CONTINUOUS MORTALITY INVESTIGATION
INSTITUTE OF ACTUARIES • FACULTY OF ACTUARIES

ISSN 2044-3145

# Continuous Mortality Investigation

# Income Protection Committee

# WORKING PAPER 46

# Background papers on the analysis of CMI Individual Income Protection Claim records

This paper forms part of a series of papers which are summarised in CMI Working Paper 48: "*An overview of the Graduation of Sickness Inception and Termination Rates for the CMI Individual Income Protection Experience for 1991-98 of Males, Occupation Class 1*."

It is recommended that Working Paper 48 is read before this paper.

July 2010

The Institute of Actuaries, the Faculty of Actuaries, the CMI and its Committee Members do not accept or assume any responsibility for the use of this document by any party in any context. This document does not provide any form of guidance and should not be relied on as such.

## Executive Summary

This Working Paper contains four short papers related to the analysis and graduation of the CMI Individual Income Protection (IP) experience. All four deal exclusively with the file of Claim records, covering some features of the data and changes to the initial processing of Claims data (Parts A and B) and reporting the results of some additional background investigations (Parts C and D).

**Part A: Note on Exclusions and some other features of the Claims data**
During the investigation and graduation of the Claim Inceptions experience for 1991-98 (published in CMI Working Paper 47), some features of the Claims data, and its initial processing, were reconsidered. Part A provides documentation and commentary on the features, and on the changes to the algorithm applied to identify and exclude unacceptable records, giving also some numerical indication of the effects.

**Part B: The Identification of Duplicates**
There is a high prevalence of 'Duplicate' records within the IP data submitted to the CMI. Duplicates typically occur when a policyholder buys additional cover of the same 'type' so that there are multiple records relating to the same underlying exposure or Claim. Overall there are 38 Duplicate records for every 100 Claim records submitted. It is important to identify and remove Duplicate records within the data as their inclusion would undermine the statistical model and may introduce bias to the graduations by affecting the weight given to observations in each data cell. Part B sets out an evaluation of alternative algorithms for identifying Duplicate Claim records.

**Part C: The Experience of Singletons and Duplicates**
Parts C and D present the results of analyses into the experience and distribution of Duplicates in the Claims data. Part C describes an investigation to compare the Claim Terminations experience of Claims with Duplicates against that of Claims that have no Duplicates ('Singletons'). The analysis shows no statistically significant differences in Claim Terminations experience between the two groups of records. We are unable to compare the Claim Inceptions experiences as we cannot currently separate the In force data for the two groups of records.

**Part D: An Analysis of the Distribution of Duplicates**
Part D reports the results of a range of investigations into the distribution of Duplicates within Claim records. The prevalence of Duplicates was found to vary by Occupation Class, Deferred Period, Age and Sex, all factors for which we would naturally subdivide the data anyway, but not for any other factors available in the data (after aggregation across Offices).

The results of the analyses in Parts C and D justify the practice of the CMI IP Committee in analysing policies which are Singletons or Duplicates together, but it is still better to exclude (extra) Duplicates from the experience analysis where possible.

CONTENTS

# Part A: Note on Exclusions and some other features of the Claims data
(the 'Exclusions Note' )

## A1 INTRODUCTION

A1.1 Several features of the processing of the CMI Income Protection (IP) Investigation have been introduced over the years, and several changes have been made to these during the investigation and graduation of the Claim Inceptions experience for 1991-98 (published in CMI Working Paper 47). We document and comment on the features and the changes in this paper, giving also some numerical indication of the effects. The basis of almost all our numbers is the file of Claims, including Duplicates, for 1991-2002, which initially contained 206,500 Claim records. (Note that this is an extended period, compared with the period used for the graduations themselves.)

## A2 FAULTY DATES

A2.1 In the first editing of the Claims file two types of error were discovered. First, although the given Dates of Sickness, Commencement of Claim and Cessation of Claim have always been checked, the checks done in the past ensured only that the day was between 1 and 31 and the month was between 1 and 12. The combination was not checked, so, for example, "31 April" was allowed as a valid date. Dates are now checked fully, but existing invalid combination dates like "31 April" are carried forward into the next month and are treated as 1 May or the equivalent, with the Claim thereafter being treated as valid. This correction affects very few cases, only five in the whole of the 1991-2002 Claims file.

A2.2 Next, the Dates of Sickness, Commencement of Claim and Cessation of Claim, are checked to ensure that they are in the correct order:

$$\text{Sickness} \leq \text{Commencement} \leq \text{Cessation.}$$

If they are not in the right order, the case is treated as an error and is processed no further. The number of such cases identified in the 1991-2002 Claims file is 37, consisting of:

| | |
|---|---|
| Commencement Date < Sick Date | 10 |
| Cessation Date < Sick Date | 3 |
| Cessation Date < Commencement Date | 24 |

In some cases two or all three of these inequalities apply, but a faulty case is identified only under one of them. After deducting these 37, the edited file consists of 206,463 Claim records.

## A3 Standard, Star and Aggregate

A3.1 In the early years of the IP Investigations, a distinction was made between “Standard” cases and the “Aggregate” data. After the coding of Occupation Class started in 1991 the Standard section was extended to a “Standard Star” (or “Standard*”) category. For the following analysis we classify cases into one of three groups, which we name “Standard”, “Star” and “Agg”. Standard Star consists of “Standard” plus “Star”, and “Agg” cases are those in the Aggregate, but excluded from Standard Star.

A3.2 “Agg” cases are defined as: those with the Location not the United Kingdom, and therefore coded as Republic of Ireland, Isle of Man or Channel Islands; also any cases with Impairment Code anything other than None (code blank or zero) or Don’t Know (code 7); also any cases with Benefit Type Code anything other than Level, Increasing or Decreasing (codes 1, 2, 3), thus including in Agg cases of Waiver (code 4), Lump Sum (code 5), or Other (code 9).

A3.3 “Star” cases are defined as: those that are not “Agg” and have Occupation Class other than Class 1 or “Class 5” (i.e. not given), thus taking in Classes 2, 3 and 4; also any cases, whatever the Occupation Class, that have the Occupation Rating coded as either Rated, or More Rated (codes 1 or 2). The cases we describe as “More Rated” are probably miscoded, but there are some hundreds of such cases in the earlier years, though none since 1994. Rather than exclude them we treat them as the same as “Rated” cases. The recording of a fuller Occupation Class began in 1990. Since then most cases with Occupation Class 2, 3 or 4, but not all, have been coded as “Rated”, whereas most cases with Occupation Class 1, but not quite all, have been coded as “Not Rated”, and those with Occupation Class “5” (“Not given”) are split. We assume that prior to 1990 those cases that would have been coded with Occupation Code 2, 3 or 4 were recorded as “Rated”, so the best way to achieve consistency throughout is to do as we have done.

A3.4 This leaves in the “Standard” category only those with Occupation Class 1 or “5”, which are coded as “Not Rated” and are not “Agg”.

A3.5 The numbers of cases in different categories are shown in Tables A1 and A2. There is considerable duplication, because a case may become “Agg” or “Star” on more than one criterion.

Table A1: Location, Impairment Code and Benefit Type

| Location | | Impairment Code | | Benefit Type | |
|---|---|---|---|---|---|
| UK | 205,646 | None | 151,087 | Level | 116,486 |
| Ireland | 741 | Don’t know | 41,069 | Increasing | 86,523 |
| Isle of Man | 29 | Hypertension | 193 | Decreasing | 3,426 |
| Channel Islands | 47 | Neurosis | 2,265 | Waiver | 19 |
| | | All other | 11,849 | Other | 9 |
| Total | 206,463 | Total | 206,463 | Total | 206,463 |

Table A2: Occupation Class and Occupation Rating

| Occupation Class | Not Rated | Rated | More Rated | Total |
|---|---|---|---|---|
| Class 1 | 141,110 | 1,034 | 2 | 142,146 |
| Class 2 | 5,711 | 14,087 | 0 | 19,798 |
| Class 3 | 5,260 | 15,539 | 0 | 20,799 |
| Class 4 | 1,779 | 11,812 | 0 | 13,591 |
| Class 5 (not given) | 8,006 | 2,121 | 2 | 10,129 |
| Total | 161,866 | 44,593 | 4 | 206,463 |

A3.6 In Table A3 we give a summary that shows that in the "Agg" cases there are 14,307 with excluded impairment codes, which account for most of the 15,104 "Agg" cases. Most of the remaining "Agg" cases must be "Not UK", but some of these may be excluded also for one or more other reasons. In the "Star" column the Occupation Classes do not overlap, so can be added and the total of 57,347 is the maximum number that could be "Star"; but some of these are already classified as "Agg" cases. In the "Standard" column we can add the Class 1 "Not Rated" and Class 5 "Not Rated", so the number of "Standard" cases can be no larger than the sum of 149,116; but a number of these are already classified as "Agg", leaving 136,942 "Standard" cases.

Table A3: "Agg", "Star" and "Standard"

| "Agg" | | "Star" | | "Standard" | |
|---|---|---|---|---|---|
| Not UK | 817 | Class 2 | 19,798 | Class 1 Not Rated | 141,110 |
| Impairment excluded | 14,307 | Class 3 | 20,799 | Class 5 Not Rated | 8,006 |
| Benefit type excluded | 28 | Class 4 | 13,591 | | |
| | | Class 5 Rated | 2,123 | | |
| | | Class 1 Rated | 1,036 | | |
| Sub-total | 15,152 | Sub-total | 57,347 | Sub-total | 149,116 |
| minus overlap | –48 | minus "Agg" cases | –2,930 | minus "Agg" cases | –12,174 |
| Total "Agg" | 15,104 | Total "Star" | 54,417 | Total "Standard" | 136,942 |

## A4 DEFERRED PERIODS

A4.1 In the input data the Deferred Period is given in integral weeks, from 0 to 52, except that cases with a "one month" Deferred Period are coded "999". The most common Deferred Periods are 1, 4, 13, 26 and 52 weeks, and in the past these have been set up as five "Grouped DP" categories, and each case with a Deferred Period that is not one of these five has been allocated to the next higher Grouped DP category, with "one month" cases treated as DP4.

A4.2 For the analysis of recovery and mortality rates, and for analysis on a Manchester Unity basis, this grouping has been satisfactory. However, for the Inceptions analysis it is less convenient, for a number of reasons. The numbers of Claims for the other Deferred Periods are as shown in Table A4. We can see that there are relatively large numbers for

Deferred Periods 0, 2 and 8 weeks, so corresponding DP categories have been created, giving now eight DP categories: DP0, DP1, DP2, DP4, DP8, DP13, DP26 and DP52. The remaining Deferred Periods all have quite few cases and it is possible that the periods with only one or two cases are miscoded, though the slightly larger numbers for Deferred Periods 16, 20, 34, 39 and 40 weeks suggest that they may be genuine.

Table A4: Numbers of Claim records for different Deferred Periods.

| Deferred Period (weeks) | Claim records |
|---|---|
| 1 | 62,617 |
| 4 | 42,319 |
| 1 month (code "999") | 936 |
| 13 | 42,073 |
| 26 | 39,024 |
| 52 | 17,546 |
| Sub-total: Common DPs | 204,515 |
| 0 | 529 |
| 2 | 162 |
| 8 | 1,118 |
| Sub-total: Less common DPs | 1,809 |
| 3 | 1 |
| 5 | 1 |
| 6 | 2 |
| 10 | 1 |
| 14 | 6 |
| 16 | 17 |
| 17 | 1 |
| 18 | 6 |
| 20 | 21 |
| 21 | 4 |
| 28 | 3 |
| 30 | 3 |
| 32 | 1 |
| 34 | 12 |
| 35 | 3 |
| 39 | 30 |
| 40 | 25 |
| 43 | 2 |
| Sub-total: Uncommon DPs (or "Odd DPs") | 139 |
| Total | 206,463 |

A4.3 Further inspection of the behaviour of the DP0, DP2 and DP8 Claims suggests that DP2 and DP8 are exactly as one would expect, since almost all Claims commence on the 15th or the 57th day respectively, counting the Date of Sickness as the first day. Most DP0 Claims commence on the Date of Sickness, like DP1 Claims, but, also like DP1, there are fairly few Terminations in the first week of each Claim, so it seemed possible that cases coded DP0 had in fact the same policy conditions as DP1 cases. However, further investigation (with the offices writing DP0 and DP1 business) showed that DP0 cases are policies with a genuine nil Deferred Period, and Claims for very few days are accepted. However, there are relatively few very short Claims, which suggests that there is a short "run-in" period for this Deferred Period, so that those who are Sick for only a few days do not bother to claim. The same investigation also showed that DP1 cases strictly have a six-day Deferred Period, not a seven-day one. That is, if the assured is Sick for exactly seven days, he or she may claim for all seven; but if the assured is Sick for only six days he or she may not claim at all. The recovery rates for these durations of Sickness were found to be consistent with this.

A4.4 The cases with a Deferred Period not exactly one of the eight classes, of which there are 139, are defined as having "Odd DPs", and are omitted from all the analyses relating to Inceptions, and will be omitted in future from the Terminations analysis.

## A5 CODING OF OCCUPATION

A5.1 All the In force cases are coded with an Occupation Class, which may be from 1 to 4, or may be not given. We denote cases where the Class is not given as "Class 5". All the Claim records are similarly coded.

A5.2 In most cases, contributing offices are able to code In force and Claim records in the same way. However, certain offices are able to give the Class only for Claims, but not for the corresponding In force. We therefore define a second Occupation Class for the Claims, "In force Class", as follows:

- for those offices that do provide Occupation Class for both In force and Claim records, In force Class is set equal to the Class given for the Claim record;
- but for those offices that cannot give the Occupation Class for the In force records, In force Class is set to Class 5 (matching the classification of the corresponding In force).

A5.3 A further number of offices are able to give records for Claims but not for the In force at all. In such cases the Claim records are coded with "In force Class" 6, and they are omitted from the Inceptions analysis but not from the Terminations analysis. The numbers of cases of each category in the 1991-2002 Claims file are shown in Table A5.

Table A5: Number of Claims in different Occupation Classes

| | Total in Claims file | Moved to Class 5 | Moved to Class 6 | Net after moves |
|---|---|---|---|---|
| Class 1 | 142,146 | 11,372 | 185 | 130,589 |
| Class 2 | 19,798 | 6,167 | 129 | 13,502 |
| Class 3 | 20,799 | 8,843 | 70 | 11,886 |
| Class 4 | 13,591 | 4,417 | 48 | 9,126 |
| Class 5 (Not given) | 10,129 | 0 | 1,589 | 39,339 |
| Class 6 (no matching In force) | 0 | 0 | 0 | 2,021 |
| Total | 206,463 | 30,799 | 2,021 | 206,463 |

## A6 False One-day Claims

A6.1 It had previously been noted (for example, in CMI Working Paper 6, Section 1.5) that for a number of Claims the Commencement Date and the Cessation Date were the same, and that the implied recovery rates on that day seemed unnaturally high. It had been suggested that, for a case with a longer Deferred Period, the insured might notify a Claim within that period, but recover before the end of it. The office, however, might record the Claim as a pending one, but dispose of it by recording it as if the Claim had started and finished on the same day. These were therefore identified as "False One-day Claims" and have been eliminated from some of the Terminations analyses. They are now also omitted as either Inceptions or Claims in the Inceptions analysis.

A6.2 The numbers of False One-day Claims for different Deferred Periods are shown in Table A6. There are hardly any for short Deferred Periods, but much larger numbers in the longer Deferred Periods, from DP4 upwards.

Table A6. Numbers of False One-day Claims by Deferred Period

| Deferred Period (weeks) | Claim records |
|---|---|
| 0 | 3 |
| 1 | 1 |
| 2 | 3 |
| 4 and "999" | 221 |
| 8 | 1 |
| 13 | 399 |
| 26 | 158 |
| 52 | 118 |
| Total | 904 |

## A7 Early Terminations

A7.1 For a long time the CMI IP programmes have identified Claims where the dates of Commencement and of Cessation were inconsistent with the given Deferred Period. To test this, the item Days Deferred is defined. This is now taken in general as seven times the Deferred Period in weeks, so that for example for DP0 it is zero days, for DP4 it is 28 days, and for DP52 it is 364 days. However, for DP1 it is taken as six days, and for cases with Deferred Period coded as "999" (one month) it is taken as 28 days. Previously the Days Deferred was taken as seven times the Grouped DP, with that for DP52 being taken as 365 days. Days Deferred is used in a number of ways as described below.

