Continuous Mortality Investigation

Institute and Faculty of Actuaries

October 2018

Changes to CMI Research data 2018

1. Introduction

Much of the CMI's research activity is based on policyholder data that we seek from insurance companies for the Annuities, Assurances and Income Protection investigations and pension scheme member data that we seek from actuarial consultancies for the SAPS investigation. We refer to this as "Research data".

In 2017 / 2018, we reviewed our approach to Research data, ahead of the implementation of the General Data Protection Regulation (GDPR) in May 2018. This led to changes in the data we collect (and how we process it). In particular, we sought to reduce the likelihood of an individual being identifiable by:

- Moving to month and year of birth, instead of exact date of birth (and similarly for date of retirement);
- Ceasing to collect the actual benefit amount, where this exceeds a specified cap; and
- Ceasing to collect postcodes, which were previously an optional field for the life insurance investigations.

Views on our provisional changes were sought from a range of data contributors; the actual changes were then summarised in a letter sent to all data contributors – and (temporarily) added to the website – in April 2018. A copy of that letter is contained in the Appendix to this document.

The purpose of this document is:

1. To provide additional detail of the rationale behind some of the changes, in particular the change from seeking exact date of birth, which is set out in Section 2.
2. To set out our intended approach to the treatment of large benefit amounts (i.e. amounts above the "cap" introduced ahead of GDPR); in Section 3.
3. To document the first version of the "CMI Postcode Mapping Tool", which enables data contributors to provide a non-personal indicator of socio-economic status, in Section 4.

We would welcome feedback on the areas discussed in this document; please send this to info@cmilimited.co.uk.

2. Rounding of dates

Background

When reviewing our "Research data" requirements ahead of the implementation of the GDPR, we considered whether we need the exact date of birth for our analyses, or if a rounded version would be sufficient and allow us to collect data with less likelihood of individuals being identifiable.

This was not a new issue under the GDPR and some data contributors were already rounding dates of birth to the nearer month or six months, due to their own policies on personal data. Consequently, the Secretariat had already undertaken some analysis to ensure that the impact of this was small.

Changes to data requirements

As part of the CMI's wider review, the Secretariat undertook a more detailed analysis of the impact of rounded dates of birth on the accuracy of CMI analyses. The results showed:

- That rounding dates of birth, for periods of up to a year, had little impact on Actual/Expected (A/E) values when the population was of a reasonable size; and
- When rounding using periods greater than a year, we no longer had exposure data for all ages in all years. e.g. if our exposure period was from 1 July to 30 June then three-year rounding would mean we would only have exposure for one-third of ages in each year, and this could have unintended consequences.

The results suggested that rounding to a year could be acceptable but that rounding to a month offered higher accuracy, particularly for very small datasets. We considered that rounding to a month did not materially increase the risk of individuals being identifiable compared with rounding to a year.

We consulted on this approach with a sample of data contributors. Feedback suggested that for the SAPS investigation, date of retirement was likely to be closely related to date of birth, as individuals often retire on their birthday, so should be subject to the same level of rounding to prevent exact dates of birth being inadvertently revealed. We think this will be less true for pension annuities in payment, but we will be happy to receive rounded date of commencement, if data contributors consider it appropriate. We will also round this date for that investigation, regardless.

As a result, we have amended our data requirements to request:

- **Month and year of birth, instead of exact date of birth, for all investigations; and**
- **Month and year of retirement, instead of exact date of retirement, for the SAPS investigation.**

In each case, the CMI will use the 16th of the month in its analyses. If a data contributor is comfortable not applying such rounding in the data they submit, then the CMI will be happy to accept such data and we will apply the rounding at an early stage of the data processing.

We do not think that other dates such as date of expiry for income protection policies would make individuals more identifiable. We therefore request exact dates for these, but are happy to accept data submissions with this rounded as for date of birth. See the data requirements documents for each investigation, available on the CMI's pages of the IFoA website[1], for more details.

Rationale

We define "T-year rounding" as splitting dates of birth into buckets of consecutive T-year periods, and rounding each date of birth to the middle of the bucket in which it falls. e.g.

- 1/12-year rounding uses months and assigns dates to the 16th of each month.
- ½-year rounding uses six-year periods and assigns dates to 1 April or 1 October.
- 1-year rounding uses calendar years, and assigns dates to 1 July of each year.
- 3-year rounding uses …, 1900-1902, 1903-1905, … and assigns dates to 1 July 1901, 1 July 1904, etc.
- 5-year rounding uses …, 1900-1904, 1905-1909, … and assigns dates to 1 July 1902, 1 July 1907, etc.

