
15. Other Musculoskeletal	–3.323	–	0.2426	0.6	0
16. R.T.A. Injuries	–0.9203	–0.06889	0.0274	0.439	0
17. Other Injuries	–0.5076	–0.04782	0.0226	0.981	0
18. All Others	–2.976	–	0.2038	0.365	1

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

We present the graphs in 4 different figures: Figures 2 to 5. The scales used in Figures 2 to 4 are more or less similar whereas the scale used in Figure 5 is completely different from the others.

As we can see from Figures 2 to 5, the approximations to the $\sigma(i)_x$ have various shapes. The approximations whose graphs are shown in Figures 2 and 3 (i.e. the approximations for categories 2, 4, 6, 7, 8, 9, 13 and 14) and the approximation for category 12 (whose graph is shown in Figure 4) are clearly increasing functions of x. The approximations for categories 1, 10, 11, 17 and 18, whose graphs are shown in Figure 5, are decreasing functions of the attained age. Finally, the approximations for categories 3, 5, 15 and 16, whose graphs are shown in Figure 4, have the same (more or less 'flat') shape as the graduation of σ_x for D1 (see Figure C1 in CMIR 12 (1991)). Note that, for most of the sickness categories, the shapes of the corresponding approximations to the $\sigma(i)_x$ are those we would expect.

As far as the levels of the approximations to the $\sigma(i)_x$ are concerned we can see that, apart from the approximations for categories 1, 10 and 17, all the approximations take on values less than 0.05. Note that we would expect the approximations for categories 1 and 10 (Other Infective and Acute Respiratory, respectively) to be among those which take on the highest values. On the other hand, we can see that the approximations for categories 2 and 9 (Malignant Neoplasms and Cerebro Vascular Disease, respectively) are among those which take on the smallest values (see Figure 2). We would also expect this to happen.

6. Final Considerations

In CMIR 12 (1991) the transition intensity μ_x has not been estimated because the data required to do so were not available. The reason for this is that the CMIB has no direct information about the mortality rates experienced by policyholders who are not making a claim.

Exactly for the same reason, we also do not estimate the mortality of the healthy intensity defined for our model. For obvious reasons, for a given deferred period d, we can propose as the approximation to this transition intensity the graduation of μ_x proposed in CMIR 12 (1991): the graduation of the force of mortality for the Male Permanent Assurances 1979-82, duration 0. We should note that this graduation is the same for the 4 deferred periods we consider.

Now, that we have proposed approximations to all the transition intensities, we have made our model fully operational. In fact, using these approximations, the formulae for basic probabilities and the numerical algorithms for the evaluation of some of the basic probabilities (derived in Cordeiro (1998, 2002)), we can calculate any quantities relevant to the study of PHI claims by cause of disability.

Tables showing the average duration of a claim and claim inception rates for the different sickness categories can help the underwriters whenever they have to make underwriting decisions concerning proposals for new entries. With the information in these tables the underwriters can make decisions which are much more well grounded.

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

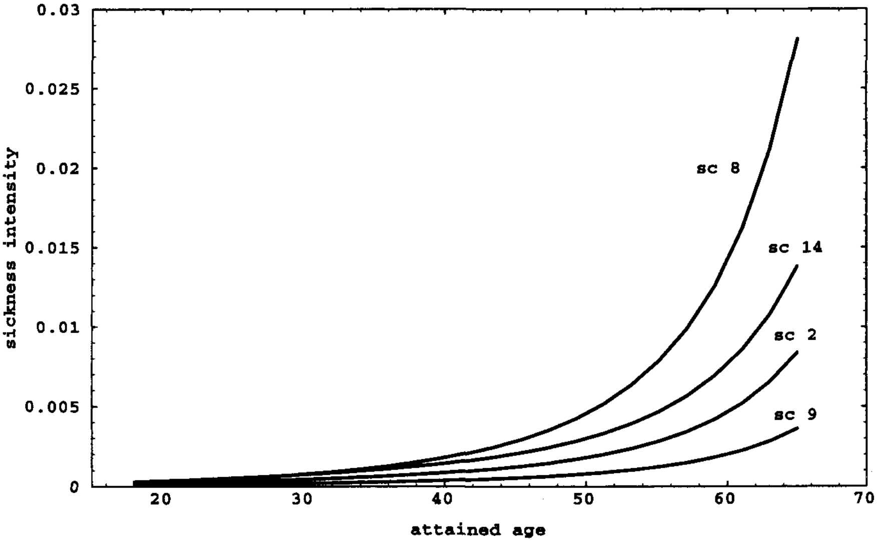

FIGURE 2: Approximations to the sickness intensities $\sigma(i)_x$ for D1 and sickness categories 2, 8, 9 and 14.