A7.2 We define also some other terms. A "New Claim" is a Claim that has a Mode of Commencement defined either as "new Claim" or as "new Claim after interruption". There are very few of the latter, only 17 in the 1991-2002 Claims file. In principle these are Inceptions, but an Inception is defined more exactly below. A "Continuation Out" is a case where the Mode of Cessation is coded as a continuation, meaning that the Claim was still In force at the end of the Investigation Year. For some purposes, like the calculation of exposure, such a case is treated as if it "exited" on 31 December of the Year. For other purposes it is treated differently. A "Continuation In" is a case where the Mode of

Commencement is coded as a continuation, meaning that the Claim was In force prior to the start of the Investigation Year. The exposure of such a case for that Year is assumed to commence on 1 January

A7.3 The first test we describe is that each Claim that is not a Continuation Out is tested to see whether the Sickness Date plus the Days Deferred is greater than the given Cessation Date. If it is, that indicates that the Claim both started and finished within the Deferred Period, which seems inconsistent. We define such a case as an "Early Termination".

A7.4 However, if the Mode of Cessation is "Benefit Change" the case is not treated as an Early Termination. If the amount of benefit is altered there should always be two matching Claim records, one coded with Mode of Cessation as Benefit Change ("Out"), the other with Mode of Commencement as Benefit Change (In). In fact the numbers of such cases are not equal as one might expect them to be, but we do not know the reason for that. But such a change may happen at any time during the Claim, and it would not be correct to assume that the Claim has in fact terminated. So a case that is a Benefit Change Out is not treated as an Early Termination.

A7.5 An Early Termination may have any Mode of Commencement, as a New Claim in that Year, a Continuation In, a Revival, or any other mode. However, cases coded as New Claims that are also Early Terminations are not treated as Inceptions in the analysis of Inceptions.

A7.6 When Days Deferred was based on the Grouped DP, rather too many cases were identified as Early Terminations, because, for example, many DP8 Claims, which were included with DP13, had started and finished before 13 weeks had expired, so were treated unnecessarily as Early Terminations. The numbers were not large, so the distortion was small. But, as already noted, it is better to keep these Deferred Periods separate in the Inceptions analysis.

A7.7 Many DP1 cases start on the Sickness Date (as day 1) and cease six days later (on day 7). When the Days Deferred was taken as seven days, these were treated as Early Terminations. Now that Days Deferred is taken as six days for DP1, these are normal cases. This affects almost 4,000 cases in all, a considerable number.

A7.8 Investigation of DP52 Claims showed that most Claims commenced on day 365, counting the Sickness Date as day one so the period was in practice really 52 weeks, not one year, as had previously been assumed. The effect of this was to treat cases that Commenced and Ceased on the 365th day as Early Terminations, rather than as False One day Claims, as they are now treated.

A7.9 Early Terminations are not treated as Inceptions in the Inceptions analysis, but are still treated as current Claims and their period of Claim is deducted from the exposure. But in the Terminations analysis they are excluded, in principle because the duration of Sickness seems unreliable, but in practice also because they have terminated before the end of the Deferred Period, so do not enter the analysis at all.

## A8 CARRIED FORWARDS AND BROUGHT FORWARDS

A8.1 Another test carried out for a long time is that, for any case coded as an Inception that is also a Continuation Out, if the Sickness Date plus the Days Deferred exceeds 31 December of the Investigation Year, the Claim is not treated as an Inception in that Year, but it is assumed that it will become an Inception in the following Year. These are defined as “Carried Forwards”. Correspondingly, for any case that is a Continuation In, if the Sickness Date plus the Days Deferred is on or after 1 January of the Investigation Year, it is assumed that this is an Inception “Brought Forward” from the previous Year. A Brought Forward is then counted as an Inception in that Investigation Year, unless it is also an Early Termination, as quite a number are.

A8.2 The test for a Carried Forward is just the same as for an Early Termination. In each case the test is whether the Cessation Date is less than the Sickness Date plus the Days Deferred, where the Cessation Date for a Continuation Out is taken as 31 December. But if the case is a Continuation Out it is defined as a Carried Forward, and otherwise it is defined as an Early Termination.

## A9 PREMATURE REVIVALS AND BENEFIT CHANGES

A9.1 A further inconsistency sometimes appears in cases with the Mode of Commencement coded as a Revival. If an individual is Sick and had made a valid Claim and then recovered, he or she may become Sick again, perhaps from the same illness (but we do not know that), and the Claim may be continued without waiting for a new Deferred Period to expire. Such a case is described as a Revival and should be so coded in the Claim record for the revived Claim. The Date of Sickness given should be the date the original Sickness commenced, so that the analysis of Terminations can take the correct duration of Sickness into account. However, it is observed that for many of these cases the given Date of Sickness seems to be the date when the Sickness recurred. We can see this because for many cases the Date of Commencement is before the Date of Sickness plus the Days Deferred, and quite often equal to the Date of Sickness. We define such cases as Premature Revivals. It may also be the case that the Date of Cessation is before the Date of Sickness plus the Days Deferred, so that the case is also an Early Termination.

A9.2 The same inconsistency appears in quite a number of the cases that are coded as Benefit Change Ins. The given Date of Sickness ought to be the original date of Sickness, the same as given for the matching Benefit Change Out record. But in quite a number of cases the given Date of Sickness is the same as the Date of Commencement, and we apply the same test as for Premature Revivals, identifying them as Premature Benefit Changes. Some are also Early Terminations.

A9.3 Premature Revivals and Premature Benefit Changes are of course not counted as Inceptions, but they are included in the Inceptions analysis, because, like Early Terminations, they are still treated as current Claims and their period of Claim is deducted from the exposure. However, in the Terminations analysis, also like Early Terminations, they are excluded because the duration of Sickness seems unreliable

A9.4 The numbers of Early Terminations, Carried Forwards, Brought Forwards, Premature Revivals and Premature Benefit Changes of different types are shown in Table A7.

Table A7: Numbers of Early Terminations, Carried Forwards, Brought Forwards, Premature Revivals and Premature Benefit Changes

| | Early Terminations | Carried Forwards | Premature cases not ETs or CFs | Others | Total |
|---|---|---|---|---|---|
| New Claims | 538 | 789 | | 76,038 | 77,365 |
| Other "Ins" | | | | 126,964 | 126,964 |
| Brought Forwards | 46 | | | 1,047 | 1,093 |
| Premature Revivals | 67 | 35 | 467 | | 569 |
| Premature Benefit Changes | 9 | 29 | 434 | | 472 |
| Total | 660 | 853 | 901 | 204,049 | 206,463 |

## A10 DEFINITION OF INCEPTIONS

A10.1 An Inception, for the purpose of the Inceptions analysis, is defined as a Claim which is either a New Claim or a Brought Forward, but in either case is not an Early Termination nor a Carried Forward nor a False One-day Claim. Revivals and Benefit Changes, whether premature or not, are not counted as Inceptions. From Table A7 we can see that the number of Inceptions is 76,038 New Claims plus 1,047 Brought Forwards, which gives 77,085 in all, excluding Early Terminations and Carried Forwards in both cases. There are 820 False One Day Claims among these Inceptions, giving a net total of 76,265 Inceptions; Table A6 showed 904 False One Day Claims in all but some are also Early Terminations, and others are not Inceptions. When the Inceptions analysis is done, the number of Inceptions may be reduced further by the exclusion of "Agg" cases.

## A11 MODE OF CESSATION

A11.1 Previously, cases coded with a Mode of Cessation as Lump Sum or Ex Gratia payment were excluded from all the analyses. After further consideration, it seemed inappropriate to do this. While the Termination is certainly neither a recovery nor a death, such cases are "exposed to the risk" of becoming a recovery or a death, even though they may be settled in some other way. Further, a case may have been coded as a Continuation Out in a previous Year, but end up as being settled in one of these unusual ways. It would not have been excluded in its earlier Years of Claim, so it seems inappropriate to exclude it in its final, or only, Year. Such cases are now included throughout, but their exit is not treated as a recovery or a death. In the 1991-2002 file there are 415 cases of cessation by Lump Sum payment and 167 by Ex Gratia payments.

## A12 INCEPTIONS ANALYSIS

A12.1 We can now define what is done in the Inceptions analysis underlying the graduation of Sickness rates for 1991-98 (published in CMI Working Paper 47). Only "Standard" and "Star" cases are included, "Agg" cases being omitted. Cases that are Odd DPs are also omitted, with DP0, DP2 and DP8 being analysed separately. These two rules apply to both the In force and the Claims. Claims that now have In force Class 6, and also

False One-day Claims cases are also omitted. All other cases are included. Inceptions are as defined in Section A10. Early Terminations, Carried Forwards, Premature Revivals and Premature Benefit Changes are included, but only to the extent that the number of days of Claim is deducted from the exposure. Although the given dates may be inconsistent, we assume that they are nevertheless Claims of some sort, and that during this Claim period they are not exposed to the risk of becoming Claims again, so should be deducted from the gross exposure.

## A13 TERMINATIONS ANALYSIS

A13.1 In the Terminations analysis, only "Standard" and "Star" cases are included, "Agg" cases being omitted. There is no need to consider the In force Occupation Class, and the Claim Occupation Class is used throughout. For the latest Terminations analyses (but not for results published before this paper), cases that are Odd DPs and False One-day Claims cases are also omitted, with DP0, DP2 and DP8 being analysed separately; Early Terminations, Carried Forwards, Premature Revivals and Premature Benefit Changes are excluded entirely, but Brought Forwards that are not Early Terminations are included.

A13.2 The total numbers on the 1991-2002 Claims file that are excluded by these various criteria are shown in Table A8. The numbers are incremental, because a case that has been excluded for an early reason in the list is not counted again if it would have been excluded for a later reason.

Table A8: Exclusions for Inceptions analysis and for Terminations analysis.

| | Inceptions analysis | Terminations analysis |
|---|---|---|
| Starting total | 206,463 | 206,463 |
| deductions: | | |
| "Agg" | 15,104 | 15,104 |
| In force Occupation = Class 6 | 1,798 | |
| Odd DPs | 137 | 137 |
| False One Day Claims | 896 | 898 |
| Early Terminations and CFs | | 1,315 |
| Premature Revivals and BCs | | 789 |
| Net total | 188,528 | 188,220 |

# Part B: The Identification of Duplicates
(the 'Duplicates Note')

## B1 INTRODUCTION

B1.1 In every actuarial demographic investigation it is highly desirable that duplicate cases among the "events" that are being counted should be eliminated. Without this, any estimates of the significance of any results are likely to be faulty. There are several reasons for this.

B1.2 First, if duplicates in any investigation are included, and it were to be assumed (unrealistically) that each individual insured had the same number of duplicate policies, then, although the ratio of actual events (deaths in a mortality investigation; Inceptions, recoveries or deaths in an Income Protection (IP) investigation, etc) to expected events in a particular "cell" (combination of Age, Sex, etc) would not be affected, the standard error of this ratio would be larger than if each insured had one policy.

B1.3 Secondly, since it is much more likely that different insureds have different numbers of policies, there may be a concentration of them in a particular cell which would result either in an unusually large number of actual events being recorded if the event happened to the insured; or, alternatively, to the exposure being increased and the observed rate of occurrence of the event being lower, if the event did not happen to the insured. Thus the observed ratio of actual to expected events has a larger variation than if each assured had only one policy recorded and may be biased either upwards or downwards, even if the expected value is unaffected.

B1.4 Thirdly, even if the expected ratio in any cell is unaffected by even a heterogeneous number of duplicates, the weight of it in any aggregate calculation of actual to expected may be unduly large.

B1.5 Finally, in any graduation of the data to provide smoothed rates, a cell with an unusual ratio resulting from an undue concentration of duplicates may have an undue weight in the calculations.

B1.6 For all these reasons it is desirable to eliminate duplicates if possible. In the IP experience, for which individual policy data is submitted to the CMI, duplicate policies appear to exist, but the limited form and depth of the available data hamper attempts to identify them: among the Claims it is possible to identify Duplicates with moderate confidence, but it is not currently possible to identify them among the In force.

B1.7 Duplicate policies in IP may arise in at least two ways:

(a) One is that a policyholder, having effected one policy for a certain amount, decides later that he or she would like to be insured for a larger amount, and effects a subsequent policy on the same terms. This might well involve additional underwriting so as to demonstrate that the policyholder is still insurable and, probably, is not currently Claiming.

(b) A second way is when an office issues a policy which is automatically incremented each year, either by a percentage amount or in accordance with some published index. Many offices would write this as a single policy, and such policies are recorded in the CMI IP files; but other offices may, or may for a time, record such a policy as if it were a series of independent policies, similar to the first type. But there would normally be no additional underwriting.

B1.8 We do not know which of these ways is the more likely in the CMI IP data, although both do appear to be present. However, the presence or absence of additional underwriting for the duplicate policies makes no difference to the current CMI IP investigations, but if we were to investigate the experience by policy duration it would be important.

B1.9 Duplicate policies have indeed been eliminated from the Claim records for IP since the investigations reported in "CMIR 12: *The Analysis of Permanent Health Insurance Data*" (1991) which re-examined the data back to 1975 using a particular set of criteria that is described below. However, during the investigation into the Claim Inceptions experience for 1991 to 1998 (see CMI Working Paper 47) an anomaly appeared. Two files of Claims, one with Duplicates included ("cumD") and one with Duplicates excluded ("exD") were used, in order to estimate a proportion of Duplicates within each "cell" (combination of Year, Sex, Age Definition, Age, Deferred Period, and Occupation). This ratio could then be applied to the In force, which necessarily included Duplicates, so as to get an estimate of the numbers of In force excluding Duplicates. However, this resulted in certain cells in the data having a non-zero number of Inceptions in the cumD file and a zero number in the exD file. This seemed inconsistent and indicated that an investigation into the process whereby Duplicates are identified was worth further consideration.

B1.10 This paper therefore describes our investigation into different criteria for identifying Duplicates, and describes what we intend to adopt as a new criterion.

## B2 CRITERIA

B2.1 The form and depth of the IP data collected by the CMI (at least up to data for calendar year 2006*) have significant limitations. In particular the data does not contain personal (policy or insured life) identifiers, and only the month and year of birth is recorded, not the full date. Thus we cannot identify individuals, and we have to use some set of matching characteristics to identify plausible Duplicates.

*[* A revised Coding Guide was published in July 2009 which, when adopted by data contributors, will lead to a much richer data supporting, inter alia, a precise identification of Duplicate policies.]*

B2.2 The CMI's present method (that is, the method in use immediately prior to this investigation) is based on a method devised at an earlier stage of the IP investigation. In the report in CMIR 12, Part B, in ¶1.2 the method of identifying Duplicate Claims was described. This included all the items in the list below except (8) "Date of Cessation of Claim", but tests indicate that this item was also used then, and it is used in the current CMI method. The method involves sorting the data for Claims into sequence by the following fields:

1 Year of Investigation
2 Sex
3 Age Definition
4 Year of Birth
5 Month of Birth
6 Deferred Period (exact weeks)
7 Date of Sickness (day, month and year)
8 Date of Cessation of Claim (day, month and year)
and then by the sequence number in the given input file.

B2.3 If two successive Claims had identical values in the first eight fields, then the Claim with the lowest sequence number was kept and others discarded. However, this meant that Claims with different Occupation Class codes or different cause codes were not treated as separate Claims, and all except one might be discarded.

B2.4 Although the In force files presumably contain Duplicate policies, there is too little information to allow Duplicates to be excluded. It is not until we identify a particular Sickness that we can attempt to exclude Duplicates.

B2.5 We assume first that the relevant section of the Claims data has been selected, in this case the Individual IP data for the years 1991 to 2002. In addition to the first eight fields noted above, the remaining relevant data that is available in the subset of the original record that is used for most of the analysis is the following:

9 "Standard Status" (described below)
10 Deferred Period (DP grouped as DP 0, 1, 2, 4, 8, 13, 26 or 52)
11 Occupation (as coded for the Inceptions analysis)
12 Occupation (as coded for the Terminations analysis)
13 Mode of Cessation of Claim
14 Date of Commencement of Claim
15 Mode of Commencement of Claim
16 Cause of Sickness
17 Year of Entry (we do not have precise dates of entry)

B2.6 By "Standard Status" we indicate that cases, both In force and Claims, are put into one of three categories, which we call "Standard", "Star" and "Agg". "Standard" is the same as what was originally called the Standard experience and is still defined the same way, excluding in particular (and *inter alia*) cases with an Occupational rating, or cases with Occupation Codes 2, 3 or 4. "Star" cases are those with Occupational ratings or Occupation Codes 2, 3 or 4, but not otherwise excluded. When these are added to Standard we have what is now called the "Standard*" or "Standard Star" experience. The "Agg" cases are those excluded from the Standard Star experience, but included in the Aggregate, though not including those cases that are excluded entirely from all the investigations as being probable errors. "Agg" cases therefore include any non-UK cases, and any cases with Impairment Codes not 0 ("none") or 7 ("don't know"), or with Benefit Type Codes 4 ("waiver"), 5 ("lump sum") or 9 ("other").