[1] https://www.actuaries.org.uk/learn-and-develop/continuous-mortality-investigation/cmi-data

As mortality rates are roughly exponential, at least over small age ranges, we can model mortality as:

$$\mu_{x+t} = \mu_x \exp(\beta t) \qquad \text{for } t \in \left(-\frac{T}{2}, +\frac{T}{2}\right) \text{ within a T-year period.}$$

Assuming a uniform distribution of dates of birth within each T-year period, the impact of rounding is to treat mortality as being constant (i.e. μ_x) in this range. We then define R as the ratio of actual mortality μ_{x+t} to assumed mortality, μ_x; so $R = \exp(\beta t)$.

Now

$$E[R] = \frac{1}{T}\int_{-T/2}^{+T/2} \exp(\beta t)\, dt = \frac{1}{\beta T}\left(\exp\left(\frac{\beta T}{2}\right) - \exp\left(-\frac{\beta T}{2}\right)\right)$$

Also

$$E[R^2] = \frac{1}{T}\int_{-T/2}^{+T/2} \exp(2\beta t)\, dt = \frac{1}{2\beta T}\left(\exp(\beta T) - \exp(\beta T)\right)$$

From this we can calculate variance as $V[R] = E[R^2] - E[R]^2$

The table below shows a 95% confidence interval for $\frac{1}{R}$ for a range of values of T and population sizes. We show $\frac{1}{R}$ (comparable with Actual/Expected calculations) and use a value of $\beta = 0.11$, which is plausible for pensioner populations. Note that the figures in Table 2.1 are based on the idealised assumption that we round to the exact midpoint of each period. In practice this cannot be done exactly as for periods with an even number of days no date falls at the exact midpoint.

Table 2.1: 95% confidence intervals for the ratio of actual mortality to assumed mortality by rounding period and population size

Population	Month	Quarter	Half-year	1-year	3 years	5 years
1	99.5%-100.5%	98.5%-101.6%	97.0%-103.2%	94.1%-106.6%	83.9%-122.4%	75.4%-143.2%
10	99.8%-100.2%	99.5%-100.5%	99.0%-101.0%	98.0%-102.0%	94.0%-105.8%	89.9%-109.5%
100	99.9%-100.1%	99.8%-100.2%	99.7%-100.3%	99.3%-100.6%	97.7%-101.4%	95.8%-101.9%
1,000	100.0%-100.0%	99.9%-100.0%	99.9%-100.1%	99.8%-100.1%	99.0%-100.1%	97.8%-99.7%
10,000	100.0%-100.0%	100.0%-100.0%	100.0%-100.0%	99.9%-100.0%	99.4%-99.7%	98.4%-99.1%
100,000	100.0%-100.0%	100.0%-100.0%	100.0%-100.0%	99.9%-100.0%	99.5%-99.6%	98.7%-98.8%
1,000,000	100.0%-100.0%	100.0%-100.0%	100.0%-100.0%	99.9%-100.0%	99.5%-99.6%	98.7%-98.8%

We can see from Table 2.1 that the variation in Actual/Expected has two components to it:

- A part that diversifies with the size of the population; and
- A systematic part that remains no matter how large the population is.

For 5-year rounding, $E\left[\frac{1}{R}\right] = 0.9875$, so there is a bias of 1.25%, no matter how large the population is. This is broadly consistent in size and direction with empirical results observed when we tested the impact of rounding on a real (large) dataset.

For monthly rounding, $E\left[\frac{1}{R}\right] = 0.9999965$ under the idealised assumption above and $E\left[\frac{1}{R}\right] = 1.000075$ once we allow for the actual distribution of month lengths and round to the 16th of each month, so there is negligible bias for a large population.

A Taylor expansion – not shown here – can be used to show that the bias varies with the square of T. It also shows that if mortality were linear (i.e. $\beta = 0$) there would be no bias at all from rounding.

Duplicates

The analysis above focused on the use of date of birth for calculating age. It did not consider its use for other purposes. In particular date of birth can be used, in conjunction with other data fields, for:

- Matching records between data submissions; and
- Identifying duplicates within data submissions.

In the first example, to date this has largely been confined to processing submissions to the Annuities and Assurances investigations. Date of birth can help match records, as part of our verification process, where the standard identifiers (policy, benefit and client id) are not available. It is not strictly necessary and the impact of not having this option available can be mitigated by encouraging data providers to include unique identifiers.