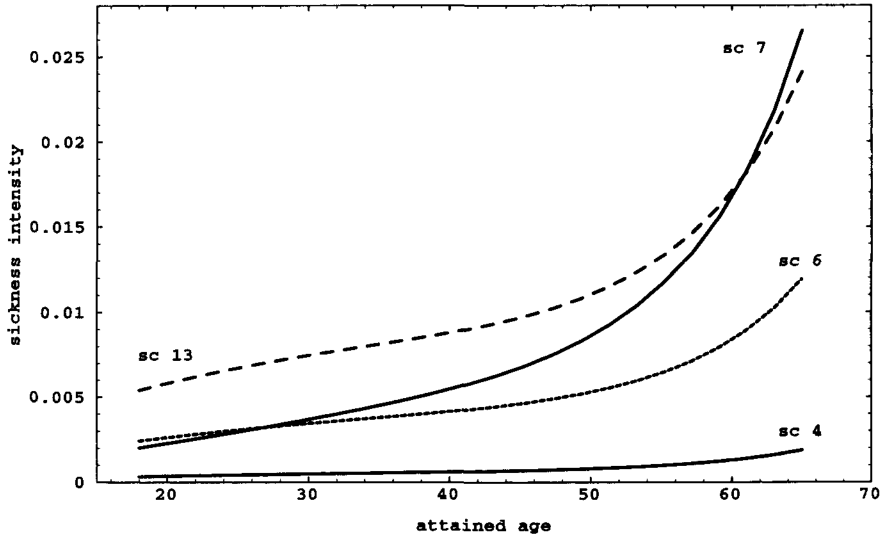

FIGURE 3: Approximations to the sickness intensities $\sigma(i)_x$ for D1 and sickness categories 4, 6, 7 and 13.

On the other hand, tables showing the average duration of a claim for the different sickness categories, for the different deferred periods and for different ages at the beginning of sickness can help the people responsible for the

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

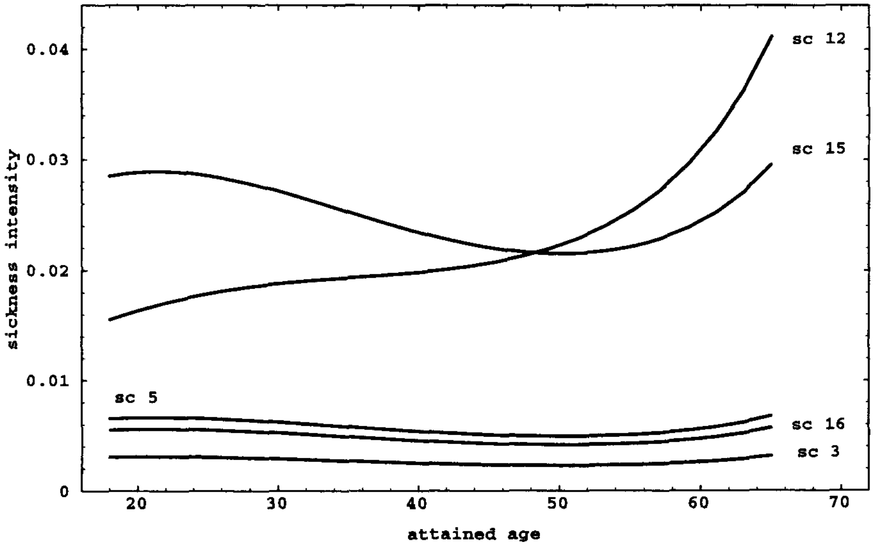

FIGURE 4: Approximations to the sickness intensities $\sigma(i)_x$ for D1 and sickness categories 3, 5, 12, 15 and 16.

