

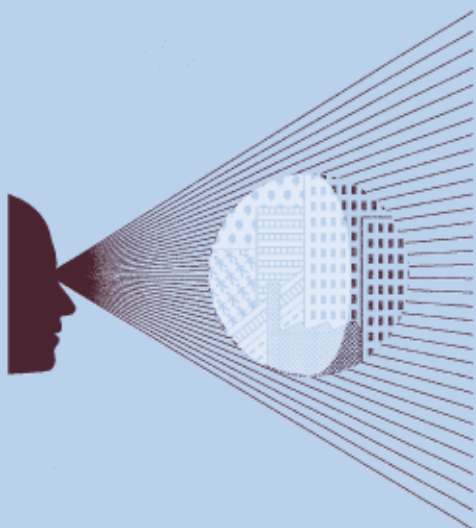
# Feasibility study for potential econometric assessment of the impact of R&D tax credits on R&D expenditure

Matching R&D credit data

Prepared for  
HM Revenue & Customs

September 2006

HM Revenue & Customs Research Report 19



© Crown copyright 2006.

Copyright in the typographical arrangement and design rests with the crown. This publication may be reproduced free of charge in any format or medium provided that it is reproduced accurately and not used in a misleading context. The material must be acknowledged as Crown copyright with the title and source of the publication specified.

The views expressed in this report are those of the authors and do not necessarily represent those of HM Revenue & Customs.

Published by HM Revenue & Customs.

Oxera Consulting Ltd is registered in England No. 2589629. Registered office at Park Central, 40/41 Park End Street, Oxford OX1 1JD, UK. Although every effort has been made to ensure the accuracy of the material and the integrity of the analysis presented herein, the Company accepts no liability for any actions taken on the basis of its contents.

Oxera Consulting Ltd is not licensed in the conduct of investment business as defined in the Financial Services and Markets Act 2000. Anyone considering a specific investment should consult their own broker or other investment adviser. The Company accepts no liability for any specific investment decision, which must be at the investor's own risk.

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                             | <b>1</b>  |
| <b>2</b> | <b>Description of the data</b>                  | <b>2</b>  |
| 2.1      | HMRC's company tax returns                      | 2         |
| 2.2      | ONS Business Enterprise R&D survey              | 3         |
| <b>3</b> | <b>Matching procedure</b>                       | <b>7</b>  |
| 3.1      | Potential identifiers for the matching          | 7         |
| 3.2      | The selection of the unique identifier          | 10        |
| 3.3      | The potential for matching                      | 11        |
| 3.4      | Matching by CRN                                 | 13        |
| 3.5      | Matching by company name                        | 14        |
| 3.6      | Caveat of the matching procedure                | 15        |
| <b>4</b> | <b>Results of the matching</b>                  | <b>16</b> |
| 4.1      | Matching based on the modified CRN              | 16        |
| 4.2      | Matching based on the company name              | 17        |
| 4.3      | Results of the matching by year                 | 18        |
| <b>5</b> | <b>Analysis of the matched data</b>             | <b>20</b> |
| 5.1      | Availability of matched observations            | 20        |
| 5.2      | Characteristics of the matched observations     | 24        |
| 5.3      | R&D expenditure across both datasets            | 26        |
| <b>6</b> | <b>Conclusions</b>                              | <b>28</b> |
|          | <b>Appendix 1 Definitions of product groups</b> | <b>30</b> |

## List of tables

|            |  |    |
|------------|--|----|
| Table 2.1  | Number of CRNs in the current CT600 extract, by year   | 3  |
| Table 2.2  | Number of CRNs in the BERD database each year          | 5  |
| Table 2.3  | The BERD sample over time                              | 6  |
| Table 3.1  | Potential identifiers for matching                     | 8  |
| Table 3.2  | Potential identifiers for matching—short form          | 9  |
| Table 3.3  | Potential identifiers for matching—long form           | 9  |
| Table 3.4  | Availability of data over time, as measured by the CRN | 11 |
| Table 3.5  | Results of the matching based on the CRN               | 13 |
| Table 3.6  | Cleaning of the company name—illustrative examples     | 14 |
| Table 4.1  | Results of matching based on the CRN                   | 16 |
| Table 4.2  | Results of matching based on the company name          | 18 |
| Table 4.3  | Results of matching, by year                           | 18 |
| Table 5.1  | Distribution of matched observations over time         | 20 |
| Table 5.2  | Frequency of matched observations                      | 23 |
| Table 5.3  | Matched observations—broad SIC (92) product groups (%) | 24 |
| Table A1.1 | Definitions of product groups                          | 30 |