B2.7 One possibility is to tailor the duplicates routine to the particular application for which it is intended to use the exD file produced. Thus, unless we were interested in the Cause of Sickness, we would not use this as a discriminating feature, and would accept

Claims with different Causes of Sickness, and otherwise matching, as Duplicates. However, if we were wishing to analyse the data by Cause of Sickness we would use this to distinguish cases as not being Duplicates. This would mean having different Duplicate routines for different purposes.

B2.8 Another possibility is to use a strict test for matching Duplicates and use the same corresponding exD file for all applications. This might, in some cases, mean accepting Claims as not being Duplicates, when, for all the factors in which we are interested, they are identical.

B2.9 The choice between these two approaches may depend on the numbers of cases involved. If the numbers of Claims affected by two different routines for identifying Duplicates is quite small, it may not matter much which is used. Therefore it is worth investigating the actual numbers.

B2.10 In order to be consistent in the method of selecting one case out of a group of presumed Duplicates we also sort the Claims file into order, using for each method all the other available fields in a prescribed order. If two cases match even then they are sufficiently identical for it to make no difference to our current investigations which one we choose.

B2.11 We have taken the cum Duplicates file for Individual IP Claims from 1991 to 2002 (12 years), including the Aggregate experience, (but omitting some cases that we presume have errors in the coding, as described in the "Exclusions Note" (Part A). The number of valid Claim records in this cumD file is 206,463. We then use 16 different methods for identifying Duplicates, from the broadest (giving the most duplicates and hence the fewest cases in the exD file) to the tightest (giving the most cases in the exD file).

B2.12 The fields (numbered as shown above) that are used in the different methods are:

Method 1:

1 Year of Investigation<br>
2 Sex<br>
3 Age Definition<br>
4 Year of Birth<br>
5 Month of Birth<br>
7 Date of Sickness (day, month and year)<br>
10 Deferred Period (DP grouped)

Method 2:

as Method 1 but using in addition:<br>
6 Deferred Period (exact weeks)

Method 3:

as Method 2 but using in addition:<br>
9 Standard Status

Method 4a:

as Method 3 but using in addition:<br>
11 Occupation (as coded for the Inceptions analysis)

Method 4b:

as Method 3 but using in addition:
12 Occupation (as coded for the Terminations analysis)

Method 5:

as Method 3 but using in addition:
11 Occupation (as coded for the Inceptions analysis)
12 Occupation (as coded for the Terminations analysis)

Method 6a:

as Method 5 but using in addition:
8 Date of Cessation of Claim

Method 7a:

as Method 6a but using in addition:
13 Mode of Cessation of Claim

Method 8a:

as Method 7a but using in addition:
14 Date of Commencement of Claim

Method 6b:

as Method 5 but using in addition:
14 Date of Commencement of Claim

Method 7b:

as Method 6b but using in addition:
15 Mode of Commencement of Claim

Method 8b:

as Method 7b but using in addition:
8 Date of Cessation of Claim

Method 9:

as Method 8a but using in addition:
15 Mode of Commencement of Claim
the same as Method 8b but using in addition:
13 Mode of Cessation of Claim

Method 10:

as Method 9 but using in addition:
16 Cause of Sickness

Method 11:

as Method 10 but using in addition:
17 Year of Entry

Method 12: (this is an artificial method, since it excludes no Duplicates, but it sorts the cases into the given sequence, and it identifies the number of cases in the original file)
as Method 11 but using in addition:
18 Sequence number

Method 13:
CMI's present method = Method 2 but using in addition
8 Date of Cessation of Claim

The combinations of fields used are also shown, for each method, in the block table below:

Table B1: Summary of fields used in each Duplicate Method
[shaded cell = field used; unshaded cell = field not used]

| Criteria (Fields) | Method for Identifying Duplicates | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4a | 4b | 5 | 6a | 7a | 8a | 6b | 7b | 8b | 9 | 10 | 11 | 12 | 13 |
| Year of Investigation | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Sex | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Age Definition | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Year of Birth | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Month of Birth | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| DP (exact weeks) | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Date of Sickness | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Date of Cessation | | | | | | | ■ | ■ | ■ | | | ■ | ■ | ■ | ■ | ■ | ■ |
| Standard Status | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| DP (grouped) | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| Occ (Inceptions) | | | | ■ | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| Occ (Terminations) | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| Mode of Cessation | | | | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | |
| Date of Commence | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| Mode of Commence | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | |
| Cause of Sickness | | | | | | | | | | | | | | ■ | ■ | ■ | |
| Year of Entry | | | | | | | | | | | | | | | ■ | ■ | |
| Sequence number | | | | | | | | | | | | | | | | ■ | |

B2.13 We note that the source data gives the exact Deferred Period (DP) of the policy in weeks, but for many purposes this is rounded (grouped) to the next higher DP, where DPs are now restricted to 0, 1, 2, 4, 8, 13, 26 and 52 weeks, and were previously restricted to 1, 4, 13, 26 and 52 weeks. Using Grouped DP allows certain cases to appear to be Duplicates which have different numbers of exact weeks Deferred Period. Using exact DP keeps such cases separate.

B2.14 We also note that for the analysis of Terminations experience we use only the Claims files, so the Occupation Code given on the Claims files record is suitable. However some offices cannot give the coded occupation in the In force records, and for these we need to shift the Claim records to the more general Occupation Class "5" (meaning no Occupation Code given) or "6" (meaning no In force data given; we do include these last cases in the Terminations analysis although they are not used for the Inception analysis). We therefore investigate, in Methods 4a, 4b and 5, the effect of using one definition, then the other, each singly, and then both.

B2.15 We also, in Methods 6a, 7a, 8a and 6b, 7b, 8b in parallel, followed by 9, investigate the effect of bringing in Dates and Modes of Cessation and of Commencement of Claim in either order.

B2.16 In Table B2 we show the numerical results. We see that, except for Methods 11 and 12, all Methods give the percentage of non-Duplicates between 68.6% and 72.7%, of the numbers in the cumD Claims file. The exceptions are Method 11, where we use Year of Entry to discriminate, and Method 12, where we remove no Duplicates. Year of Entry is not an item that it is useful to use, because one of the ways in which an individual may accumulate multiple policies is by taking out incremental policies in successive years. It is therefore not surprising that if we treat policies in different Years as different, we eliminate most of the apparent Duplicates.

B2.17 The present CMI method (Method 13) shows intermediate results. Adding Standard Status shows the biggest jump in non-Duplicates (excluding Year of Entry); adding either Date of Commencement or Date of Cessation of Claim is next; adding the other adds only a little; the present CMI method does use Date of Cessation of Claim.

Table B2: Numbers of non-Duplicates (in exD File) and of Duplicates and incremental numbers, for each Duplicate Method.

| Method | Items included | Number in exD file | Incremental number | Number of Duplicates omitted | Percentage retained |
|---|---|---|---|---|---|
| 1 | First 7 items | 141,675 | | 64,788 | 68.6 |
| 2 | 1 + Deferred period (exact) | 141,697 | 22 | 64,766 | 68.6 |
| 3 | 2 + Standard status | 146,344 | 4,647 | 60,119 | 70.9 |
| 4a | 3 + Occupation for Inceptions | 146,813 | 469 | 59,650 | 71.1 |
| 4b | 3 + Occupation for Terminations | 146,895 | 551 | 59,568 | 71.1 |
| 5 | 3 + Both Occupations | 147,104 | 291 or 209 | 59,359 | 71.2 |
| 6a | 5 + Date of Cessation | 149,722 | 2,618 | 56,741 | 72.5 |
| 7a | 6a + Mode of Cessation | 149,736 | 14 | 56,727 | 72.5 |
| 8a | 7a + Date of Commencement | 149,885 | 149 | 56,578 | 72.6 |
| 6b | 5 + Date of Commencement | 149,504 | 2,400 | 56,959 | 72.4 |
| 7b | 6b + Mode of Commencement | 149,519 | 15 | 56,944 | 72.4 |
| 8b | 7b + Date of Cessation | 149,876 | 357 | 56,587 | 72.6 |
| 9 | 8a + Mode of Commencement<br>= 8b + Mode of Cessation | 149,888 | 3 or 12 | 56,575 | 72.6 |
| 10 | 9 + Cause of Sickness | 150,026 | 138 | 56,437 | 72.7 |
| 11 | 10 + Year of Entry | 199,058 | 49,032 | 7,405 | 96.4 |
| 12 | No Duplicates | 206,463 | 7,405 | 0 | 100.0 |
| 13 | CMI present method<br>= 2 + Date of Cessation | 144,492 | | 61,971 | 70.0 |

## B3 CONCLUSION

B3.1 Occupation, defined in one or other way, or both, is essential if we are to avoid the problem, noted in B1.9, of there being no apparent exD cases in a cell where there are cumD ones. So going up to at least Method 6a or 6b seems very desirable. If we were to move to the tightest reasonable definition of Duplicates, Method 10, which includes Cause of Sickness, we would add 5,534 Claim records to the exD file in comparison with the present CMI method, but fewer than about 500 in comparison with either of methods 6a or 6b. Having a single definition for all purposes is a convenience, so we intend to use Method 10 for all future CMI work, unless a new investigation using different data requires a reconsideration of this. We have also therefore used Method 10 throughout the investigation and graduation of the Claim Inceptions experience for 1991-98 (published in CMI Working Paper 47).

# Part C: The Experience of Singletons and Duplicates

## C1 INTRODUCTION

C1.1 The CMI Income Protection (IP) Committee has revised the method used for identifying Duplicate policies in the individual IP investigations. The revised method is described in the paper "*The Identification of Duplicates*" (Part B). Duplicate policies occur when it appears that one policyholder has a number of separate Claim records with sufficiently similar conditions for it to be better to treat them as one Claim rather than as several. An analysis was carried out using the file of Claims for 1991 to 2002, with, in the first place, all Claims included. Then the Claim records were sorted so that records with a matching set of characteristics, described in Part B, were identified, and a specified one of these was selected to represent the bundle of assumed Duplicates.

C1.2 If it were the case that the experience of those with Duplicate policies was significantly different from those with only one policy, it would be desirable to consider their experiences separately. We do not know the Duplicates in the In force files, so we cannot investigate whether the Inception experiences differ. However, for the Claim Terminations we do know how many Duplicates there are, so we can investigate their experiences, and indeed can do so in relation to recoveries and deaths separately.

## C2 INVESTIGATION

C2.1 We divide the Claims into three categories:

(a) those with only one policy, which we call "Singletons";

(b) the first policy of a bundle of Duplicates (which is retained in the ex Duplicates section), which we call "First Duplicates";

(c) the subsequent policies of a bundle of Duplicates, which we call "Extra Duplicates".

C2.2 The "exD" Claims file excludes Extra Duplicates and so consists of Singletons and First Duplicates; First Duplicates and Extra Duplicates together form "All Duplicates"; and all three categories together form the "cumD" Claims file.

C2.3 For any bundle of assumed Duplicates the Claim Terminations experience is the same, because they all have the same Sex, Deferred Period, Occupation Class and Age, and the same dates of Sickness Commencement and Cessation. However, if the experience differed according to how many Duplicate policies there were, the experience of Extra Duplicates could be different from that of First Duplicates, and if the experience of All Duplicates differed from that of Singletons, this would also be apparent.

C2.4 In Tables C1, C3, C5 and C7 we show the numbers of Terminations of Claims in 1991-2002, for Males Recoveries, Females Recoveries, Males Deaths and Females Deaths respectively, subdivided by Occupation Class and Deferred Period (DP), and by Duplicates category. We omit DPs 0, 2 and 8 and all Odd DPs. In Tables C2, C4 and C6 we show the ratio 100 × $A/E$ (using IPM 1991-98) for corresponding categories, but only where the number of Singletons and the number of All Duplicates are both at least 100. If the number of events is smaller than this the confidence intervals for the ratio are so large that results would be misleading. This results in many sections being omitted, including Females Deaths entirely (so there is no Table C8).

C2.5 We can see that much the largest number of events with Duplicates is in the category Males, Recoveries, Occupation Class 1, DP1, where there are 2,256 recoveries for Singletons, 3,391 for First Duplicates and 7,461 for Extra Duplicates. The overall experience ratios, 100 × $A/E$, for these three categories are respectively 102, 102 and 101, so there is no evidence of significant differences between them. Even if the numbers of cases were larger, so the ratios were statistically significantly different, such small differences are not important.

C2.6 DP4 in the same category (Males, Recoveries, Occupation Class 1) also has quite a large number of recoveries, and the ratios are 97, 100 and 96, also not significantly different. DP13 has fewer cases and the ratios differ more, being 117, 103 and 106 respectively; but these are still not significantly different.

C2.7 DP26 (Males, Recoveries, Occupation Class 1) has fewer recoveries still, and there are only 124 in the All Duplicates category, with fewer than 100 in both First and Extra Duplicates, with 322 Singletons. The 100 × $A/E$ ratios of Singletons and All Duplicates are 114 and 93, further apart than for the other DPs, and the Expected numbers are 282.2 and 133.2. The standard deviations of the 100 × $A/E$ ratio, $r$, may be estimated as $\sqrt{(100.r)} / \sqrt{E}$ (equivalently 100 × $\sqrt{A} / E$). So, for this DP26 example they are respectively 6.4 and 8.4. A little more calculation shows that the possible ranges for the ratios overlap considerably, so the ratios are not significantly different. The numbers of recoveries for DP52 among All Duplicates are only 46, so no results are shown.

C2.8 For Males Recoveries we find large enough numbers in Occupation Classes 2, 3 and 4 only for DP4. For Occupation Class 2 the 100 × $A/E$ ratio for Singletons is 85, with a 95% range of 79 to 91; for All Duplicates it is 127 with a 95% range of 108 to 146. For Occupation Class 3 the ratios are 96 and 108, with smaller numbers so wider ranges, and for Occupation Class 4 they are 104 and 87. So sometimes the All Duplicates category has a higher ratio than Singletons, sometimes lower.

C2.9 Inspection of the results for Females Recoveries and for Males Deaths shows likewise that the 100 × $A/E$ ratios for Singletons and Duplicates are seldom far apart and, because of the relatively small numbers of events, we cannot say that Duplicates have any clearly different experience from Singletons.

## C3 CONCLUSION

C3.1 This investigation of the Terminations experience of Claims, classified as Singletons, First Duplicates and Extra Duplicates, has shown that:

(a) where the numbers of cases were large, the differences in the experience of recoveries and deaths were negligible; and

(b) although, where the numbers were smaller, there were noticeable differences between the experiences, these were nowhere statistically significant.

C3.2 This result supports the CMI practice of analysing the three Claim categories - Singletons, First Duplicates and Extra Duplicates - together.