The CMI's approach to dealing with duplicates differs by investigation:

- The SAPS investigation, which uses a prescribed data format with relatively few data fields, currently removes records that are determined as duplicates by a match on all data fields, including date of birth. Collecting only month and year of birth will increase the risk of records being incorrectly identified as duplicates, particularly in large submissions although:
  - This would not affect amounts-weighted results; and
  - It should not introduce any bias into lives-weighted results.
- The Annuities and Assurances investigations do not attempt to systematically remove duplicates, but seek to combine 'increments' with the parent policy. The number of fields available to match on, should mean that using a rounded date of birth does not materially impact our ability to do this accurately.
- The Income Protection investigation de-duplicates data as part of its analysis methodology, but rounded dates of birth do not present a new problem as data is already collected and processed in this form.

3. Treatment of large benefit amounts

Background

The review of our "Research data" requirements identified that data records with very large benefit amounts carried a potentially increased risk of individuals being identified. To reduce this risk, we have specified caps – detailed below – for each investigation, above which exact benefit amounts will not be requested.

Capping benefit amounts will have an impact on amounts-weighted analyses, so we have investigated how best to deal with this.

Changes to data requirements

We believe that the potentially increased risk of individuals being identified arises from only the very largest benefits so the caps we have chosen for each investigation, shown in Table 3.1, are designed to capture only a very small proportion of data records. As for date of birth, if a data contributor is comfortable submitting uncapped amounts, then we will apply this cap ourselves.

Table 3.1 also shows the amounts that we will use initially in place of the capped amounts in our analyses. These were derived pragmatically as described in the section below.

Table 3.1: Capped amounts and assumed replacements for each CMI investigation

Investigation	Capped amount	Replacement amount
SAPS	£100,000 pa	£160,000 pa
Annuities	£100,000 pa	£160,000 pa
Assurances (death benefits)	£1,000,000	£2,200,000
Assurances (critical illness)	£1,000,000	£1,700,000
Income protection	£75,000 pa	n/a (see below)

Rationale

In deriving assumed amounts to use in place of capped amounts for each investigation, we have adopted an overriding principle of simplicity with a secondary aim of minimising the impact on amounts-weighted Actual/Expected (A/E) values. However, we also recognised that – irrespective of data protection considerations – capping very large amounts is arguably a preferable approach as it reduces the volatility in amounts-weighted results and, by definition, individuals with high benefits who have already died cannot die again in the future.

SAPS investigation

We performed an analysis of amounts larger than the threshold for capping (£100,000 pa) in the 2009-2016 dataset. The average pension amounts for such cases were approximately £170,000 pa for males and £150,000 pa for females. We also calculated average benefit amounts for each five-year age band but found there was limited variation at most ages. As the results were fairly insensitive to changes in the assumed pension amount (as described below), we considered it spurious to introduce this additional complexity.

We used these figures in place of the actual pension amount to assess the impact on the amounts-weighted 100A/E as follows:

- For each group, we calculated an amounts-weighted A/E for the age range 60-95 using the (uncapped) dataset with:
  - Actual deaths over the period 2009-2016; and
  - Expected deaths based on the proposed S3PMA and S3PFA tables.

- We did the same for the subset of pensioners in that group that had pensions over the £100,000 pa cap, on both a lives and amounts basis.
- We adjusted the actual and expected deaths for this subset by multiplying the lives-weighted figures by the assumed pension amount to get these on an amounts basis with the cap applied.
- We used these adjusted figures to calculate a new amounts-weighted A/E for the overall group and compare with the original results.

For male pensioners, using an assumed pension amount of £170,000 pa for capped records gave the same overall value of 100A/E (to one decimal place) as the original dataset. This was also true when using an assumed pension amount of £150,000 pa for female pensioners. Changing the assumed amount by £10,000 pa had very little impact on 100A/E values (at most 0.1) so the results were fairly insensitive to this assumption, leading us to conclude that it would be reasonable to assume the same average amount of £160,000 pa for both males and females.

We checked how the results varied for subgroups of the SAPS dataset when pensions were capped and replaced by assumed amounts of £160,000 pa. For Normal health pensioners, and Ill-health pensioners, the 100A/Es differed by at most 0.1. For the Very Light bands (which are new for S3 and are made up of pensioners with pensions over £40,000 pa (males) and £16,000 pa (females)), the 100A/Es differed by 0.1 for females and 0.2 for males.