## List of figures

|            |   |    |
|------------|---|----|
| Figure 3.1 | Availability of data over time, as measured by the CRN  | 12 |
| Figure 3.2 | Length of CRNs in CT600 and BERD, as a percentage of the total (%)  | 13 |
| Figure 5.1 | Distribution of matched short-form observations over time   | 21 |
| Figure 5.2 | Distribution of matched long-form observations over time  | 22 |
| Figure 5.3 | Distribution of matched observations over time  | 23 |
| Figure 5.4 | Turnover of the matched companies, relative to the CT600 population (% of the total number of matched companies each year)        | 25 |
| Figure 5.5 | R&D expenditure of the matched companies, relative to the CT600 population (% of the total number of matched companies each year) | 26 |
| Figure 5.6 | Comparisons of R&D expenditure across BERD and the CT600 extract  | 27 |

# 1 Introduction

Oxera has matched HM Revenue and Customs' (HMRC) extract from the company tax returns (CT600) with both the short and long forms from the ONS' Business Enterprise R&D survey (BERD) over the period between 2000 and 2004. The CT600 extract contains information from companies' returns that include claims for research and development (R&D) tax credits and/or enhanced expenditure. The BERD database includes more detailed information on R&D expenditure, drawn from a sample of companies that undertake R&D.

This report describes the matching procedure and discusses the results from the matching that has been undertaken, including a brief analysis of the matched observations. The structure of the report is as follows.

- Section 2 describes the extract from HMRC's company tax returns and both the short and long forms from BERD.
- Section 3 discusses the matching procedure, highlighting the potential overlap between the databases, and as such, the level of matching that can be expected.
- Section 4 presents the results from the matching, which has predominantly been based on the unique combination of the CRN and the time variable. The approaches that have been taken to ensure the robustness and accuracy of the matching are highlighted.
- Section 5 analyses the matched data and attempts to identify whether there is a link between the matching rate and the companies that have been matched. Potential gaps in the matched dataset that could have implications for any future econometric analysis are also examined.
- Section 6 concludes by discussing possible reasons behind the current level of matching.

This report presents only aggregate results and does *not* disclose any confidential data. Throughout the report, examples that do *not* reflect real cases have been provided for illustrative purposes. Appendix 2, referred to in this report, is *not* available for circulation as it contains confidential data.

## 2 Description of the data

This section describes the extract from HMRC's CT600 and the availability of the ONS' BERD data.

### 2.1 HMRC's company tax returns

The extract from the CT600 database contains returns from companies that submit at least one claim for R&D tax credits or enhanced expenditure during the period covered by the return. The data is based at the claims level, and includes all claims for R&D tax credits recorded in companies' tax returns that were received by HMRC before January 16th 2006. The data is arranged according to the accounting period end date, and includes the period between the financial years, 1999/2000 and 2005/06.<sup>1</sup>

The number of CRNs in the CT600 extract varies by year. In the CT600 extract—as at January 16th 2006—the number of companies claiming R&D tax credits or enhanced expenditure peaked in 2003/04 after the introduction of the large companies' scheme in 2002. Since 2003/04, the number of companies recorded in the current extract has declined. This may reflect delays in companies submitting their tax returns after the end of the financial year, and as a consequence, the full number of claims for the financial years, 2004/05 and 2005/06, may not have been received by HMRC by January 16th 2006.

The extract from the CT600 database contains company-specific financial data, requested as part of the company tax returns, such as turnover, trading profits or losses, profits chargeable to company tax and capital allowances. In addition, the CT600 extract includes data relating to any claim (or claims) for R&D tax credits or enhanced expenditure. This covers four R&D schemes—vaccines research relief, R&D that is sub-contracted to a small to medium-sized enterprise (SME) by a non-SME, the SME scheme and the large companies' scheme.

As it is possible that companies may claim under more than one scheme, the data has been collapsed to a company level—defined by the CRN—while retaining the number of schemes applied for, before any matching with the BERD data was attempted. It was found that, in some instances, if companies submit more than one claim per year, at different times of the year, the values of expenditure reported in the boxes of the tax return are different. For example, if a company submits two claims at different times of the year, the value of company-specific variables, such as turnover and profits, may not be consistent in the two returns. It would be difficult to make a judgement about whether data for the year where the company submits more than one claim should comprise the total, the latest or the average of the values reported in the boxes of the tax return, and as such, this would affect the robustness of the future econometric analysis.