Table C1: Numbers of Recoveries, Males, 1991-2002,
subdivided by Occupation Class and Duplicates Category.

| | DP 1 | DP 4 | DP 13 | DP 26 | DP 52 | All DPs |
|---|---|---|---|---|---|---|
| Occupation Class 1 | | | | | | |
| Singletons | 2,256 | 1,675 | 661 | 322 | 137 | 5,132 |
| First Duplicates | 3,391 | 476 | 102 | 46 | 19 | 4,044 |
| Extra Duplicates | 7,461 | 755 | 179 | 78 | 27 | 8,510 |
| ex D | 5,647 | 2,151 | 763 | 368 | 156 | 9,176 |
| All Duplicates | 10,852 | 1,231 | 281 | 124 | 46 | 12,554 |
| cum D | 13,108 | 2,906 | 942 | 446 | 183 | 17,686 |
| Occupation Class 2 | | | | | | |
| Singletons | 23 | 734 | 363 | 88 | 30 | 1,256 |
| First Duplicates | 0 | 63 | 24 | 4 | 0 | 92 |
| Extra Duplicates | 0 | 108 | 28 | 7 | 0 | 144 |
| ex D | 23 | 797 | 387 | 92 | 30 | 1,348 |
| All Duplicates | 0 | 171 | 52 | 11 | 0 | 236 |
| cum D | 23 | 905 | 415 | 99 | 30 | 1,492 |
| Occupation Class 3 | | | | | | |
| Singletons | 3 | 1,714 | 351 | 78 | 29 | 2,253 |
| First Duplicates | 0 | 99 | 11 | 8 | 2 | 120 |
| Extra Duplicates | 0 | 114 | 11 | 12 | 3 | 140 |
| ex D | 3 | 1,813 | 362 | 86 | 31 | 2,373 |
| All Duplicates | 0 | 213 | 22 | 20 | 5 | 260 |
| cum D | 3 | 1,927 | 373 | 98 | 34 | 2,513 |
| Occupation Class 4 | | | | | | |
| Singletons | 3 | 1,109 | 418 | 73 | 15 | 1,648 |
| First Duplicates | 0 | 57 | 17 | 6 | 1 | 81 |
| Extra Duplicates | 0 | 65 | 23 | 6 | 1 | 95 |
| ex D | 3 | 1,166 | 435 | 79 | 16 | 1,729 |
| All Duplicates | 0 | 122 | 40 | 12 | 2 | 176 |
| cum D | 3 | 1,231 | 458 | 85 | 17 | 1,824 |
| Occupation Class 5 | | | | | | |
| Singletons | 2 | 269 | 281 | 87 | 36 | 724 |
| First Duplicates | 0 | 7 | 13 | 8 | 0 | 28 |
| Extra Duplicates | 0 | 7 | 15 | 10 | 0 | 32 |
| ex D | 2 | 276 | 294 | 95 | 36 | 752 |
| All Duplicates | 0 | 14 | 28 | 18 | 0 | 60 |
| cum D | 2 | 283 | 309 | 105 | 36 | 784 |
| All Occupation Classes | | | | | | |
| Singletons | 2,287 | 5,501 | 2,074 | 648 | 247 | 11,013 |
| First Duplicates | 3,391 | 702 | 167 | 72 | 22 | 4,365 |
| Extra Duplicates | 7,461 | 1,049 | 256 | 113 | 31 | 8,921 |
| ex D | 5,678 | 6,203 | 2,241 | 720 | 269 | 15,378 |
| All Duplicates | 10,852 | 1,751 | 423 | 185 | 53 | 13,286 |
| cum D | 13,139 | 7,252 | 2,497 | 833 | 300 | 24,299 |

Table C2: Ratios 100 × *A*/*E* for Recoveries, Males, 1991-2002, subdivided by Occupation Class and Duplicates Category. *E* is calculated using IPM 1991-98.

| | DP 1 | DP 4 | DP 13 | DP 26 | DP 52 | All DPs |
|---|---|---|---|---|---|---|
| Occupation Class 1 | | | | | | |
| Singletons | 102 | 97 | 117 | 114 | .. | 101 |
| First Duplicates | 102 | 100 | 103 | .. | .. | 101 |
| Extra Duplicates | 101 | 96 | 106 | .. | .. | 100 |
| ex D | 102 | 98 | 115 | 111 | .. | 101 |
| All Duplicates | 101 | 98 | 105 | 93 | .. | 100 |
| cum D | 101 | 97 | 113 | 107 | .. | 101 |
| Occupation Class 2 | | | | | | |
| Singletons | .. | 85 | .. | .. | .. | 94 |
| First Duplicates | .. | .. | .. | .. | .. | .. |
| Extra Duplicates | .. | 129 | .. | .. | .. | 121 |
| ex D | .. | 87 | .. | .. | .. | 95 |
| All Duplicates | .. | 127 | .. | .. | .. | 121 |
| cum D | .. | 91 | .. | .. | .. | 97 |
| Occupation Class 3 | | | | | | |
| Singletons | .. | 96 | .. | .. | .. | 96 |
| First Duplicates | .. | .. | .. | .. | .. | 101 |
| Extra Duplicates | .. | 105 | .. | .. | .. | 97 |
| ex D | .. | 97 | .. | .. | .. | 97 |
| All Duplicates | .. | 108 | .. | .. | .. | 99 |
| cum D | .. | 97 | .. | .. | .. | 97 |
| Occupation Class 4 | | | | | | |
| Singletons | .. | 104 | .. | .. | .. | 109 |
| First Duplicates | .. | .. | .. | .. | .. | .. |
| Extra Duplicates | .. | .. | .. | .. | .. | .. |
| ex D | .. | 103 | .. | .. | .. | 109 |
| All Duplicates | .. | 87 | .. | .. | .. | 109 |
| cum D | .. | 102 | .. | .. | .. | 109 |
| All Occupation Classes | | | | | | |
| Singletons | 101 | 97 | 125 | 115 | 93 | 102 |
| First Duplicates | 102 | 103 | 120 | .. | .. | 102 |
| Extra Duplicates | 101 | 99 | 116 | 104 | .. | 100 |
| ex D | 102 | 98 | 124 | 114 | 88 | 102 |
| All Duplicates | 101 | 100 | 118 | 105 | .. | 101 |
| cum D | 101 | 98 | 124 | 113 | 82 | 101 |

Table C3: Numbers of Recoveries, Females, 1991-2002, subdivided by Occupation Class and Duplicates Category.

| | DP 1 | DP 4 | DP 13 | DP 26 | DP 52 | All DPs |
|---|---|---|---|---|---|---|
| Occupation Class 1 | | | | | | |
| Singletons | 692 | 734 | 307 | 182 | 91 | 2,019 |
| First Duplicates | 381 | 88 | 29 | 15 | 5 | 518 |
| Extra Duplicates | 566 | 115 | 36 | 18 | 4 | 739 |
| ex D | 1,073 | 822 | 336 | 197 | 96 | 2,537 |
| All Duplicates | 947 | 203 | 65 | 33 | 9 | 1,257 |
| cum D | 1,639 | 937 | 372 | 215 | 100 | 3,276 |
| Occupation Class 2 | | | | | | |
| Singletons | 5 | 258 | 115 | 45 | 38 | 467 |
| First Duplicates | 0 | 8 | 3 | 0 | 1 | 12 |
| Extra Duplicates | 0 | 9 | 6 | 0 | 1 | 16 |
| ex D | 5 | 266 | 118 | 45 | 39 | 479 |
| All Duplicates | 0 | 17 | 9 | 0 | 2 | 28 |
| cum D | 5 | 275 | 124 | 45 | 40 | 495 |
| Occupation Class 3 | | | | | | |
| Singletons | 0 | 55 | 20 | 22 | 9 | 107 |
| First Duplicates | 0 | 7 | 1 | 3 | 2 | 13 |
| Extra Duplicates | 0 | 8 | 1 | 3 | 2 | 14 |
| ex D | 0 | 62 | 21 | 25 | 11 | 120 |
| All Duplicates | 0 | 15 | 2 | 6 | 4 | 27 |
| cum D | 0 | 70 | 22 | 28 | 13 | 134 |
| Occupation Class 4 | | | | | | |
| Singletons | 0 | 2 | 11 | 6 | 3 | 22 |
| First Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| Extra Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| ex D | 0 | 2 | 11 | 6 | 3 | 22 |
| All Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| cum D | 0 | 2 | 11 | 6 | 3 | 22 |
| Occupation Class 5 | | | | | | |
| Singletons | 2 | 18 | 40 | 32 | 10 | 117 |
| First Duplicates | 0 | 1 | 3 | 0 | 0 | 4 |
| Extra Duplicates | 0 | 1 | 4 | 0 | 0 | 5 |
| ex D | 2 | 19 | 43 | 32 | 10 | 121 |
| All Duplicates | 0 | 2 | 7 | 0 | 0 | 9 |
| cum D | 2 | 20 | 47 | 32 | 10 | 126 |
| All Occupation Classes | | | | | | |
| Singletons | 699 | 1,067 | 493 | 287 | 151 | 2,732 |
| First Duplicates | 381 | 104 | 36 | 18 | 8 | 547 |
| Extra Duplicates | 566 | 133 | 47 | 21 | 7 | 774 |
| ex D | 1,080 | 1,171 | 529 | 305 | 159 | 3,279 |
| All Duplicates | 947 | 237 | 83 | 39 | 15 | 1,321 |
| cum D | 1,646 | 1,304 | 576 | 326 | 166 | 4,053 |

Table C4: Ratios 100 × *A*/*E* for Recoveries, Females, 1991-2002, subdivided by Occupation Class and Duplicates Category. *E* is calculated using IPM 1991-98.

| | DP 1 | DP 4 | All DPs |
|---|---|---|---|
| Occupation Class 1 | | | |
| Singletons | 101 | 92 | 98 |
| First Duplicates | 111 | .. | 113 |
| Extra Duplicates | 107 | 117 | 109 |
| ex D | 104 | 94 | 101 |
| All Duplicates | 109 | 119 | 110 |
| cum D | 105 | 96 | 102 |
| | | | |
| All Occupation Classes | | | |
| Singletons | 101 | 89 | 96 |
| First Duplicates | 111 | 116 | 113 |
| Extra Duplicates | 107 | 114 | 108 |
| ex D | 104 | 90 | 99 |
| All Duplicates | 109 | 115 | 110 |
| cum D | 105 | 92 | 101 |

Table C5: Numbers of Deaths, Males, 1991-2002,
subdivided by Occupation Class and Duplicates Category.

| | DP 1 | DP 4 | DP 13 | DP 26 | DP 52 | All DPs |
|---|---|---|---|---|---|---|
| Occupation Class 1 | | | | | | |
| Singletons | 59 | 154 | 201 | 149 | 71 | 645 |
| First Duplicates | 80 | 52 | 46 | 36 | 16 | 231 |
| Extra Duplicates | 191 | 98 | 88 | 49 | 35 | 462 |
| ex D | 139 | 206 | 247 | 185 | 87 | 876 |
| All Duplicates | 271 | 150 | 134 | 85 | 51 | 693 |
| cum D | 330 | 304 | 335 | 234 | 122 | 1,338 |
| Occupation Class 2 | | | | | | |
| Singletons | 0 | 55 | 69 | 39 | 12 | 178 |
| First Duplicates | 0 | 5 | 4 | 3 | 1 | 13 |
| Extra Duplicates | 0 | 10 | 5 | 6 | 1 | 22 |
| ex D | 0 | 60 | 73 | 42 | 13 | 191 |
| All Duplicates | 0 | 15 | 9 | 9 | 2 | 35 |
| cum D | 0 | 70 | 78 | 48 | 14 | 213 |
| Occupation Class 3 | | | | | | |
| Singletons | 0 | 65 | 55 | 30 | 9 | 164 |
| First Duplicates | 0 | 1 | 3 | 1 | 1 | 6 |
| Extra Duplicates | 0 | 1 | 4 | 1 | 1 | 7 |
| ex D | 0 | 66 | 58 | 31 | 10 | 170 |
| All Duplicates | 0 | 2 | 7 | 2 | 2 | 13 |
| cum D | 0 | 67 | 62 | 32 | 11 | 177 |
| Occupation Class 4 | | | | | | |
| Singletons | 0 | 42 | 45 | 15 | 6 | 110 |
| First Duplicates | 0 | 2 | 1 | 0 | 1 | 4 |
| Extra Duplicates | 0 | 3 | 1 | 0 | 2 | 6 |
| ex D | 0 | 44 | 46 | 15 | 7 | 114 |
| All Duplicates | 0 | 5 | 2 | 0 | 3 | 10 |
| cum D | 0 | 47 | 47 | 15 | 9 | 120 |
| Occupation Class 5 | | | | | | |
| Singletons | 0 | 37 | 56 | 42 | 18 | 158 |
| First Duplicates | 0 | 1 | 1 | 2 | 0 | 4 |
| Extra Duplicates | 0 | 1 | 1 | 3 | 0 | 5 |
| ex D | 0 | 38 | 57 | 44 | 18 | 162 |
| All Duplicates | 0 | 2 | 2 | 5 | 0 | 9 |
| cum D | 0 | 39 | 58 | 47 | 18 | 167 |
| All Occupation Classes | | | | | | |
| Singletons | 59 | 353 | 426 | 275 | 116 | 1,255 |
| First Duplicates | 80 | 61 | 55 | 42 | 19 | 258 |
| Extra Duplicates | 191 | 113 | 99 | 59 | 39 | 502 |
| ex D | 139 | 414 | 481 | 317 | 135 | 1,513 |
| All Duplicates | 271 | 174 | 154 | 101 | 58 | 760 |
| cum D | 330 | 527 | 580 | 376 | 174 | 2,015 |

Table C6: Ratios 100 × *A*/*E* for Deaths, Males, 1991-2002, subdivided by Occupation Class and Duplicates Category. *E* is calculated using IPM 1991-98.

| | DP 4 | DP 13 | DP 26 | All DPs |
|---|---|---|---|---|
| Occupation Class 1 | | | | |
| Singletons | 92 | 101 | .. | 90 |
| First Duplicates | .. | .. | .. | 82 |
| Extra Duplicates | .. | .. | .. | 85 |
| ex D | 89 | 100 | .. | 88 |
| All Duplicates | 87 | 104 | .. | 84 |
| cum D | 89 | 102 | .. | 87 |
| All Occupation Classes | | | | |
| Singletons | 75 | 83 | 84 | 80 |
| First Duplicates | .. | .. | .. | 78 |
| Extra Duplicates | 85 | .. | .. | 82 |
| ex D | 75 | 83 | 83 | 80 |
| All Duplicates | 81 | 94 | 68 | 81 |
| cum D | 77 | 85 | 79 | 80 |

Table C7: Numbers of Deaths, Females, 1991-2002,
subdivided by Occupation Class and Duplicates Category.

| | DP 1 | DP 4 | DP 13 | DP 26 | DP 52 | All DPs |
|---|---|---|---|---|---|---|
| Occupation Class 1 | | | | | | |
| Singletons | 5 | 22 | 43 | 42 | 27 | 141 |
| First Duplicates | 4 | 8 | 6 | 8 | 2 | 28 |
| Extra Duplicates | 8 | 11 | 7 | 9 | 3 | 38 |
| ex D | 9 | 30 | 49 | 50 | 29 | 169 |
| All Duplicates | 12 | 19 | 13 | 17 | 5 | 66 |
| cum D | 17 | 41 | 56 | 59 | 32 | 207 |
| Occupation Class 2 | | | | | | |
| Singletons | 0 | 6 | 14 | 4 | 4 | 28 |
| First Duplicates | 0 | 0 | 1 | 2 | 0 | 3 |
| Extra Duplicates | 0 | 0 | 1 | 3 | 0 | 4 |
| ex D | 0 | 6 | 15 | 6 | 4 | 31 |
| All Duplicates | 0 | 0 | 2 | 5 | 0 | 7 |
| cum D | 0 | 6 | 16 | 9 | 4 | 35 |
| Occupation Class 3 | | | | | | |
| Singletons | 0 | 1 | 6 | 1 | 3 | 11 |
| First Duplicates | 0 | 0 | 0 | 1 | 0 | 1 |
| Extra Duplicates | 0 | 0 | 0 | 1 | 0 | 1 |
| ex D | 0 | 1 | 6 | 2 | 3 | 12 |
| All Duplicates | 0 | 0 | 0 | 2 | 0 | 2 |
| cum D | 0 | 1 | 6 | 3 | 3 | 13 |
| Occupation Class 4 | | | | | | |
| Singletons | 0 | 0 | 0 | 0 | 0 | 1 |
| First Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| Extra Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| ex D | 0 | 0 | 0 | 0 | 0 | 1 |
| All Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| cum D | 0 | 0 | 0 | 0 | 0 | 1 |
| Occupation Class 5 | | | | | | |
| Singletons | 0 | 1 | 6 | 10 | 3 | 20 |
| First Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| Extra Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| ex D | 0 | 1 | 6 | 10 | 3 | 20 |
| All Duplicates | 0 | 0 | 0 | 0 | 0 | 0 |
| cum D | 0 | 1 | 6 | 10 | 3 | 20 |
| All Occupation Classes | | | | | | |
| Singletons | 5 | 30 | 69 | 57 | 37 | 201 |
| First Duplicates | 4 | 8 | 7 | 11 | 2 | 32 |
| Extra Duplicates | 8 | 11 | 8 | 13 | 3 | 43 |
| ex D | 9 | 38 | 76 | 68 | 39 | 233 |
| All Duplicates | 12 | 19 | 15 | 24 | 5 | 75 |
| cum D | 17 | 49 | 84 | 81 | 42 | 276 |

# Part D: An Analysis of the Distribution of Duplicates

## D1 INTRODUCTION

D1.1 The CMI Income Protection (IP) Committee has revised the method used for identifying Duplicate policies in the individual IP investigations. The revised method is described in the paper "*The Identification of Duplicates*" (Part B). Duplicate policies occur when it appears that one policyholder has a number of separate Claim records with sufficiently similar conditions for it to be better to treat them as one Claim rather than as several. An analysis was carried out using the file of Claims for 1991 to 2002, with, in the first place, all Claims included. Then the Claim records were sorted so that records with a matching set of characteristics, described in Part B, were identified, and a specified one of these was selected to represent the bundle of assumed Duplicates.