When the results were broken down by industry, 100A/Es differed by at most 0.1 for females but the impact was larger for males in some cases (up 0.9). However, except for one industry, these differences were smaller than the corresponding lives-weighted standard deviations. As confidence intervals are wider when weighted by amounts rather than lives, we can infer that the differences are all within the 95% confidence intervals for their respective industries; i.e. the impact on 100A/E of using the proposed assumed pension amounts is materially lower than the uncertainty due to the volume of available data.

When the results were broken down by year, 100A/Es differed by at most 0.3 over the period 2009-2016. We do not propose varying these amounts by year for the following reasons:

- The average pension largely depends on which schemes have submitted data. When the data is broken down by year the average pension is volatile and does not show a steady increase.
- Although pensions for members already above the cap would typically increase, there would also be new records joining this subset (e.g. a pension of £100,001 pa) which lower the average.
- Over time, while the average salary in the pension scheme population might increase, the average service is likely to fall due to the cessation of accrual and closure of pension schemes. It is not clear whether the overall impact would be to increase or decrease pension amounts.

Annuities investigation

We performed a similar analysis on the Annuities investigation's 2011-14 dataset for pension annuities in payment. We found that the assumed amount that would leave the overall 100A/E unchanged was £154,000 pa. Interestingly, the figure for males was slightly lower (£151,000 pa) and for females it was higher (£166,000 pa). This supported using a single figure for both genders; applying a figure of £160,000 pa, consistent with the SAPS investigation, changed the overall 100A/E by 0.1.

We found that the assumed amount required to leave the A/E unchanged varied little for the product categories and retirement types with significant volumes of data, so it is reasonable to use a single figure in all cases.

Assurances investigation

In a similar analysis on the Assurance's investigation's 2011-14 term assurances dataset, we considered the three benefit types of death benefits (DB), accelerated critical illness (ACI) and stand-alone critical illness (SCI) separately. We found that the assumed amounts that would leave the overall 100A/Es unchanged were £2.2m for DB, £1.7m for ACI and £1.8m for SCI.

There were some differences in these values by gender, by product type and by distribution channel. However, we did not consider these material. The ACI and SCI figures were not significantly different and pointed towards using the same figure of £1.7m for both. Applying this figure to the DB dataset increased its 100A/E by 0.6, compared with a lives-weighted standard deviation of 0.5, justifying the use of a different figure for DB data.

Income Protection investigation

The Income Protection investigation has never produced amounts-weighted experience results. Working Paper 102 included results by benefit amount band, but these bands are not currently included in standard analyses. That paper also noted limitations with the existing data, as the additional factors had not been used in “all offices” results, and so the checks performed on those fields during data processing were not as stringent as on some other data fields. In the 2003-10 dataset, the relatively small number of claim records with amounts larger than £75,000 pa had a surprisingly high average size of £325,000 pa, perhaps because data had been supplied with annual amounts treated as weekly or monthly.

As a result, and as it is not required in current work, we have deferred deriving a replacement amount to use where benefit amounts are capped. If the CMI seeks to report income protection experience by amounts in future we will need to consider our approach, and appropriate data checks, at that time.

Future updates and secondary investigations

We do not propose updating the assumed amounts used to replace capped amounts on an annual basis but they will need reviewing in future from time to time. As data collected in future will be capped, the CMI will not be able to perform its own analyses to determine appropriate updated amounts to use. Our current intention is to occasionally seek information from data contributors about the average size of benefit amounts above the cap to assess whether the figures need to change.

We have also not yet determined appropriate assumed amounts to use in ‘secondary’ investigations such as enhanced annuities, life annuities or Whole life assurances. We intend to adopt a similar approach of asking data contributors to advise their average amounts (for benefits above the cap) when we collect data for these investigations.

4. Analysis of experience by socio-economic status

Full postcode, intended to allow socio-economic analyses, had previously been withdrawn from SAPS data requirements as few data contributors were submitting it, but it remained an optional field in our data requirements for life insurers. However, the option for life insurers to supply a number of alternatives and the lack of consistency between these made it difficult to produce combined results and, in any event, we considered full postcode to be inappropriate under the GDPR, as it would increase the likelihood of an individual being identifiable. After seeking views from a sample of data contributors, we have instead decided to collect two measures based on the Index of Multiple Deprivation (IMD), a population-based statistic described below, for each record.

The first version of the CMI Postcode Mapping Tool (2018 v01) is available to CMI Subscribers alongside this document on the CMI pages of the IFoA website. Data contributors can use it to provide a measure of socio-economic status for each life in their submissions without the need to submit full postcodes. Once we have a meaningful volume of data, this will allow the CMI to perform analyses of mortality and/or morbidity experience by socio-economic status for any of our investigations.