To ensure the accuracy of any quantitative analysis, Oxera has excluded data for companies in the year in which they submit more than one claim. This removes 1,426 observations—data for a *particular* company in a *specific* year. To check whether the removal of these observations affects the results of the matching, Oxera has repeated the same matching procedure (which is discussed in the following sections) without excluding observations for companies in the year in which they submit more than one claim. However, the number of

<sup>1</sup> The financial year includes the accounting period end date.

matched observations (discussed in section 4) does not alter substantially—only another 34 observations are matched.

Table 2.1 shows the distribution of the number of companies, as measured by the CRN, in the CT600 extract over time.

- The small number of companies—the 55 CRNs—included in the extract in 1999/2000 are likely to be errors in the database, as R&D tax credits were not introduced until 2000, when the SME scheme was introduced.
- The availability of data increases substantially in 2002/03—this is likely to reflect the continuing increase in the number of companies claiming under the SME scheme, and may also partially reflect the introduction of the large companies' scheme in 2002.
- Limited data is currently available for 2005/06 and the full set of claims may not have yet been received by HMRC for 2004/05 and 2005/06—this may be due to companies delaying submitting their tax returns at the end of their financial year. As a result, the current extract from the CT600 dataset—as at January 16th 2006—may not include the full number of claims that HMRC expects to receive for 2004/05 and 2005/06. It should be noted that the BERD database does *not* cover the financial year 2005/06, and as a result, the 464 CRNs in the current CT600 extract for 2005/06 will not be able to be matched.

**Table 2.1 Number of CRNs in the current CT600 extract, by year**

| Financial year | Number of CRNs |
|----------------|----------------|
| 1999/00        | 55             |
| 2000/01        | 1,497          |
| 2001/02        | 2,903          |
| 2002/03        | 4,730          |
| 2003/04        | 5,554          |
| 2004/05        | 4,827          |
| 2005/06        | 464            |

Note: The financial year includes the accounting period end date.  
Sources: Oxera's calculations and HMRC's CT600 extract.

## 2.2 ONS Business Enterprise R&D survey

Conducted annually by the ONS, BERD covers expenditure and employment relating to R&D activity undertaken by UK businesses. As only a small proportion of businesses undertake R&D, and a comprehensive list of those businesses that undertake R&D is not available, the sample of businesses that receive the survey is stratified according to the amount of R&D that is undertaken. Two different forms of the survey are sent to businesses, depending on the amount of R&D that they undertake.

- A *'long' form*—this requests a detailed breakdown of R&D activity. This form is sent to all businesses identified, either through the previous survey or from information gained by

ONS from other sources, as spending more than £3m on R&D. Businesses receiving the long form account for approximately 80% of total R&D expenditure.<sup>2</sup>

- A ‘short’ form—this requests only R&D expenditure and employment totals from the smaller R&D performers. These companies were identified as those businesses that spend less than £3m on R&D expenditure in the previous survey, or those that responded positively to R&D questions in other ONS inquiries, as well as other identified potential R&D performers. However, not all these businesses are sent a short form. A stratified sampling technique is used to select a sample of these smaller R&D performers, depending on the size of the business—as measured by the total number of employees—and industry grouping.

Data from the BERD’s short forms—for small companies which are re-sampled each year, the same business cannot be interviewed more than once every four years—and long forms—which cover large companies and tracks them through time—are discussed below.

### 2.2.1 Short form

The short form covers the period between 1999 and 2004, and is arranged by calendar year, or the nearest 12-month period for which figures are available. The short form contains data relating to turnover, civil and defence intramural R&D expenditure (work carried out within the company in the UK), civil and defence extramural R&D expenditure (work conducted outside the company but funded by the business), the number of employees engaged in R&D and sources of R&D funding.

HMRC identified that the number of short-form records obtained from the ONS is lower than it expected, based on records from the Business Monitor. The ONS has acknowledged that this is due to a number of returned short forms—over 1,000 forms—being excluded from the records, as a result of companies failing to undertake R&D during the reporting period, despite having been identified as R&D performers.<sup>3</sup>

### 2.2.2 Long form

The long form covers the same period between 1999 and 2004, and is also arranged by calendar year, or the nearest 12-month period for which figures are available. The long form incorporates data on the same variables (outlined above) from the short form, as well as more detailed information, including expenditure on basic and applied research and experimental development and capital expenditure on land and buildings.

The long-form BERD data was received from the ONS in separate Excel spreadsheets—three spreadsheets for each year—and as such, needed to be matched together initially before any matching with the CT600 extract could be attempted.