D1.2 It is of interest to investigate whether the frequency distribution of the number of these Duplicates could be assumed to follow a recognised statistical distribution. In this paper we give the results of our investigations. In Appendix F of the paper "*Sickness Experience 1975-78 for Individual PHI Policies*" in CMIR 7, it was suggested that the frequency of the number of Duplicates might follow a geometric distribution, with different parameters for those cases with different Deferred Periods. We investigate the geometric distribution as a candidate, but also consider other distributions.

D1.3 In Section D2 we describe the fitting process, the assumptions and the statistical distributions that we have tried, and how these have been applied to each Deferred Period separately. In Section D3 we consider comparisons between categories on a non-parametric basis, that is, without making any assumptions about distributions. We summarise our conclusions in Section D4.

## D2 FITTING DISTRIBUTIONS

### D2.1 *Preliminary*

D2.1.1 We start by showing, in Table D2.1.1, the overall numbers of cases with 1, 2, 3, etc similar Claim records (numbers excluding Duplicates or "exD") along with the total number of Claim records (including Duplicates or "cumD"). We could alternatively describe those with only one Claim record as "Singletons" (of which there are 120,563), those with two Claim records as having one Duplicate, etc.

D2.1.2 We can see that the great majority of cases, even among the cumD cases, are Singletons. The numbers with multiple Duplicates reduce steadily up to three cases with 13 policies each, and then there are three outliers, one with 21 policies and two with 22 policies. These are not impossible numbers if a policyholder were to have effected a new policy every year for a number of years.

Table D2.1.1. Number of cases with 1, 2, 3, …matching policies (Claim records)

| Number of policies | Number of cases, exD | Number of cases, cumD |
|---|---|---|
| 1 | 120,563 | 120,563 |
| 2 | 15,964 | 31,928 |
| 3 | 6,583 | 19,749 |
| 4 | 3,311 | 13,244 |
| 5 | 1,871 | 9,355 |
| 6 | 1,075 | 6,450 |
| 7 | 387 | 2,709 |
| 8 | 131 | 1,048 |
| 9 | 73 | 657 |
| 10 | 36 | 360 |
| 11 | 16 | 176 |
| 12 | 10 | 120 |
| 13 | 3 | 39 |
| … | | |
| 21 | 1 | 21 |
| 22 | 2 | 44 |
| Total | 150,026 | 206,463 |

D2.1.3 The mean number of policies (Claim records) per life claiming IP benefit, counting all cases, is 1.38. This is equal to the number of cases in the cumD file (206,463) divided by the number in the exD file (150,026). We refer to this statistic as "Mean Policies".

D2.1.4 We can also calculate another mean, the mean number of Duplicate policies among those cases that have at least one Duplicate. There are 29,463 such cases, 19.6% of all the exD cases, and they have 85,900 policies in all, or 56,437 *extra* policies between them, an average of 1.92. We refer to this statistic as "Mean Duplicates". Both means are of some interest.

D2.1.5 In CMIR 7 it was suggested that the frequency of Duplicates for the different Deferred Periods might have the same (geometric) distribution, but with different parameters so we too investigate the data separated by Deferred Period. In Table D2.1.2 we show the numbers of Claims, exD, for DP1, DP4 (including cases coded as "999" for "one month"), DP13, DP26, DP52 and all other Deferred Periods (DP0, DP2, DP8 and uncommon or "Odd Deferred Periods", described collectively as "Other DPs").

Table D2.1.2. Number of cases (ex D) with 1, 2, 3, …matching policies (Claim records)

| Number of policies | DP1 | DP4 | DP13 | DP26 | DP52 | Other DPs | Total |
|---|---|---|---|---|---|---|---|
| 1 | 11,825 | 31,497 | 33,539 | 28,788 | 13,074 | 1,840 | 120,563 |
| 2 | 6,741 | 3,176 | 2,348 | 2,491 | 1,154 | 54 | 15,964 |
| 3 | 3,824 | 910 | 657 | 836 | 356 | 0 | 6,583 |
| 4 | 2,331 | 352 | 226 | 316 | 86 | 0 | 3,311 |
| 5 | 1,536 | 117 | 82 | 91 | 45 | 0 | 1,871 |
| 6 | 890 | 69 | 39 | 54 | 23 | 0 | 1,075 |
| 7 | 326 | 6 | 24 | 17 | 14 | 0 | 387 |
| 8 | 82 | 18 | 4 | 24 | 3 | 0 | 131 |
| 9 | 41 | 8 | 0 | 20 | 4 | 0 | 73 |
| 10 | 15 | 0 | 3 | 10 | 8 | 0 | 36 |
| 11 | 1 | 1 | 0 | 9 | 5 | 0 | 16 |
| 12 | 0 | 0 | 2 | 0 | 8 | 0 | 10 |
| 13 | 2 | 0 | 0 | 1 | 0 | 0 | 3 |
| … | | | | | | | |
| 21 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 22 | 0 | 0 | 2 | 0 | 0 | 0 | 2 |
| Total | 27,614 | 36,154 | 36,927 | 32,657 | 14,780 | 1,894 | 150,026 |

D2.1.6 We can quickly see that the number of Duplicates is higher for DP1 than for any other Deferred Period. Each of the first few numbers in the DP1 column is roughly half the number above it, whereas for all other Deferred Periods they are much less than half. For Other DPs there is hardly any distribution at all, with only a few cases having two policies, and none more than two. We show statistics for the separate Deferred Periods in Table D2.1.3. We see that the mean numbers and percentage of Duplicates are much bigger in DP1 than in any other Deferred Period; that DP4, DP13, DP26 and DP52 have numbers that are not very far apart; and that Other DPs has very few Duplicates.

Table D2.1.3. Statistics for Deferred Periods

| | DP1 | DP4 | DP13 | DP26 | DP52 | Other DP | Total |
|---|---|---|---|---|---|---|---|
| Number cumD | 62,617 | 43,255 | 42,073 | 39,024 | 17,546 | 1,948 | 206,463 |
| Number exD | 27,614 | 36,154 | 36,927 | 32,657 | 14,780 | 1,894 | 150,026 |
| Mean Policies | 2.27 | 1.20 | 1.14 | 1.19 | 1.19 | 1.03 | 1.38 |
| Number with $j \geq 2$ | 15,789 | 4,657 | 3,388 | 3,869 | 1,706 | 54 | 29,463 |
| Percent with Duplicates | 57.2 | 12.9 | 9.2 | 11.8 | 11.5 | 2.9 | 19.6 |
| Mean Duplicates | 2.22 | 1.52 | 1.52 | 1.65 | 1.62 | 1.00 | 1.92 |

## D2.2 *Assumptions and distributions*

D2.2.1 Many of the standard statistical distributions of discrete data give a probability that there are $k$ “events” (in this case policies or Duplicates), with $k$ = 0, 1, 2, …, increasing through the integers either to some limit, $K$, or indefinitely. We obviously do not have any record of cases with zero policies, but there is a large population of persons who do have zero IP policies. We use a number of different ways of modelling the distribution of Duplicates.

Throughout we denote the number of cases with $j$ policies, $j = 1, 2, 3, \dots$ as $C_j$, and the number of cases in the corresponding distribution as $N_k$, $k = 0, 1, 2, \dots$

D2.2.2 Model 1: The first possibility is that we count Duplicates, counting Singletons as having zero Duplicates ($k = 0$), those with two policies as having one Duplicate ($k = 1$) and so on. Thus we make $C_j$ ($j = 1, 2, 3, \dots$ ) correspond with $N_{j-1}$.

D2.2.3 Model 2: Next we assume that we have records only of cases with $k = 1, 2, 3, \dots$ events and that there is an unknown number of cases with $k = 0$ events, so we have what we describe as a "diminished" distribution. In a diminished distribution we assume that we observe cases with $k$ events, $k = 1, 2, 3, \dots$, but that there is an unknown number, $N_0$, of cases with zero events. We denote the probabilities of $k$ events in the full distribution as $p(k)$, $k = 0, 1, 2, \dots$ . Thus the proportion of unobserved cases is $p(0)$ and the proportion observed is $(1 - p(0))$. So in the diminished distribution the probability of $k$ events is

$$p^*(k) = \quad p(k) / (1 - p(0)) \qquad\qquad k = 1, 2, 3, \dots$$

We can estimate the parameters of such a diminished distribution and we can then estimate the number of cases with zero events, $N_0$, who might be considered an unobserved part of the population.

D2.2.4 So for our second model, we use the full data and assume a diminished distribution, counting those with one policy as having one event, etc. so that we make $C_j$ ($j = 1, 2, 3, \dots$ ) correspond with $N_j$, and we estimate $N_0$.

D2.2.5 Model 3: The third possibility is that we ignore the Singletons, and count those with two policies as having zero "excess Duplicates", those with three policies as having one excess Duplicate, and so on. So we make $C_j$ ($j = 2, 3, 4, \dots$ ) correspond with $N_{j-2}$. We then fit a full distribution. Another way of looking at this is to assume a Bernoulli distribution of Singletons and Duplicates, for which we can easily estimate the parameter, with a subsidiary distribution for the number of policies if the case is a Duplicate.

D2.2.6 Model 4: Finally, we assume that the observed population is split into those who would never have Duplicates, and those that might have Duplicates, but have not yet effected a second policy. (The proportions might well vary by the office concerned.) So again we use a diminished distribution, but starting with those who have two policies, treating them as having their first "Duplicate" $k = 1$, and later estimating the number of Singletons that might have had Duplicates, but have not got any yet. Thus we make $C_j$ ($j = 2, 3, 4, \dots$ ) correspond with $N_{j-1}$, and we estimate with $N_0$ the number of Singletons that might have had Duplicates, but have none so far.

D2.2.7 The distributions we consider are the Poisson, geometric, negative binomial and binomial distributions. Each of these distributions of $k$ cases starts at $k = 0$, and increases through the integers, $k = 1, 2, 3, \dots$ , except for the binomial, which stops at some integer $r$.

D2.2.8 The probability function of the Poisson distribution, with parameter $\lambda > 0$, is:

$$p(k) \quad = \quad \exp(-\lambda).\lambda^k / k!$$

so that $p(0) = \exp(-\lambda)$ and thereafter $p(k) = p(k - 1).\ \lambda / k$.

D2.2.9 The probability function of the geometric distribution, with parameter $p$, $0 < p < 1$, is:

$$p(k) = p.(1-p)^k$$

so that $p(0) = p$ and subsequent values reduce geometrically, with $p(k) = p(k-1).q$, where $q = 1 - p$.

D2.2.10 The probability function of the negative binomial distribution, with parameters $p$ and $r$, $0 < p < 1$ and $0 < r$, is:

$$p(k) = [\, \Gamma(r+k) / (k!\, \Gamma(r)) \,].p^r.(1-p)^k$$

so that $p(0) = p^r$ and thereafter $p(k) = p(k-1).(r + k - 1).(1 - p) / k$. Note that if $r = 1$, this simplifies to the geometric distribution.

D2.2.11 The probability function of the binomial distribution, with parameters $p$ and $r$, $0 < p < 1$ and $0 < r$, with $r$ integral, is:

$$p(k) = [\, r! / (k!\, (r-k)!) \,].p^k\, .(1-p)^{r-k}$$

so that $p(0) = (1 - p)^r$ and thereafter $p(k) = p(k-1).\ [\ (r - k + 1) / k\ ].\ p /(1 - p)$, as far as $p(r)$.

We choose the integral value of $r$ that gives the maximum log likelihood. Note that, if $p$ and $r$ increase without limit, but so that their product $p.r$ tends to some constant $\lambda$, then this tends to a Poisson distribution with parameter $\lambda$. In practice, since the value of $r$ is unknown, it is better to parameterise the distribution in terms of $r$ and $\lambda = p.r$, because the maximum likelihood estimate of $\lambda$ is independent of $r$ and is just the same as for the Poisson distribution, being the mean number. We can then choose the integral value of $r$ which maximises the likelihood.

D2.2.12 The probability functions of the corresponding diminished distributions are calculated from $p^*(k) = p(k) / (1 - p(0))$, where $p(k)$ is for the basic distribution, $p^*(k)$ for the diminished distribution. If $N = \Sigma_{j\geq 1} N_j$ is the total number of observed cases with one or more events, then an estimate of $N_0$, the number of cases with zero events, is:

$$N_0 \approx N.p(0) / [\, 1 - p(0) \,]$$

So,

diminished Poisson: $p^*(k) = [\, \exp(-\lambda).\lambda^k / k!\, ] / [\, 1 - \exp(-\lambda) \,]$

$N_0 = N.\ \exp(-\lambda) / [\, 1 - \exp(-\lambda) \,]$

diminished geometric: $p^*(k) = [\, p.(1-p)^k \,] / (1-p)$
$= p.(1-p)^{k-1}$

$N_0 = N.\ p / (1-p)$

This is just another geometric distribution, with the same value of the parameter $p$.

diminished negative binomial: $p^*(k) \quad = \quad [\, \Gamma(r + k) / (k!\, \Gamma(r)\, ].p^r.(1 - p)^k \; / (1 - p^r)$

$$N_0 \quad = \quad N.\, p^r / (1 - p^r)$$

Normally the parameters of the diminished distributions have the same ranges as the parameters of the corresponding basic distributions. But, curiously, the possible range for the diminished negative binomial is different, as we discuss in paragraph D2.5.3.

diminished binomial: $p^*(k) \quad = \quad [\, r! / (k!\, (r - k)!))\, ].p^k \,.\, (1 - p)^{r-k} / (1 - p)^r$

$$N_0 \quad = \quad N.(1 - p)^r / [\, 1 - (1 - p)^r\, ]$$

D2.2.13 The parameters are estimated in each case in a suitable way, sometimes algebraically, sometimes by successive approximation.

D2.2.14 Another way of deriving the diminished distributions would be to assume an arbitrary value $a$ ($0 < a$) for $p(1)$, then to use the relevant recurrence relation to derive formulae for $p(j)$, $j > 1$, then to sum these values, in which $a$ is a common factor, and finally to chose the value of $a$ that makes the sum of the probabilities unity. This makes no assumptions about $N_0$, but it avoids the apparent problem that arises later when the maximum likelihood estimate of the value of $r$ for the negative binomial distribution is less than zero. We could start at any higher number of cases, say $j_0$, for example $j_0 = 2$, and do the calculations similarly. For the (diminished) negative binomial the permitted range of $r$ is $r > -j_0$.

## D2.3 *The fitting process*

D2.3.1 We fit each of the four distributions to each of our four models, to each of the Deferred Periods DP1, DP4, DP13, DP26 and DP52, and also to the Other DPs on their own. We also try the combination of DPs 4, 13, 26 and 52, because they are rather similar, and the sum of all Deferred Periods, with and without the Other DPs. We also investigate whether the outliers with 21 and 22 Duplicate Claims in DP13 make any difference by fitting DP13 and any combination including DP13 both with and without these outliers.

D2.3.2 We find that it makes very little difference whether or not we include the outliers with 21 or 22 Claims in DP13. It also makes rather little difference whether we add the Other DPs in with all the common Deferred Periods or not. Adding DP1 to the other common Deferred Periods gives a noticeably worse fit, whereas the combination of DPs 4, 13, 26 and 52 is of interest.

D2.3.3 The following remarks apply to all the distributions with one trivial exception, the Other DPs, which we discuss in Section D2.9. For all the others, the Poisson distribution is nowhere at all a satisfactory fit. The binomial distribution finds a maximum likelihood fit with an extremely large value of $r$, so that it is indistinguishable from the Poisson, and is just as bad a fit. The geometric is a special case of the negative binomial, so the latter is always a better fit. In fact the negative binomial is always a significantly better fit, though sometimes the geometric is not exceptionally bad.

D2.3.4 It is not so easy to decide which of the four different models is most satisfactory. Sometimes the diminished negative binomial fits better than the full version; sometimes it is the other way around. Sometimes omitting the first entry gives a substantial improvement in the fit, sometimes not. Each case is different. For the geometric distribution, the diminished version is always identical with the full version. We now discuss the Deferred Periods in turn.