The mapping tool enables socio-economic indicators to be determined from an individual's postcode via mapping to a census data area for which IMD measures are calculated. The indicators in the tool are IMD deciles, which we consider to be sufficiently broad to avoid any risk that the measure could be reverse-engineered to identify the individual.

Method for collecting IMD-based data

The mapping tool consists of a text file that provides a mapping from postcode to two measures of socio-economic status, both based on the IMD produced for each of the nations of the UK. Specifically, these are:

- Deciles within each of the four nations' own measures of IMD, combined with an indicator for the 12 'NUTS 1' regions of the UK – Northern Ireland, Scotland, Wales and nine English regions[2] – plus the Channel Islands and Isle of Man; and
- Deciles within the UK, based on a UK-wide measure of IMD.

Note that this approach differs from that on which we sought views from a sample of data contributors. At that stage, we envisaged mapping postcodes to semi-deciles, which would give greater flexibility than deciles but are sufficiently broad to avoid any risk that the IMD measure could be reverse-engineered to identify the individual. Our revised approach incorporates a regional indicator – which we think has the potential to add value to our analyses – but uses deciles – to avoid overly narrow data segmentation.

UK-wide and nation-specific IMDs

Each of the four nations of the UK produces its own IMD. These are used primarily by local governments/ authorities to target funding at the most deprived areas. However, they are also widely used in other fields to provide an indicator of socio-economic status, and the ONS has published population mortality data by IMD deciles[3].

The construction of each country's IMD is different but follows a similar approach, based on a series of seven or eight 'domains', of which income and employment levels carry the most weight (currently 45%-56%). IMD scores are published for geographical areas used in national censuses. The absolute values of these scores have little meaning, but they are used to rank the areas according to their level of deprivation. The CMI Postcode Mapping Tool includes deciles for every postcode, with 1 representing the most deprived and 10 the least deprived (the order commonly used by the ONS and others).

Note: As the IMD scores are calculated independently for each nation, the countries' deciles are not comparable. For example, England decile 2 does not necessarily imply the same level of deprivation as Scotland decile 2.

[2] NUTS (https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics) is an EU standard. The English regions are: South West, South East, London, East of England, West Midlands, East Midlands, Yorkshire and the Humber, North West, and North East.

[3] For example, population and deaths by single year of age and IMD decile for England and Wales, 2001-2016: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/007710numberofdeathsandpopulationsindeprivationdecileareasbysexandsingleyearofageenglandandwalesregisteredyears2001to2016

Each country's IMD tends to be updated every few years, at a different time. The latest version of each[4] is listed in Table 4.1, together with the version included in the UK-wide adjusted IMD (described below).

Table 4.1: Latest IMD for each UK country and the version included in the UK-wide adjusted IMD

Country	Latest IMD version	UK-wide adjusted IMD version
England	2015	2015
Northern Ireland	2017	2010
Scotland	2016	2012
Wales	2014	2014

Nation-specific measures of IMD

The underlying motive for capturing IMD is to allow us to analyse mortality (and morbidity) experience for 'similar' groups of lives. By mapping postcodes to deciles for each country separately, we might then find that an appropriate grouping for the lightest mortality/morbidity rates comprises (say) English deciles 8-10, Northern Irish decile 10, Scottish deciles 8-10 and Welsh deciles 9 and 10.

Given the disparity in size of the four countries, we have chosen to incorporate a regional indicator in the mapping file, to further sub-divide England. This uses the 12 'NUTS 1' regions of the UK and could allow us to explore regional differentials in English insured and pensioner mortality if data volumes permit. For the avoidance of doubt, the IMD deciles for England are national measures, not regional measures – e.g. in population data, we might expect a greater weighting to less-deprived deciles in the South East than in the North East.

UK-wide measure of IMD

A method for standardising IMD scores of the four countries, based on the underlying elements for income and employment, was published by Abel, Barclay and Payne (BMJ, 2016). This updated previous work by two of these authors, published by the Office for National Statistics (Health Statistics Quarterly 53, 2012). Using this method allows us to collect consistent data for the UK as a whole. However, it does have some drawbacks. In particular, it relies on:

- The IMD calculation methodologies being similar for each country; and
- The background levels of each domain (e.g. employment rates) being stable over time.

The first of these appeared to be a reasonable assumption for the latest available indices at the time of Abel et al's latest report, shown in Table 4.1.