- The *long\_form1* spreadsheet contained data on R&D and company-specific financial variables, such as turnover. The spreadsheet included an IDBR reference number, the CRN and the company name. The data was arranged by the IDBR reference number, and there were no duplicate IDBR reference numbers.
- The *long\_groupdata* spreadsheet contained data on R&D, which (in contrast to *long\_form1*) was split between civil and defence expenditure. The *long\_groupdata* spreadsheet contained some duplicated CRNs—those companies undertaking either (or both) civil or defence R&D in more than one product group, as defined by SIC(92). The

<sup>2</sup> Business Monitor MA14 (2005), ‘Research and Development in UK Businesses, 2004’, National Statistics, January.

<sup>3</sup> If companies fail to undertake R&D, they will not be required to complete the short-form survey in the following year.

*long\_groupdata* spreadsheet contained an IDBR reference number as well as the CRN, but did *not* include the company name. Data in the *long\_groupdata* spreadsheet was re-arranged by the IDBR reference number, so that there were no duplicated IDBR reference numbers.

- The *long\_form2* spreadsheet identified whether companies performed R&D at more than one site, and contained the percentage of R&D performed at the company’s various sites. The *long\_form2* spreadsheet contained an IDBR reference number, but did *not* contain the CRN or company name. The data was re-arranged, so that there were no duplicated IDBR reference numbers.

The long-form BERD data was matched together internally, based on the unique combination of the IDBR reference number (as this was common to all Excel sheets) and the time variable. The *long\_form1* and *long\_groupdata* spreadsheets contained some common variables relating to R&D expenditure. There were instances where the value of these common variables differed across the two Excel sheets, for the *same* CRN in the *same* time period, which were not a consequence of any rounding error. These differences were noted, and would need to be investigated further, before any econometric analysis is undertaken.

There are 85 observations—relating to a *particular* CRN in a *particular* year—that were found to be present in both the short and the long form. These observations are evenly distributed over time, and represent 24 different companies. To ensure the accuracy of any future econometric analysis, these observations have subsequently been excluded. However, to maximise the potential number of matched observations, data has only been excluded in the year that a company appears in both the short and long form. The final database does *not* contain any instances where the same company—as defined by the CRN—appears more than once in the same year.

The following table shows the number of companies in both the short- and long-form BERD database. The figures reported in the table below are based on the cleaned data. It should be noted that the ONS has not provided any long-form and only a very limited number of short-form records for Northern Ireland.<sup>4</sup>

**Table 2.2 Number of CRNs in the BERD database each year**

| Year | Short form | Long form | Total |
|------|------------|-----------|-------|
| 1999 | 1,821      | 306       | 2,127 |
| 2000 | 1,760      | 291       | 2,051 |
| 2001 | 1,743      | 320       | 2,063 |
| 2002 | 1,810      | 331       | 2,141 |
| 2003 | 1,845      | 309       | 2,154 |
| 2004 | 1,901      | 299       | 2,200 |

Note: These results are based on the cleaned data. The year variable is defined according to BERD’s own definition, which closely approximates the start of the financial year that includes the accounting period end date. Source: Oxera’s calculations and ONS’ BERD database.

<sup>4</sup> Only one short-form record per year is available for Northern Ireland between 2001 and 2004.

### 2.2.3 The BERD sample

The ONS targets around 4,000 companies known to perform R&D, which are sent either a short or long form. In 2004, the target sample size increased to 4,500 businesses. Over the period between 2000 and 2004, the average response rate by companies has been high—around 95%.<sup>5</sup> Table 2.3 illustrates changes in the BERD sample over time. The final column shows the completed forms as a percentage of the total BERD population—around 10,000 companies that have been identified as performing R&D. The number of completed forms as a percentage of the total BERD population has averaged around 38% between 2000 and 2004.

**Table 2.3 The BERD sample over time**

| Year | Total number of forms dispatched | Total number of forms completed | Response rate (%) | Completed forms as % of the total BERD population |
|------|----------------------------------|---------------------------------|-------------------|---|
| 2000 | 3,996                            | 3,669                           | 91.82             | 36.69   |
| 2001 | 3,856                            | 3,624                           | 93.98             | 36.24   |
| 2002 | 3,983                            | 3,844                           | 96.51             | 38.44   |
| 2003 | 3,949                            | 3,720                           | 94.20             | 37.20   |
| 2004 | 4,535                            | 4,354                           | 96.01             | 43.54   |

Note: The total BERD population is 10,000.

Source: Oxera's calculations and ONS' Business Monitor.

<sup>5</sup> Business Monitor MA14 (2000, 2001, 2002, 2003, 2004, 2005), 'Research and Development in UK Businesses', National Statistics.

## 3 Matching procedure

This section describes the matching that Oxera has undertaken, beginning with a description of possible identifiers on which the matching could be based. The potential overlap between the CT600 extract and the BERD database is also discussed, before the matching procedure is described in greater detail towards the end of this section.