## D2.4 *DP1*

D2.4.1 We consider first DP1. This has 27,614 cases (exD), distributed as shown in Table D2.1.2. The results of fitting negative binomial and diminished negative binomial distributions are very much better than for any of the other distributions, and they are the only ones considered, for this and for all other Deferred Periods. Values of the parameters, fitted by maximum likelihood, are shown in Table D2.4.1, along with standard errors (in parentheses), and the correlation coefficient ($\rho$) between the estimates of $p$ and $r$. Also shown are the sum of the squares of the standardised differences, adjusted for continuity corrections ($X^2$), and the number of degrees of freedom (df) when $X^2$ is compared with a $\chi^2$ distribution. Finally, for models 2 and 4, we also show the estimated values of $N_0$, the implied number of cases with $k = 0$ for the diminished distributions.

D2.4.2 We see that the values of the parameters are broadly similar in each case, though they are significantly different when the standard errors are taken into account. The large number of cases means that the standard errors are relatively small. The parameter estimates are highly correlated. The values of $X^2$ are large compared with the numbers of degrees of freedom, showing that the fit is a very long way from being perfect. It is not worth showing the $p$-values for the $\chi^2$ test, because they are so small.

Table D2.4.1. Parameters for fitted negative binomial and diminished negative binomial to DP1 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $N_0$ |
|---|---|---|---|---|---|---|
| Model 1 | 0.4924 | 1.2298 | | 396.1 | 11 | |
| | (0.0054) | (0.0251) | (0.94) | | | |
| Model 2 | 0.5286 | 1.6611 | | 367.6 | 11 | 14,665.1 |
| | (0.0077) | (0.0663) | (0.97) | | | |
| Model 3 | 0.5479 | 1.4747 | | 308.2 | 9 | |
| | (0.0079) | (0.0447) | (0.95) | | | |
| Model 4 | 0.6120 | 2.4441 | | 263.4 | 8 | 6,806.8 |
| | (0.0113) | (0.1403) | (0.98) | | | |

D2.4.3 Details of the fits are shown in Table D2.4.2a and Table D2.4.2b, which show: the actual number of cases with $k$ policies (Claim records), $A$; the expected number according to each of the distributions, $E$; the difference between $A$ and $E$, $D = A - E$; and the value of the squared standardised and adjusted difference, $z^2 = (D \pm \text{adjustment})^2/E$, which sum to $X^2$. Values of $j$ are grouped so that the value of $E$ in each cell is at least 5.

D2.4.4 When the models including Singletons are compared, Model 2 is seen to be rather better than Model 1, but in both cases E is much bigger than A for $j = 2$ and 3, much smaller for $j = 5$ and 6, and rather bigger for all $j \geq 7$. For the diminished distribution, the estimate of $N_0$ is 14,665.1. If we consider this as an estimate of the number of persons with no IP policies, which obviously is enormously much larger than this, it seems poor. But as a filler to complete a hypothetical distribution, it does not seem too bad.

D2.4.5 When the models omitting Singletons are compared, Model 4 is seen to be rather better than Model 3, but the discrepancies are much the same as for Models 1 and 2. The values of $X^2$ for Models 3 and 4 cannot be directly compared with those for Models 1 and 2, but if we deduct the values of $z^2$ for $j$ = 1 from the values of $X^2$ for the latter models, i.e. 6.3 from 396.1, and 10.2 from 367.6, we see that the fit of Models 3 and 4 for the case with $j \geq 2$ is rather better, with Model 4 better than Model 3. The estimate of $N_0$ for Model 4 is 6,806.8 and is not ridiculous compared with the observed value of 11,825.

Table D2.4.2a. Results of fitting negative binomial and diminished negative binomial to DP1 data, Models 1 and 2, including "Singletons".

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number of policies | Actual *A* | Expected *E* | *D=A–E* | $z^2$ | Expected *E* | *D=A–E* | $z^2$ |
| 0 | | | | | [*14,665.1*] | | |
| 1 | 11,825 | 11,555.3 | 269.7 | 6.3 | 11,482.1 | 342.9 | 10.2 |
| 2 | 6,741 | 7,212.8 | –471.8 | 30.8 | 7,201.0 | –460.0 | 29.3 |
| 3 | 3,824 | 4,081.6 | –257.6 | 16.2 | 4,142.1 | –318.1 | 24.4 |
| 4 | 2,331 | 2,230.4 | 100.6 | 4.5 | 2,275.1 | 55.9 | 1.4 |
| 5 | 1,536 | 1,197.1 | 338.9 | 95.7 | 1,214.1 | 321.9 | 85.1 |
| 6 | 890 | 635.5 | 254.5 | 101.5 | 635.3 | 254.7 | 101.7 |
| 7 | 326 | 334.9 | –8.9 | 0.2 | 327.7 | –1.7 | 0.0 |
| 8 | 82 | 175.6 | –93.6 | 49.3 | 167.2 | –85.2 | 42.9 |
| 9 | 41 | 91.7 | –50.7 | 27.5 | 84.6 | –43.6 | 22.0 |
| 10 | 15 | 47.7 | –32.7 | 21.8 | 42.5 | –27.5 | 17.2 |
| 11 | 1 | 24.8 | –23.8 | 21.9 | 21.2 | –20.2 | 18.4 |
| 12 | 0 | 12.8 | –12.8 | 11.9 | 10.6 | –10.6 | 9.6 |
| 13 | 2 | 6.6 | –4.6 | 2.6 | 5.2 | –3.2 | 1.4 |
| > 13 | 0 | 7.1 | –7.1 | 6.1 | 5.1 | –5.1 | 4.1 |
| Total | 27,614 | 27,614.0 | 0.0 | 396.1 | 27,614.0 | 0.0 | 367.6 |

Table D2.4.2b. Results of fitting negative binomial and diminished negative binomial to DP1 data, Models 3 and 4, excluding "Singletons".

| | | Negative binomial | | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|---|
| Number of policies | Actual *A* | Expected *E* | *D=A–E* | $z^2$ | | Expected *E* | *D=A–E* | $z^2$ |
| 1 | | | | | | [*6,806.8*] | | |
| 2 | 6,741 | 6,501.3 | 239.7 | 8.8 | | 6,453.8 | 287.2 | 12.7 |
| 3 | 3,824 | 4,334.6 | –510.6 | 60.0 | | 4,311.4 | –487.4 | 55.0 |
| 4 | 2,331 | 2,424.9 | –93.9 | 3.6 | | 2,477.7 | –146.7 | 8.6 |
| 5 | 1,536 | 1,269.8 | 266.2 | 55.6 | | 1,308.2 | 227.8 | 39.5 |
| 6 | 890 | 642.2 | 247.8 | 95.2 | | 654.1 | 235.9 | 84.8 |
| 7 | 326 | 317.9 | 8.1 | 0.2 | | 314.8 | 11.2 | 0.4 |
| 8 | 82 | 155.1 | –73.1 | 34.0 | | 147.3 | –65.3 | 28.5 |
| 9 | 41 | 74.9 | –33.9 | 14.9 | | 67.5 | –26.5 | 10.0 |
| 10 | 15 | 35.9 | –20.9 | 11.6 | | 30.4 | –15.4 | 7.3 |
| 11 | 1 | 17.1 | –16.1 | 14.2 | | 13.5 | –12.5 | 10.7 |
| 12 | 0 | 8.1 | –8.1 | 7.1 | } | 10.4 | –8.4 | 6.0 |
| ≥ 13 | 2 | 7.2 | –5.2 | 3.0 | } | | | |
| Total | 15,789 | 15,789.0 | 0.0 | 308.2 | | 15,789.0 | 0.0 | 263.4 |

## D2.5 *DP4*

D2.5.1 DP4 has 36,154 cases (exD), distributed as shown in Table D2.1.2. The same statistics are shown in Table D2.5.1 as are shown for DP1 in Table D2.4.1, with the addition of *p*-values for the $\chi^2$ test, which are small, but are worth showing. The values of $X^2$ are very much smaller than for DP1, but still significantly large.

D2.5.2 Details are shown in Table D2.5.2a and Table D2.5.2b. The four Models all show much better fits than for DP1, with a reasonable fluctuation of signs of the differences. Only one value of *j* stands out. There are only 6 cases with 7 Duplicate Claims compared with over 20 expected on each of the Models.

D2.5.3 We note, however, that the value of *r* for Model 2 is negative, which is unacceptable for a negative binomial distribution, but turns out to be acceptable for a diminished negative binomial. For a proper negative binomial we start with $p(0) = p^r$. If $r < 0$, then $p^r > 1$, so the divisor for the diminished distribution, $1 – p^r$, is negative. Further, from the recurrence relation:

$$p(k) \quad = \quad p(k–1).(r + k – 1).(1 – p) / k$$

we see that $p(1) = p(0).\ r\ (1 – p)$, which is also negative. Successive values of $p(k)$ are negative if $–1 < r < 0$. When these are divided by $(1 – p^r)$ they are all positive. If $r < –1$ then some or all values of $p(k)$ for $k > 1$ are positive, so become negative when divided by $(1 – p^r)$. Thus, for $0 < p < 1$ and $–1 < r < 0$, although the negative binomial distribution is not a proper one, the diminished version is satisfactory. However, it results in the estimate of $N_0$ being negative, so that it is not a sensible estimate of the numbers of cases with zero Duplicates. It turns out that this “improper” diminished negative binomial is a reasonably good fit to the data in a number of cases, as in Model 2 for DP4.

D2.5.4 When we compare the Models for DP4 we see that Model 2 is rather poorer than Model 1, and Models 1, 3 and 4 have almost the same values of $X^2$, even allowing for the omission of the data for $j = 1$ in the later Models.

Table D2.5.1. Parameters for fitted negative binomial and diminished negative binomial to DP4 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $p(\chi^2)$ | $N_0$ |
|---|---|---|---|---|---|---|---|
| Model 1 | 0.5114 | 0.2056 | | 24.34 | 6 | 0.0005 | |
| | (0.0087) | (0.0062) | (0.88) | | | | |
| Model 2 | 0.3895 | –0.6632 | | 30.99 | 7 | 0.0001 | –77,764.1 |
| | (0.0112) | (0.0104) | (0.89) | | | | |
| Model 3 | 0.5481 | 0.6365 | | 24.01 | 5 | 0.0002 | |
| | (0.0172) | (0.0407) | (0.92) | | | | |
| Model 4 | 0.5030 | 0.1724 | | 23.67 | 7 | 0.0003 | 37,016.6 |
| | (0.0235) | (0.0845) | (0.95) | | | | |

Table D2.5.2a. Results of fitting negative binomial and diminished negative binomial to DP4 data, Models 1 and 2, including “Singletons”.

| | | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 0 | | | | | | [*–77,764.1*] | | |
| 1 | 31,497 | | 31,498.0 | –1.0 | 0.00 | 31,485.7 | 11.3 | 0.00 |
| 2 | 3,176 | | 3,164.0 | 12.0 | 0.04 | 3,237.1 | –61.1 | 1.13 |
| 3 | 910 | | 931.8 | –21.8 | 0.49 | 880.6 | 29.4 | 0.95 |
| 4 | 352 | | 334.7 | 17.3 | 0.84 | 314.1 | 37.9 | 4.46 |
| 5 | 117 | | 131.1 | –14.1 | 1.40 | 128.0 | –11.0 | 0.86 |
| 6 | 69 | | 53.9 | 15.1 | 3.98 | 56.5 | 12.5 | 2.56 |
| 7 | 6 | | 22.8 | –16.8 | 11.68 | 26.3 | –20.3 | 14.89 |
| 8 | 18 | | 9.9 | 8.1 | 5.86 | 12.7 | 5.3 | 1.80 |
| 9 | 8 | } | 7.9 | 1.1 | 0.05 | 6.3 | 1.7 | 0.22 |
| ≥ 10 | 1 | } | | | | 6.8 | –5.8 | 4.12 |
| Total | 36,154 | | 36,154.0 | 0.0 | 24.34 | 36,154.0 | 0.0 | 30.99 |

Table D2.5.2b. Results of fitting negative binomial and diminished negative binomial to DP4 data, Models 3 and 4, excluding “Singletons”.

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 1 | | | | | [*37,016.6*] | | |
| 2 | 3,176 | 3,176.0 | 0.0 | 0.00 | 3,172.6 | 3.4 | 0.00 |
| 3 | 910 | 913.6 | –3.6 | 0.01 | 924.4 | –14.4 | 0.21 |
| 4 | 352 | 337.8 | 14.2 | 0.55 | 332.7 | 19.3 | 1.06 |
| 5 | 117 | 134.2 | –17.2 | 2.07 | 131.1 | –14.1 | 1.42 |
| 6 | 69 | 55.1 | 13.9 | 3.25 | 54.4 | 14.6 | 3.66 |
| 7 | 6 | 23.1 | –17.1 | 11.93 | 23.3 | –17.3 | 12.12 |
| 8 | 18 | 9.8 | 8.2 | 6.04 | 10.2 | 7.8 | 5.20 |
| ≥ 9 | 9 | 7.4 | 1.6 | 0.16 | 8.3 | 0.7 | 0.00 |
| Total | 4,657 | 4,657.0 | 0.0 | 24.01 | 15,789.0 | 0.0 | 23.67 |

### D2.6 *DP13*

D2.6.1 DP13 has 36,927 cases, as shown in Table D2.1.2. The same statistics as before are shown in Table D2.6.1, with details in Table D2.6.2a and Table D2.6.2b. We now get very good fits for all the Models, with quite satisfactory values of $p(\chi^2)$. Now the values of $r$ for both Model 2 and Model 4 are negative, though the value for the latter is close to zero. Changing to a small positive value would make a small difference. Although $r$ = 0 is an unacceptable value, because $(1 - p^r)$ becomes zero, there is otherwise continuity across the zero boundary.

D2.6.2 We observe that Model 1 shows a lower value of $X^2$ than does Model 2, but Model 4 shows a lower one than Model 3. Model 3 is a little worse than Model 1, but Model 4 is rather better than Model 3. But all fit well, so there is little reason to prefer any Model to any other.

D2.6.3 If we omit the outliers with 21 and 22 Claims each (1 and 2 cases respectively) the fitted distributions are very close to those found when these outliers are included, which are the ones shown.

Table D2.6.1. Parameters for fitted negative binomial and diminished negative binomial to DP13 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $p(\chi^2)$ | $N_0$ |
|---|---|---|---|---|---|---|---|
| Model 1 | 0.4955 | 0.1369 | | 6.54 | 6 | 0.37 | |
| | (0.0099) | (0.0047) | (0.87) | | | | |
| Model 2 | 0.3566 | –0.7801 | | 11.62 | 7 | 0.11 | –66,825.7 |
| | (0.0124) | (0.0074) | (0.87) | | | | |
| Model 3 | 0.5045 | 0.5283 | | 8.36 | 5 | 0.14 | |
| | (0.0193) | (0.0366) | (0.90) | | | | |
| Model 4 | 0.4334 | –0.0691 | | 6.92 | 5 | 0.23 | –60,396.8 |
| | (0.0253) | (0.0686) | (0.93) | | | | |

Table D2.6.2a. Results of fitting negative binomial and diminished negative binomial to DP13 data, Models 1 and 2, including "Singletons".

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 0 | | | | | [*–66,825.7*] | | |
| 1 | 33,539 | 33,543.2 | –4.2 | 0.00 | 33,537.8 | 1.2 | 0.00 |
| 2 | 2,348 | 2,316.2 | 31.8 | 0.42 | 2,372.5 | –24.5 | 0.24 |
| 3 | 657 | 664.2 | –7.2 | 0.07 | 620.7 | 36.3 | 2.07 |
| 4 | 226 | 238.7 | –12.7 | 0.62 | 221.6 | 4.4 | 0.07 |
| 5 | 82 | 94.4 | –12.4 | 1.51 | 91.8 | –9.8 | 0.95 |
| 6 | 39 | 39.4 | –0.4 | 0.00 | 41.5 | –2.5 | 0.10 |
| 7 | 24 | 17.0 | 7.0 | 2.46 | 19.9 | 4.1 | 0.64 |
| 8 | 4 | 7.5 | –3.5 | 1.22 | 10.0 | –6.0 | 3.00 |
| 9 | 0 } | 6.3 | 1.7 | 0.24 | 5.1 | –5.1 | 4.19 |
| ≥ 10 | 8 } | | | | 6.0 | 2.0 | 0.36 |
| Total | 36,927 | 36,927.0 | 0.0 | 6.54 | 36,927.0 | 0.0 | 11.62 |

Table D2.6.2b. Results of fitting negative binomial and diminished negative binomial to DP13 data, Models 3 and 4, excluding "Singletons".