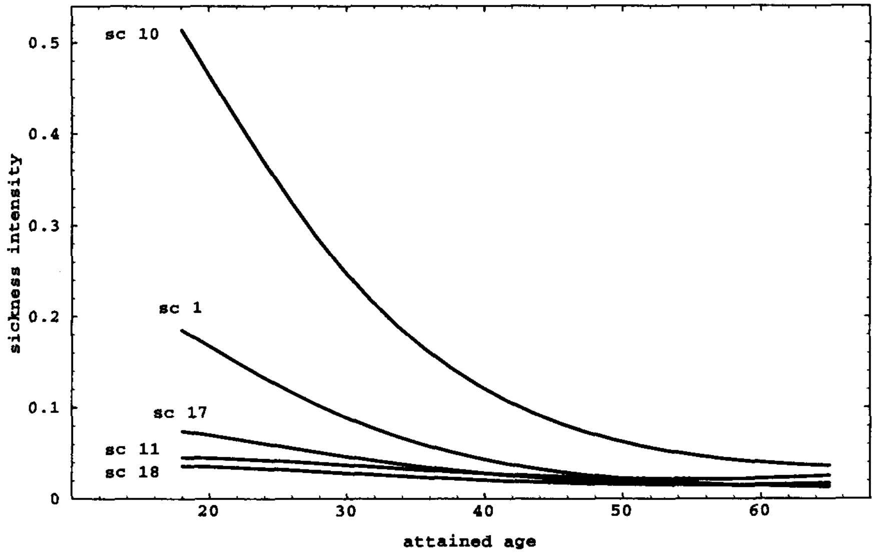

FIGURE 5: Approximations to the sickness intensities $\sigma(i)_x$ for D1 and sickness categories 1, 10, 11, 17 and 18.

claims control process in keeping a tighter control over the claims which are being paid and, therefore, in reducing claim recovery time.

Examples of the tables mentioned above can be found in Cordeiro (2002).

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press

Acknowledgements

I am grateful to the CMIB which has made available the data used in this work. I would also like to thank Professor Howard Waters and two anonymous referees for their invaluable comments and suggestions.

References

Continuous Mortality Investigation Committee (1986) Cause of Disability Experience Individual PHI Policies 1975-78. *Continuous Mortality Investigation Reports*, **8**, pp. 65-88. The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1991) The Analysis of Permanent Health Insurance Data. *Continuous Mortality Investigation Reports*, **12**. The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1993) Calculation of Continuation Tables and Allowance for Non-Recorded Claims Based on the PHI Experience 1975-78. *Continuous Mortality Investigation Reports*, **13**, pp. 123-130. The Institute of Actuaries and the Faculty of Actuaries.

Cordeiro, I.M.F. (1998) *A Stochastic Model for the Analysis of Permanent Health Insurance Claims by Cause of Disability*. Ph.D. Thesis. Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, U.K.

Cordeiro, I.M.F. (2002) A Multiple State Model for the Analysis of Permanent Health Insurance Claims by Cause of Disability. *Insurance: Mathematics & Economics*, forthcoming.

Dobson, A.J. (1990) *An Introduction to Generalized Linear Models*. Chapman & Hall, London.

Francis, B., Green, M. and Payne, C. (1993) *The GLIM System, Release 4 Manual*. Clarendon Press, Oxford.

Hoem, J.M. (1987) Statistical Analysis of a Multiplicative Model and its Application to the Standardization of Vital Rates: a Review. *International Statistical Review*, **55/2,** pp. 119-152.

Macdonald, A.S. (1996) An Actuarial Survey of Statistical Models for Decrement and Transition Data, I: Multiple State, Binomial and Poisson Models. *British Actuarial Journal*, **2**, pp. 129-155.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. Chapman & Hall, London.

Schou, G. and Vaeth, M. (1980) A Small Sample Study of Ocurrence/Exposure Rates for Rare Events. *Scandinavian Actuarial Journal*, 1980, pp. 209-225.

Sverdrup, E. (1965) Estimates and Test Procedures in Connection with Stochastic Models for Deaths, Recoveries and Transfers Between Different States of Health. *Scandinavian Actuarial Journal*, 1965, pp. 184-211.

Isabel Maria Ferraz Cordeiro.
Escola de Economia e Gestão
Universidade do Minho
Campus Universitário de Gualtar
4710-057 Braga
Portugal
Tel: ++351-253-604546
Fax: ++351-253-676375
E-mail: icordeiro@eeg.uminho.pt

https://doi.org/10.2143/AST.32.2.1033 Published online by Cambridge University Press