However, it is not necessarily the case that future versions of each country's index will retain similar methodologies. Indeed, the latest version of Northern Ireland's index (2017, known as the Multiple Deprivation Measure) introduced a different definition for its income domain, based on median incomes in Northern Ireland alone.

The second assumption, that the background levels of each domain are stable, is also far from certain, particularly given economic conditions over the past decade. This may explain, at least in part, why updating Abel et al's paper to include the latest version of Scotland's IMD (2016) results in a materially higher proportion

[4] The latest versions of each country's IMD are available at the following locations:
England: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015
Northern Ireland: https://www.nisra.gov.uk/statistics/deprivation/northern-ireland-multiple-deprivation-measure-2017-nimdm2017
Scotland: http://www.gov.scot/Topics/Statistics/SIMD
Wales: https://gov.wales/statistics-and-research/welsh-index-multiple-deprivation/?lang=en

of the Scottish population in the least deprived quintile (26% of data zones, compared with 20% previously) of the UK population as a whole.

As a result, and the risk of similar future changes, we have chosen to base our UK-wide measure on the indices shown in Table 4.1 (including corrections to the Welsh IMD data made after Abel et al's work was completed), and not more recent versions. We do not currently intend updating this measure in future, as noted on the next page, unless/until an updated UK-wide measure becomes available.

We consider that the different measures offer their own benefits. In summary:

- The UK-wide measure, using normalised data:
  - offers a simpler starting point for analysing mortality experience and building models.
  - may be useful to Subscribers as a 'standard' normalised UK-wide IMD dataset that they can use for their own analysis.
- However the nation-specific measure:
  - avoids spurious accuracy that could arise from the published method of adjustment.
  - can be updated for each country separately when its IMD is updated.
  - allows more direct comparison with population mortality experience for England and Wales, at least, as the ONS has released exposure and deaths data by IMD decile. This is not (currently) possible on a UK-wide basis.
  - combined with the 'NUTS 1' regions, might enable the CMI to explore regional differentials in English insured and pensioner mortality if data volumes permit.

Postcodes

The postcodes included in the CMI Postcode Mapping Tool are those listed in the Office for National Statistics Postcode Directory (ONSPD)[5]. At August 2018 there were around 2.6m postcodes in the directory. Not all of these relate to UK residential addresses or remain in use but we have included them all in the tool for completeness. The ONSPD is updated quarterly.

The ONSPD includes postcodes in three different formats:

i. 7-character fixed length, where the third and fourth characters may be blank

ii. 8-character fixed length, where the fifth character is always blank and the third and fourth characters may be blank.

iii. Up to 8-character variable length, with a single space between the two 'halves', known as inward and outward codes.

For ease of use, we have included each of these in the tool, e.g.:

- "AB1 0AA", "AB1 0AA", "AB1 0AA"; and
- "B1 1AA", "B1   1AA", "B1 1AA".

How do we get from postcode to IMD decile?

IMDs are calculated and published for each:

- Lower-layer super output area (LSOA, comprising around 1,500 people on average) in England and Wales;
- Super output area (SOA, around 2,000 people) in Northern Ireland; and
- Data zone (around 750 people) in Scotland.

The ONSPD provides the applicable output area for each postcode, allowing for a straightforward map from postcode to IMD data. The nine 'NUTS 1' regions of England are similarly obtained.

Some postcodes do not have an assigned census output area, for example because the postcode does not apply to a geographical location or because it is no longer in use. These are relatively few in number (less than

[5] The ONSPD as at August 2018 is available from:
https://ons.maps.arcgis.com/home/item.html?id=870f2b5e0fb1441ab38afd145214e9f4

1%); we have retained them in our mapping file, with an 'unknown' IMD decile. They will thus be treated in the same way as unknown, incomplete or non-existent postcodes.

Updating the mapping tool

Our current thinking is that we will issue an updated version of the CMI Postcode Mapping Tool on an annual basis. Our plan is for this to be early in Q1, using the November version of the ONSPD, ready for that year's data submissions. In particular, this would incorporate postcodes that have been introduced or altered during the year, and also allow for any updated nation-specific IMDs. We currently do not intend updating the UK-wide measure unless/until an updated UK-wide measure becomes available.

We will generally be happy to accept data mapped using an earlier version of the tool; in particular if the only change is to the ONSPD. We will consider the materiality of changes in IMD classification when determining whether we can accept data mapped using versions based on IMDs that have been superseded and will advise data contributors accordingly when we issue the updated mapping tool.