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 1 | | | | | [*–60,396.8*] | | |
| 2 | 2,348 | 2,360.3 | –12.3 | 0.06 | 2,362.9 | –14.9 | 0.09 |
| 3 | 657 | 617.9 | 39.1 | 2.42 | 623.2 | 33.8 | 1.78 |
| 4 | 226 | 234.0 | –8.0 | 0.24 | 227.3 | –1.3 | 0.00 |
| 5 | 82 | 97.7 | –15.7 | 2.36 | 94.3 | –12.3 | 1.49 |
| 6 | 39 | 42.7 | –3.7 | 0.24 | 42.0 | –3.0 | 0.15 |
| 7 | 24 | 19.2 | 4.8 | 0.98 | 19.6 | 4.4 | 0.79 |
| 8 | 4 | 8.7 | –4.7 | 2.06 | 9.4 | –5.4 | 2.55 |
| ≥ 9 | 8 | 7.6 | 0.4 | 0.00 | 9.3 | –1.3 | 0.07 |
| Total | 3,388 | 3,388.0 | 0.0 | 8.36 | 3,388.0 | 0.0 | 6.92 |

## D2.7 *DP26*

D2.7.1 DP26 has 32,657 cases as shown in Table D2.1.2. The same statistics as before are shown in Table D2.7.1, with details in Table D2.7.2a and Table D2.7.2b. The values of $X^2$ are higher than for DP4 and DP13, so the fits are not so good, though much better than for DP1. The estimated value of *r* for Model 2 is negative, but that for Model 4 is positive, though quite close to zero.

D2.7.2 The values of $X^2$ show that the fits of Models 2 and 4 are distinctly better than are those for Models 1 and 3, and Model 2 is better than Model 4. This is a little surprising, because one would expect that with less data the fit would improve.

D2.7.3 The deviations for $j = 5$ are always large and negative, and those for $j \geq 9$ are all large and positive.

Table D2.7.1. Parameters for fitted negative binomial and diminished negative binomial to DP26 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $N_0$ |
|---|---|---|---|---|---|---|
| Model 1 | 0.4411 | 0.1539 | | 65.67 | 7 | |
| | (0.0088) | (0.0047) | (0.85) | | | |
| Model 2 | 0.3037 | –0.7445 | | 48.32 | 7 | –55,524.4 |
| | (0.0108) | (0.0076) | (0.85) | | | |
| Model 3 | 0.4677 | 0.5672 | | 70.62 | 6 | |
| | (0.0162) | (0.0329) | (0.89) | | | |
| Model 4 | 0.4014 | 0.0160 | | 54.87 | 6 | 262,812.2 |
| | (0.0211) | (0.0622) | (0.92) | | | |

Table D2.7.2a. Results of fitting negative binomial and diminished negative binomial to DP26 data, Models 1 and 2, including “Singletons”.

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 0 | | | | | [–*55,524.4*] | | |
| 1 | 28,788 | 28,792.5 | –4.5 | 0.00 | 28,781.5 | 6.5 | 0.00 |
| 2 | 2,491 | 2,476.2 | 14.8 | 0.08 | 2,560.0 | –69.0 | 1.83 |
| 3 | 836 | 798.4 | 37.6 | 1.72 | 745.9 | 90.1 | 10.75 |
| 4 | 316 | 320.4 | –4.4 | 0.05 | 292.9 | 23.1 | 1.75 |
| 5 | 91 | 141.2 | –50.2 | 17.48 | 132.8 | –41.8 | 12.82 |
| 6 | 54 | 65.6 | –11.6 | 1.86 | 65.6 | –11.6 | 1.87 |
| 7 | 17 | 31.5 | –14.5 | 6.20 | 34.3 | –17.3 | 8.21 |
| 8 | 24 | 15.5 | 8.5 | 4.18 | 18.7 | 5.3 | 1.26 |
| 9 | 20 | 7.7 | 12.3 | 17.93 | 10.5 | 9.5 | 7.78 |
| 10 | 10 } | 8.1 | 11.9 | 16.16 | 6.0 | 4.0 | 2.01 |
| ≥ 11 | 10 } | | | | 8.9 | 1.1 | 0.04 |
| Total | 32,657 | 32,567.0 | 0.0 | 65.67 | 32,567.0 | 0.0 | 48.32 |

Table D2.7.2b. Results of fitting negative binomial and diminished negative binomial to DP26 data, Models 3 and 4, excluding “Singletons”.

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 1 | | | | | [*262,812.2*] | | |
| 2 | 2,491 | 2,514.1 | –23.1 | 0.20 | 2,518.8 | –27.8 | 0.30 |
| 3 | 836 | 759.1 | 76.9 | 7.68 | 765.9 | 70.1 | 6.32 |
| 4 | 316 | 316.7 | –0.7 | 0.00 | 308.1 | 7.9 | 0.18 |
| 5 | 91 | 144.3 | –53.3 | 19.29 | 139.0 | –48.0 | 16.26 |
| 6 | 54 | 68.5 | –14.5 | 2.85 | 66.8 | –12.8 | 2.28 |
| 7 | 17 | 33.3 | –16.3 | 7.50 | 33.5 | –16.5 | 7.61 |
| 8 | 24 | 16.4 | 7.6 | 3.02 | 17.2 | 6.8 | 2.30 |
| ..9 | 20 | 8.2 | 11.8 | 15.51 | 9.0 | 11.0 | 12.13 |
| ≥ 10 | 20 | 8.4 | 11.6 | 14.57 | 10.6 | 9.4 | 7.50 |
| Total | 3,869 | 3,869.0 | 0.0 | 70.62 | 3,869.0 | 0.0 | 54.87 |

## D2.8 *DP52*

D2.8.1 DP52 has 14,780 cases as shown in Table D2.1.2. The same statistics as before are shown in Table D2.8.1, with details in Table D2.8.2a and Table D2.8.2b. The values of $X^2$ are similar to those for DP26, so the fits are not so good, though much better than for DP1. The values of $r$ for Models 2 and 4 are both negative. The fits for Models 2 and 4 are quite a lot better than for Models 1 and 3, and that for Model 2 is better than for Model 4.

D2.8.2 The main discrepancies are a rather small number of actual cases for $j$ = 4, and a rather large number for $j \geq 10$.

Table D2.8.1. Parameters for fitted negative binomial and diminished negative binomial to DP52 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $N_0$ |
|---|---|---|---|---|---|---|
| Model 1 | 0.4485 | 0.1522 | | 80.50 | 6 | |
| | (0.0133) | (0.0070) | (0.85) | | | |
| Model 2 | 0.3089 | –0.7500 | | 40.34 | 6 | –25,237.4 |
| | (0.0164) | (0.0112) | (0.85) | | | |
| Model 3 | 0.3977 | 0.4102 | | 50.60 | 5 | |
| | (0.0337) | (0.0333) | (0.86) | | | |
| Model 4 | 0.3039 | –0.2788 | | 46.04 | 6 | –6,037.4 |
| | (0.0281) | (0.0586) | (0.89) | | | |

Table D2.8.2a. Results of fitting negative binomial and diminished negative binomial to DP52 data, Models 1 and 2, including “Singletons”.

| | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 0 | | | | | [*–25,237.4*] | | |
| 1 | 13,074 | 13,082.1 | –8.1 | 0.00 | 13,080.9 | –6.9 | 0.00 |
| 2 | 1,154 | 1,098.0 | 56.0 | 2.80 | 1,130.0 | 24.0 | 0.49 |
| 3 | 356 | 348.9 | 7.1 | 0.13 | 325.4 | 30.6 | 2.79 |
| 4 | 86 | 138.0 | –52.0 | 19.24 | 126.5 | –40.5 | 12.64 |
| 5 | 45 | 60.0 | –15.0 | 3.50 | 56.8 | –11.8 | 2.26 |
| 6 | 23 | 27.5 | –4.5 | 0.58 | 27.8 | –4.8 | 0.67 |
| 7 | 14 | 13.0 | 1.0 | 0.02 | 14.4 | –0.4 | 0.00 |
| 8 | 3 | 6.3 | –3.3 | 1.25 | 7.8 | –4.8 | 2.36 |
| ≥.9 | 25 | 6.3 | 18.7 | 52.99 | 10.4 | 14.6 | 19.14 |
| Total | 14,780 | 14,780.0 | 0.00 | 80.50 | 14,780.0 | 0.0 | 40.34 |

Table D2.8.2b. Results of fitting negative binomial and diminished negative binomial to DP52 data, Models 3 and 4, excluding “Singletons”.

| | | | Negative binomial | | | Diminished negative binomial | | |
|---|---|---|---|---|---|---|---|---|
| Number<br>of policies | Actual<br>*A* | | Expected<br>*E* | *D*=*A*–*E* | $z^2$ | Expected<br>*E* | *D*=*A*–*E* | $z^2$ |
| 1 | | | | | | [*–6,037.4*] | | |
| 2 | 1,154 | | 1,168.7 | –14.7 | 0.17 | 1,171.8 | –17.8 | 0.26 |
| 3 | 356 | | 288.8 | 67.2 | 15.43 | 294.1 | 61.9 | 12.82 |
| 4 | 86 | | 122.6 | –36.6 | 10.65 | 117.5 | –31.5 | 8.16 |
| 5 | 45 | | 59.3 | –14.3 | 3.23 | 55.6 | –10.6 | 1.84 |
| 6 | 23 | | 30.5 | –7.5 | 1.60 | 28.8 | –5.8 | 0.98 |
| 7 | 14 | | 16.2 | –2.2 | 0.18 | 15.8 | –1.8 | 0.10 |
| 8 | 3 | | 8.8 | –5.8 | 3.19 | 9.0 | –6.0 | 3.34 |
| 9 | 4 | } | 11.1 | 13.9 | 16.16 | 5.2 | –1.2 | 0.11 |
| ≥.10 | 25 | } | | | | 8.2 | 12.8 | 18.44 |
| Total | 1,706 | | 1,706.0 | 0.0 | 50.60 | 1,706.0 | 0.0 | 46.04 |

### D2.9 *Other DPs*

D2.9.1 There is rather little data for the Other DPs, 1,840 Singletons and 54 cases with two matching Claim records. A simple binomial with $r$ = 1 and $p$ = 0.0285 fits perfectly, as does a diminished binomial with $r$ = 2 and $p$ = 0.1109. But it is unreasonable to assume that the number of Duplicates is limited in this way. Cases with more Duplicate Claims could presumably occur. A Poisson with $\lambda$ = 0.0285, or a geometric distribution with $p$ = 0.9723 also fit almost exactly. There is no need to consider this further.

### D2.10 *Combinations of Deferred Periods*

D2.10.1 Since the parameters for DP4, DP13, DP26 and DP52 are moderately similar, it is worth fitting distributions to the combined data for these Deferred Periods. The resulting statistics are shown in Table D2.10.1. However, when we apply the collective parameters to the data for the individual Deferred Periods the results are always worse, sometimes a great deal worse (Models 1 and 2 for DP4 and DP13), sometimes only a bit worse, but significantly so. We do not show details.

Table D2.10.1. Parameters for fitted negative binomial and diminished negative binomial to combined DP4, DP13, DP26 and DP52 data, Models 1 to 4

| | $p$ | $r$ | $\rho$ | $X^2$ | df | $N_0$ |
|---|---|---|---|---|---|---|
| Model 1 | 0.4768 | 0.1616 | | 112.11 | 8 | |
| | (0.0049) | (0.0027) | (0.86) | | | |
| Model 2 | 0.3422 | –0.7358 | | 51.93 | 9 | –220,856.3 |
| | (0.0061) | (0.0044) | (0.87) | | | |
| Model 3 | 0.4880 | 0.5430 | | 105.21 | 7 | |
| | (0.0091) | (0.0179) | (0.90) | | | |
| Model 4 | 0.4196 | –0.0334 | | 66.95 | 6 | –476,742.4 |
| | (0.0121) | (0.0339) | (0.93) | | | |

D2.10.2 We have also tried fitting to the combined data including DP1, either including or excluding the Other DPs. These latter make almost no difference. The best fitting distributions are always the negative binomials. However, the fit for each Model is sufficiently much worse for each Deferred Period that there is little advantage in pursuing this approach.

## D3 NON-PARAMETRIC COMPARISONS

### D3.1 *Comparisons: method*

D3.1.1 Independently of any assumptions about a particular distribution, we can compare the empirical distributions of any two categories within the data, such as Male and Female, DP1 and DP4, or different Years of Investigation. We can use the cumulative distributions and the equivalent of the Kolmogorov-Smirnov test for this.

D3.1.2 We denote two categories by the subscripts A and B, and use subscript $j$, $j$ = 1 to J (J = 22) to indicate the number of matching policies. Let $n_{Aj}$ be the number of cases in A with $j$ policies. Let $N_A = \sum_{j=1}^{J} n_{Aj}$ be the total number of cases (counting the exD cases) in A. We calculate the cumulative distribution function for A as $F_{Aj} = \left(\sum_{k=1}^{j} n_{Ak}\right)/N_A F_{Aj} = \left(\sum_{k=1}^{j} n_{Ak}\right)/N_A$. Do the same for B. We then calculate the signed maximum difference between the distribution functions for A and B, which we later refer to as the "K-S distance" statistic, and also calculate the absolute maximum difference, $D_{AB}$ = Max $|F_{Aj} - F_{Bj}|$. Finally, we calculate the K-S statistic, $K_{AB} = D_{AB} / \surd\{1/N_A + 1/N_B\}$. The K-S statistic has a unique distribution, and we can look up the complement of the K-S probability, which we denote p$K_{AB}$ from it. The value of p$K_{AB}$ is high when the maximum distance is small and the distributions are close. It is low if they are far apart.

D3.1.3 We can also compare the distribution functions of A and B for all values of $j$. If $F_{Aj} < F_{Bj}$ for all $j < J$, so that $F_A$ is increasing more slowly than $F_B$, then we can describe A as "dominating" B. This is technically "first order stochastic dominance". The mean of A is necessarily greater than the mean of B. Because there are three outliers with $j$ = 21 and 22, and there are few cases with $j$ > 10 it is possible that $F_{Aj} < F_{Bj}$ only up to say $j$ = 10. In that case we can say that A "practically dominates" B.

D3.1.4 There are many factors that can be used in the comparisons. Some, such as Male/Female have only two categories, so the K-S test is quite valid. Others, such as Deferred Period, have several categories, and yet others, such as Age, have many categories, so the K-S probabilities need to be interpreted with care. When many comparisons are being made, it would not be surprising for some of them to appear "statistically significant" even just by chance. We comment below only on the interesting comparisons. Note especially that we show only one-way marginal factors. There are many interactions between the factors, which only a full model allowing for all factors at once would explain. We do not do so here.

D3.1.5 For most of our comparisons we look at the Deferred Periods separately as well as at the overall numbers, since we know that their distributions are different, and there may be many interactions between Deferred Period and other features.

### D3.2 *Deferred Periods*

D3.2.1 We use this method first to compare the Deferred Periods. To demonstrate the method, we show in Table D3.2.1 detailed calculations for comparing DP4 and DP13. For each Deferred Period we show the numbers of cases with $j$ policies, the cumulative numbers, and the cumulative proportions, expressed as percentages. Finally we show the differences in the percentages as "DP4 – DP13". We observe that these change sign, which indicates that neither Deferred Period dominates the other DP4 has rather more Duplicates than DP13 on average, but DP13 has a longer tail, including the cases with 21 and 22 policies.

D3.2.2 The largest (absolute) percentage difference is for $j$ = 1, and is 3.71%. We divide this by $\surd[\ 1/N_A + 1/N_B\ ] = \surd[\ 1/36{,}154 + 1/36{,}927\ ] = 0.00739864$. The result is 5.01. This is well beyond a plausible number for the Smirnov distribution. The 5% point of this distribution is about 1.36 and the 1% point about 1.60, so there is no doubt that the data for two Deferred Periods cannot be treated as samples from the same original distribution. Nevertheless, the two distributions are not so far apart as some others, as we shall see.

Table D3.2.1. Comparison of DP4 and DP13: calculations for K-S test

| Number of policies | DP4 Number exD | DP4 Cumulative | DP4 Proportion % | DP13 Number exD | DP13 Cumulative | DP13 Proportion % | Difference (DP4–DP13) |
|---|---|---|---|---|---|---|---|
| 1 | 31,497 | 31,497 | 87.12 | 33,539 | 33,539 | 90.83 | –3.71 |
| 2 | 3,176 | 34,673 | 95.90 | 2,348 | 35,887 | 97.18 | –1.28 |
| 3 | 910 | 35,583 | 98.42 | 657 | 36,544 | 98.96 | –0.54 |
| 4 | 352 | 35,935 | 99.39 | 226 | 36,770 | 99.57 | –0.18 |
| 5 | 117 | 36,052 | 99.72 | 82 | 36,852 | 99.80 | –0.08 |
| 6 | 69 | 36,121 | 99.91 | 39 | 36,891 | 99.90 | 0.01 |
| 7 | 6 | 36,127 | 99.93 | 24 | 36,915 | 99.97 | –0.04 |
| 8 | 18 | 36,145 | 99.98 | 4 | 36,919 | 99.98 | 0.00 |
| 9 | 8 | 36,153 | 100.00 | 0 | 36,919 | 99.98 | 0.02 |
| 10 | 0 | 36,153 | 100.00 | 3 | 36,922 | 99.99 | 0.01 |
| 11 | 1 | 36,154 | 100.00 | 0 | 36,922 | 99.99 | 0.01 |
| 12 | 0 | 36,154 | 100.00 | 2 | 36,924 | 99.99 | 0.01 |
| 13 | 0 | 36,154 | 100.00 | 0 | 36,924 | 99.99 | 0.01 |
| … | | | | | | | |
| 21 | 0 | 36,154 | 100.00 | 1 | 36,925 | 99.99 | 0.01 |
| 22 | 0 | 36,154 | 100.00 | 2 | 36,927 | 100.00 | 0.00 |
| Total | 36,154 | | | 36,927 | | Max (abs) | 3.71 |
| | | | | | | K-S statistic | 5.01 |

D3.2.3 We have already noted that DP1 has many more Duplicates than other Deferred Periods; that Other DPs has very few; and that the other common Deferred Periods are not very different. In Table D3.2.2 we show comparison statistics, comparing each Deferred Period with each other. We show two numbers: first the K-S distance, the signed maximum difference between the cumulative distributions, as a percentage; and secondly the K-S statistic. We can see that DP1 is very different from any other Deferred Period, with huge values of the K-S statistic; also that "Other DPs" is also significantly different from the common ones, but in the other direction, as we can observe from the data. The common Deferred Periods are not so very far apart, and the statistics show that DP4, DP26 and DP52 are all not significantly different from each other, at a 5% level, because the K-S statistics are less than (or only just above) the 5% point of 1.36. DP13 is a bit more different. This distinction is apparent also if we look back to Table D2.1.3, which shows that the mean numbers of policies are close for DP4, DP26 and DP52, and a bit smaller for DP13, in spite of its long tail.

D3.2.4 If, in the analysis for DP4 and DP13 in Table 3.2.2, we were to omit the Singletons and to redo the calculations from two policies upwards, the resulting maximum absolute difference would be 1.11% (at two policies) and the K-S statistic would be 0.049, will within a reasonable probability for the Smirnov distribution. So the two distributions would not be significantly different. However, in deciding whether to analyse Singletons and Duplicates separately or together, it is the whole distribution that matters.

Table D3.2.2. Comparison of Deferred Periods with K-S test

| | | DP4 | DP13 | DP26 | DP52 | Other DPs |
|---|---|---|---|---|---|---|
| DP1 v | K-S distance | –44.30% | –48.00% | –45.33% | –45.63% | –54.33% |
| | K-S statistic | 55.43 | 60.34 | 55.45 | 44.78 | 22.87 |
| DP4 v | K-S distance | | –3.71% | –1.03% | –1.34% | –10.03% |
| | K-S statistic | | 5.01 | 1.35 | 1.37 | 4.25 |
| DP13 v | K-S distance | | | 2.67% | 2.37% | –6.32% |
| | K-S statistic | | | 3.52 | 2.43 | 2.68 |
| DP26 v | K-S distance | | | | –0.48% | –9.00% |
| | K-S statistic | | | | 0.49 | 3.81 |
| DP52 v | K-S distance | | | | | –8.69% |
| | K-S statistic | | | | | 3.56 |

## D3.3 *Males and Females*

D3.3.1 We next compare Males and Females. Overall there are 175,461 Male cumD cases, and 31,002 Female; these reduce to 123,145 and 26,881 exD cases respectively. The mean numbers of policies are 1.42 and 1.15. Of the Males, 26,628 of the exD cases, or 21.6%, have some Duplicates with a mean of 1.96 Duplicates each; of the Females, 2,835, or 10.5%, have Duplicates with a mean of 1.45 Duplicates each. The K-S statistic is a huge 16.45 with almost zero K-S probability. The distribution for Males wholly dominates that for Females. So there is no doubt that Males have many more Duplicates than Females. We summarise the numbers in Table D3.3.1.

Table D3.3.1. Comparison of Males and Females: statistics

| | Males | Females | Total |
|---|---|---|---|
| Number cumD | 175,461 | 31,002 | 206,463 |
| Number exD | 123,145 | 26,881 | 150,026 |
| Mean Policies | 1.42 | 1.15 | 1.38 |
| Number with $j \geq 2$ | 26,628 | 2,835 | 29,463 |
| Percent with Duplicates | 21.6 | 10.6 | 19.6 |
| Mean Duplicates | 1.96 | 1.45 | 1.92 |
| K-S distance | | –11.08% | |
| K-S statistic | | 16.45 | |

D3.3.2 The proportions of Males and Females differ between different Deferred Periods, so it is worth comparing the Sexes for each Deferred Period separately. The key statistics are shown in Table D3.3.2. We find that in each of the common Deferred Periods, Males have more Duplicates than Females, dominating absolutely in each case. The average numbers of Duplicates are much higher in DP1 than in the other common Deferred Periods, for both Sexes, but the other common Deferred Periods do not differ among themselves so much. The K-S test shows a very significant difference between the Sexes for DP1, and a smaller but still significant difference for the other common Deferred Periods, but not for the Other DPs, where the numbers are too small to show significance, in spite of the fact that the rather few Females have no Duplicates at all.

Table D3.3.2 Comparison of Males and Females: average numbers of policies, and K-S statistics for each Deferred Period separately.

| | | Mean Policies | | K-S | K-S |
|---|---|---|---|---|---|
| | Males | Females | Total | distance | statistic |
| DP1 | 2.37 | 1.53 | 2.27 | –24.96% | 13.52 |
| DP4 | 1.21 | 1.13 | 1.20 | –4.03% | 2.65 |
| DP13 | 1.15 | 1.07 | 1.14 | –3.77% | 2.65 |
| DP26 | 1.22 | 1.11 | 1.19 | –6.03% | 4.69 |
| DP52 | 1.23 | 1.08 | 1.19 | –8.32% | 4.60 |
| Other DPs | 1.03 | 1.00 | 1.03 | –3.29% | 0.49 |
| All | 1.42 | 1.15 | 1.38 | –11.08% | 16.45 |

## D3.4 *Year of Claim*

D3.4.1 We can compare individual Years of Claim, but in Table D3.4.1 we group the results by quadrennium. There is a small drift downwards by Year in the mean number of policies. However, although the proportion of cases with Duplicates reduces, the distribution of Duplicates among those cases stays much the same, as can be seen by considering the mean number of excess Duplicates. Because of this none of the quadrennia clearly dominates any of the others, though some of the K-S statistics are significant.

D3.4.2 However, when we look at Deferred Periods separately, we see a different result. Within each Deferred Period the average number of Duplicates, as shown in Table D3.4.2, has often been increasing, though by very little, and the K-S statistics are mixed, although sometimes not significant. The decrease overall is because of a reducing proportion of DP1 business among the Claims.

Table D3.4.1. Comparison of Quadrennia of Claim

| | | 1991-1994 | 1995-1998 | 1999-2002 | Total |
|---|---|---|---|---|---|
| Number cumD | | 64,613 | 66,945 | 74,905 | 206,463 |
| Number exD | | 45,319 | 47,487 | 57,220 | 150,026 |
| Mean Policies | | 1.43 | 1.41 | 1.31 | 1.38 |
| Number with $j \geq 2$ | | 10,134 | 10,219 | 9,110 | 29,463 |
| Percent with Duplicates | | 22.4 | 21.5 | 15.9 | 19.6 |
| Mean Duplicates | | 1.90 | 1.90 | 1.94 | 1.92 |
| 1991-94 v: | K-S distance | | 0.84% | 6.44% | |
| | K-S statistic | | 1.28 | 10.24 | |
| 1995-98 v: | K-S distance | | | 5.60% | |
| | K-S statistic | | | 9.02 | |

Table D3.4.2 Comparison of Quadrennia of Claim,
average numbers of policies for each Deferred Period separately.

| | 1991-1994 | 1995-1998 | 1999-2002 | Total |
|---|---|---|---|---|
| DP1 | 2.20 | 2.27 | 2.37 | 2.27 |
| DP4 | 1.17 | 1.22 | 1.21 | 1.20 |
| DP13 | 1.14 | 1.16 | 1.12 | 1.14 |
| DP26 | 1.22 | 1.22 | 1.17 | 1.19 |
| DP52 | 1.20 | 1.21 | 1.17 | 1.19 |
| Other DPs | 1.04 | 1.02 | 1.03 | 1.03 |
| All | 1.43 | 1.41 | 1.31 | 1.38 |

## D3.5 *Year of Entry, Year of Birth and Age*

D3.5.1 We can analyse by Year of Entry, which ranges from 1950 to 2002 with a few unspecified, but this analysis is complex. In any group of Duplicates, the Claim or policy with the earliest Entry Year is kept and the others are excluded. Those that have the earliest Entry Years have the largest average number of Duplicate policies, but the ratio of cumD to exD policies is close to 1.0, because all the Claims are retained. Later Entry Years show increasing ratios of cumD to exD policies, because Duplicates which had the first policy in an earlier Entry Year are rejected, but they show a reducing average number of Duplicates because, where the retained policy has that later Entry Year, it not surprisingly has fewer Duplicates following it. The rise in the average numbers of policies goes on till about 1980, and then falls, since those who first entered most recently have acquired fewer Duplicates. The obvious conclusion is that cases whose first Year of Entry is early have more Duplicates than those cases whose first Year of Entry is late.

D3.5.2 The same effect is seen, but more straightforwardly, when we look at Year of Birth. Not surprisingly, those with earlier Years of Birth have rather more Duplicates; younger policyholders have many fewer. However, when we look in detail, we see that those born before 1940 have on average a bit less than 2 Duplicates each, whereas those born from 1940 to 1952 or so have more than 2.0 each; this number falls, till those born after 1970 have hardly any Duplicates. The same is shown, though in reverse, when we look at Age, which is taken as age at Commencement of Sickness. Younger claimants have few Duplicates, older have many more.

## D3.6 *Occupation Class*

D3.6.1 In Table D3.6.1 we show results by Occupation Class. This is the Occupation Class for the Claim records, without the adjustment to make the Claim records correspond with the In force records. Note that “OC 5” indicates those with no Occupation Class given. We can see that Occupation Class 1 has many more Duplicates than the other Occupation Classes, dominating all of them. The others are all quite similar, and the K-S statistics are either not significant, or only marginally so.

D3.6.2 We could attribute the high weight of Duplicates in OC 1 to the fact that it is particularly heavy in DP1, and this indeed accounts for a lot of the difference. However, when we analyse the Occupation Classes within Deferred Periods, we see the same tendency, as shown in Table D3.6.2, except where there are too few cases, as in DP1, to see any trend. Almost everywhere OC 1 has many Duplicates, and the other Occupation Classes have few.

Table D3.6.1. Comparison of Occupation Class

| | | OC 1 | OC 2 | OC 3 | OC 4 | OC 5 | Total |
|---|---|---|---|---|---|---|---|
| Number cum D | | 142,146 | 19,798 | 20,799 | 13,591 | 10,129 | 206,463 |
| Number exD | | 89,675 | 18,380 | 19,324 | 12,968 | 9,679 | 150,026 |
| Mean Policies | | 1.59 | 1.08 | 1.08 | 1.05 | 1.05 | 1.38 |
| Number with $j \geq 2$ | | 26,466 | 977 | 1,152 | 519 | 349 | 29,463 |
| Percent with Duplicates | | 29.5 | 5.3 | 6.0 | 4.0 | 3.6 | 19.6 |
| Mean Duplicates | | 1.98 | 1.45 | 1.28 | 1.20 | 1.29 | 1.92 |
| OC 1 v: | K-S distance | | –24.20% | –23.55% | –25.51% | –25.91% | |
| | K-S statistic | | 29.89 | 29.70 | 27.15 | 24.21 | |
| OC 2 v: | K-S distance | | | 0.65% | –1.31% | 1.71% | |
| | K-S statistic | | | 0.63 | 1.15 | 1.36 | |
| OC 3 v: | K-S distance | | | | –1.96% | –2.36% | |
| | K-S statistic | | | | 1.73 | 1.89 | |
| OC 4 v: | K-S distance | | | | | –0.40% | |
| | K-S statistic | | | | | 0.30 | |

Table D3.6.2. Comparison of Occupation Class
average numbers of policies for each Deferred Period separately.
(Cells where there are fewer than 100 cumD cases are marked with an asterisk.)

| | OC1 | OC2 | OC3 | OC4 | OC5 | Total |
|---|---|---|---|---|---|---|
| DP1 | 2.27 | * | * | * | * | 2.27 |
| DP4 | 1.36 | 1.10 | 1.07 | 1.07 | 1.03 | 1.20 |
| DP13 | 1.25 | 1.08 | 1.04 | 1.04 | 1.03 | 1.14 |
| DP26 | 1.27 | 1.06 | 1.13 | 1.04 | 1.08 | 1.19 |
| DP52 | 1.25 | 1.05 | 1.15 | 1.03 | 1.02 | 1.19 |
| Other DPs | 1.05 | 1.05 | 1.02 | 1.00 | 1.01 | 1.03 |
| All | 1.59 | 1.08 | 1.08 | 1.05 | 1.05 | 1.38 |

## D3.7 *Dates of Sickness, Commencement of Claim and Cessation of Claim*

D3.7.1 It is not surprising that neither the Day nor the Month of Commencement of Sickness shows any difference in respect of Duplicates. However, the Year of Sickness does. This ranges from 1965 to 2002. The older Years show rather few Duplicates, the recent Years many more. But one must expect the recent Years to have many more short-term Sicknesses, which are much more common in DP1 business. Further, a Claim that commenced Sickness in 1965 and was still Sick at least until 1991 must have been relatively young at the start of Sickness, so could have accumulated few Duplicates prior to falling Sick.

D3.7.2 The Day and Month of Commencement of Claim likewise show no differences, with the exception that for cases where the Mode of Commencement is a Continuation from the previous Year there are fewer Duplicates. Again this is associated with DP1 business, where short-term Claims are more likely to recover within the Year, and therefore are not carried over to the next Investigation Year. The same is true for Day and Month of Cessation of Claim, and Mode of Cessation. Cases continued to the next Year show fewer Duplicates than those that recover within the Year.

### D3.8 *Cause of Sickness*

D3.8.1 There is considerable variation in the numbers of Duplicates for different Causes of Sickness, but this varies greatly with the Deferred Period, DP1 having many more Sicknesses with causes from which one may recover quickly. We do not show details.

## D4 CONCLUSION

D4.1 In this note we have investigated possible statistical distributions to describe the numbers of Duplicates in our file of Claims from 1991 to 2002. We use different Models to allow for starting position and different theoretical distributions. We find that in every case, a negative binomial distribution or a “diminished negative binomial distribution” (a distribution where there is no explicit probability of zero events), fits best, though not necessarily perfectly; the different Deferred Periods require different values of the parameters; and in some cases the best fit is an “improper diminished negative binomial” (a distribution where one of the parameters, which in a negative binomial distribution must be positive, is negative, but the probabilities of one or more events are still positive).

D4.2 We then investigated the differences between the cumulative numbers of Duplicates in many different categories, using a non-parametric method, the Kolmogorov-Smirnov test, and found that there were big differences between the distributions of Duplicates for Males and Females, for different Deferred Periods, for different Occupation Classes, and for different Ages, but not for any category by which we did not already subdivide the data. Duplicates are much more common in DP1 business, and therefore in categories associated with DP1 business. In addition, but not surprisingly, those who have effected their first policies a longer time ago, and are now older, have had more chance to add Duplicates than more recent, still young, entrants.

D4.3 All this justifies the practice of the CMI IP Committee in analysing policies which are Singletons or Duplicates together.

D4.4 However it is still better where possible to exclude (extra) Duplicates and use the exD subset of the Claims file rather than the cumD. Within the CMI IP analysis, Duplicates can be identified with reasonable confidence in the Claim records, and so the Claim Terminations analysis is based on the exD file. Duplicates cannot be identified in the In force data, and so the Claim Inceptions analyses must use the cumD files, but separate scaling allowances are made for the prevalence of Duplicates in each cell defined by the combination of Age, Year, Sex, Deferred Period and Occupation Class.