Using the mapping tool

The tool is intended to be very simple for data contributors (and other CMI Subscribers) to use. It is provided in a comma-separated value (CSV) format that is widely used and compatible with a wide range of systems. We envisage that data contributors will use it as a 'look-up' tool.

The fields contained in the tool are described in Table 4.2 (Note: the three postcode fields are named consistently with the ONSPD). By using the postcode data field(s) that correspond to the format used in their own systems, data contributors can add three additional fields to their data submissions – nation-specific IMD decile, region and UK-wide IMD decile.

Table 4.2: CMI Postcode Mapping Tool data format

Field	Description
PCD	Postcode; 7-character fixed length, where the third and fourth characters may be blank
PCD2	Postcode; 8-character fixed length, where the fifth character is always blank and the third and fourth characters may be blank.
PCDS	Postcode; up to 8-character variable length, with a single space between the two 'halves', known as inward and outward codes.
NS_Decile	Nation-specific IMD decile, where 1 is most deprived and 10 is least deprived.
Region	'NUTS 1' region: Northern Ireland, Scotland, Wales and nine English regions.
UK_Decile	UK-wide IMD decile, where 1 is most deprived and 10 is least deprived.

The first few records from the mapping tool are shown below:

PCD	PCD2	PCDS	NS_Decile	Region	UK_Decile
AB1 0AA	AB1 0AA	AB1 0AA	10	SC	9
AB1 0AB	AB1 0AB	AB1 0AB	10	SC	9
AB1 0AD	AB1 0AD	AB1 0AD	10	SC	9
AB1 0AE	AB1 0AE	AB1 0AE	8	SC	8

Producing results

We recognise that there will be some movement in postcodes being mapped to each decile as countries update their IMD versions.

Consequently aggregate results for any single year may be based on data mapped using more than one version of the tool; however we expect the overall impact of such changes on the analysis to be limited from year to year.

Appendix: Copy of letter sent to Data Contributors in April 2018

April 2018

CMI Research data – our intentions in relation to the GDPR

Background

Much of the CMI's research activity is based on policyholder data that we seek from insurance companies for the Annuities, Assurances and Income Protection investigations and pension scheme member data that we seek from actuarial consultancies for the SAPS investigation. We refer to this as "Research data".

We have reviewed our approach to Research data, ahead of the implementation of the General Data Protection Regulation (GDPR) in May 2018; this document sets out our intentions.

Is CMI Research data covered by the GDPR?

Under the GDPR, *"'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."*

As we do not ask for the name, address or National Insurance number in Research data, the vast majority of records within a CMI dataset could not be related to a particular individual. However, a few may and we are therefore:

1. Amending the data we seek, to minimise this likelihood, and
2. Continuing to treat all Research data as if it were personal data from a legal and data security perspective.

Lawful basis for processing personal data

Under the GDPR, data controllers need to:

- identify the lawful basis for processing personal data,
- document this, and
- advise individuals of this in their privacy notice.

The GDPR also states that: *"The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. … further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes."*

We have been advised that the CMI's Research data should fall within this definition; consequently it may not be ***necessary*** for data controllers to explicitly mention in their privacy notice that information may be used for research purposes. However – for transparency – it may be desirable to do so, if there is potential that the data being submitted to CMI could be considered personal data.

The following wording may be suitable for life insurers submitting policyholder data:

> *We may also use your personal data in research into the mortality and morbidity experience of policyholders in general. This may include the provision of data, anonymised as far as possible, to a recognised external authority, such as the Continuous Mortality Investigation (CMI) (a subsidiary of the Institute and Faculty of Actuaries), which analyses experience on our behalf.*

The following wording may be suitable in the case of pension scheme members:

> *The Scheme Actuary may also use your personal data in research which assists actuaries in providing actuarial advice to pension schemes – for example research into the mortality experience (life expectancy) of pension scheme members in general. This may include the provision of data, anonymised as far as*

possible, to a recognised external authority, such as the Continuous Mortality Investigation (CMI) (a subsidiary of the Institute and Faculty of Actuaries), which analyses mortality experience on our behalf.

CMI's status under the GDPR

Under the CMI's current Terms & Conditions for Data Contributors, to the extent that Research Data submitted to the CMI includes any personal data, the CMI will act as a data processor on behalf of the data contributor. We have been advised that – with some minor modifications to current practices – this position remains valid under the GDPR; we therefore intend to reflect this positioning in our updated Terms & Conditions.

Going forward, as a processor, the CMI will use such personal data only for the purposes of the insurer- or scheme-specific analysis carried out for that data contributor. Where the CMI undertakes further research and analysis to prepare industry-wide analysis, it will do that only with anonymised and aggregated data created in the course of the insurer- or scheme-specific research. The GDPR does not therefore apply to such further research activities.

Data minimisation

We have reviewed a number of aspects of our data requirements. As an example, our discussion of the need for the exact date of birth (DoB) concluded that:

a) In some cases, exact DoB does increase the chances of an individual becoming identifiable.

b) With regard to accuracy:
   - Only collecting the month and year of birth is adequate; it is also easy for data contributors to implement.
   - Data with quarterly-or half-yearly rounding would still be acceptable.
   - Less precise dates – such as annual dates – unacceptably reduce the accuracy of the analyses.

Consequently, we are amending our data requirements for all investigations to encourage data contributors to submit only month and year of birth. We will continue to accept and process data submissions containing exact values, but will only retain and use month and year. If data contributors wish to round dates of birth to the nearest quarter- or half-year, we will also accept and process such data however we will not process any data that is rounded to annual dates.

We also thought it could be helpful to share our thinking on certain other aspects of the data requirements;

1. **Date of retirement**. As retirement often occurs on the 60th or 65th birthday, use of the exact date could inadvertently reveal the exact DoB. We will therefore treat this consistently with DoB for both Annuities and SAPS data.
2. **Other dates**. We also reviewed the need for exact date of death (DoD) and agreed that we should continue to request this; as the rounding of both DoB and DoD would make age at death less accurate. In addition, it was not obvious that using a rounded DoD would significantly impact on the potential identifiability of individuals. We do not think that other exact dates, such as date of exit, would be well-known, so they should not make a person more identifiable. We therefore see no need to change our current practice of seeking exact dates here.
3. **Very high benefit amounts**. As with exact DoB, we recognise that very high amounts potentially increase the chance of certain people becoming identifiable. In addition, we currently make limited use of the exact amount, as (to date) our analyses have grouped data into relatively wide amounts bands.

   However there is no simple pragmatic approach, akin to collecting only the month of birth, as capping extreme amounts would reduce the weighting of this group in amounts-weighted analyses and the use of an average would be complicated as this could be insurer- or scheme-specific, vary with age and change over time.

   For SAPS, we intend to encourage data contributors to provide a figure of £999,999 for any record with an annual amount over £100,000 and we will then re-weight these values in our analyses. If (some) data contributors are comfortable supplying the exact amount then we will treat that consistently with other large amount data in "all schemes" analyses.

   We intend using the same value for Annuities, for consistency, and values of £1,000,000 for Assurances and £75,000pa for Income Protection.

4. **Very high ages**. Although there is an increased chance of identifiability at very high ages – say 105+ – due to the low numbers of lives at such ages in the population we intend to continue to request such data as this is an area of increasing relevance to longevity risk.

Socio-economic indicators

CMI has been keen to expand the quality of its analyses by collecting some form of socio-economic indicator. Currently:

- Full postcode is an optional field in our data requirements for life insurers for this purpose, but so too are several alternatives and the lack of consistency between these has made it difficult to produce combined results.
- We do not seek any such indicator for SAPS. Previously we sought full postcode but we withdrew this when it became apparent that too few data contributors were submitting this.

We consider full postcode to be inappropriate under the GDPR and have sought views from a sample of data contributors on a proposal that the CMI should instead collect a measure based on the Index of Multiple Deprivation (IMD) for each record. We are pleased that this was widely-supported.

We therefore intend to make available a tool that will enable the value to be determined from an individual's postcode, via mapping to a census data area[6] for which IMD measures are calculated[7]. Our intention is that we will map to semi-deciles, which give greater flexibility than deciles but are sufficiently broad to avoid any risk that the IMD measure could be reverse-engineered to identify the individual.

Next steps

This document sets out our current intentions.

We intend to issue revised Terms & Conditions, Data Handling Protocols and data requirements documents later this month. We will also issue an IMD mapping, and supporting documentation, as soon as we can.

[6] For example, lower-layer super output areas (LSOAs) in England and Wales.
[7] For example, the latest data for England is available here: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015. As the four constituent countries of the UK each produce their own IMDs using different methodologies, they are not directly comparable. We are currently considering using the method for standardising scores, based on the underlying elements for income and employment, published by Abel, Barclay and Payne (BMJ, 2016) in a CMI mapping tool to collect consistent data for the UK.