### 3.1 Potential identifiers for the matching

#### 3.1.1 HMRC's company tax returns

There are several variables in the CT600 extract that could potentially be used to identify companies, and as such, could form part of the matching procedure. These variables include:

- CRN;
- company name;
- taxpayer reference number;
- HMRC's unique identifier;
- company postcode;
- financial year that includes the accounting period end date.

Table 3.1 shows the number of missing or zero values associated with the above variables, as well as the minimum and maximum number of characters (*'length'*). Missing or zero values would limit the potential for these variables to form part of the matching procedure. The minimum and maximum number of characters includes embedded blanks—blank spaces between words.

As an illustrative example, the company name in BERD may contain several embedded blanks between words, such as *Alpha Beta Gamma Ltd*. The same name in the CT600 extract may only contain one embedded blank between words, such as *Alpha Beta Gamma Ltd*. Before any matching was undertaken, the *additional* blank spaces between words, as well as *any* embedded blanks in the CRNs, across both databases were removed. The statistics reported in Table 3.1 have been obtained from the CT600 data when it is arranged at a claims level, before it has been cleaned and subsequently collapsed to the level of the company.

**Table 3.1 Potential identifiers for matching**

|                                  | Total  | Number of missing | Number of zero | Minimum length | Maximum length |
|----------------------------------|--------|-------------------|----------------|----------------|----------------|
| <b>CRN</b>                       | 21,539 | 40                | 41             | 3              | 10             |
| <b>Company name</b>              | 21,539 | 33                | 0              | 4              | 56             |
| <b>Taxpayer reference number</b> | 21,539 | 0                 | 0              | 10             | 10             |
| <b>HMRC's unique identifier</b>  | 21,539 | 0                 | 0              | 16             | 16             |
| <b>Company postcode</b>          | 21,539 | 3,776             | 0              | 2              | 7              |
| <b>Finance year</b>              | 21,539 | 0                 | 0              | 7              | 7              |

Note: The total includes missing and zero observations. The figures reported for the minimum and maximum length describe the length of the variables, once missing or zero values have been excluded, and include embedded blanks—blank spaces between words.

Sources: Oxera's calculations and HMRC's CT600 extract.

As shown in Table 3.1, the CRNs (from the CT600 extract) vary in length, from a minimum character length of 3 to a maximum of 10 characters. Around 6% of the CRNs have been found to contain non-numeric characters at the start of the CRN, including:

- E;
- FC;
- IP;
- NI;
- NO;
- SC;
- SF;
- ZC.

### 3.1.2 ONS Business Enterprise R&D Survey

Both the short and long forms contain the following variables, which could potentially be used as part of the matching procedure:

- CRN;
- company name;
- IDBR reference number;
- year variable that closely approximates the start of the financial year that includes the end date of the accounting period;
- address including postcode.

Tables 3.2 and 3.3 below show the summary statistics associated with the above variables from both the short and long forms, including the number of missing and zero observations. These statistics have been obtained from the raw data, before the data has subsequently been cleaned.

**Table 3.2 Potential identifiers for matching—short form**

|   | Total  | Number of missing | Number of zero | Minimum length | Maximum length |
|---|--------|-------------------|----------------|----------------|----------------|
| CRN                                       | 11,798 | 678               | 1              | 4              | 8              |
| Company name                              | 11,798 | 0                 | 0              | 3              | 35             |
| IDBR reference number                     | 11,798 | 0                 | 0              | 11             | 11             |
| Calendar year                             | 11,798 | 0                 | 0              | 4              | 4              |
| First line of company address (main site) | 11,798 | 0                 | 0              | 6              | 8              |
| Company postcode (main site)              | 11,798 | 5                 | 0              | 1              | 30             |

Note: The total includes missing and zero observations. The figures reported for the minimum and maximum length describe the length of the variables, once missing or zero values have been excluded, and include embedded blanks—blank spaces between words.

Sources: Oxera’s calculations and ONS’ BERD database.

**Table 3.3 Potential identifiers for matching—long form**

|   | Total | Number of missing | Number of zero | Minimum length | Maximum length |
|---|-------|-------------------|----------------|----------------|----------------|
| CRN                                       | 2,153 | 54                | 10             | 5              | 8              |
| Company name                              | 2,153 | 0                 | 0              | 5              | 35             |
| IDBR reference number                     | 1,804 | 0                 | 0              | 11             | 11             |
| Calendar year                             | 2,153 | 0                 | 0              | 4              | 4              |
| First line of company address (main site) | 2,153 | 13                | 0              | 3              | 30             |
| Company postcode (main site)              | 2,153 | 0                 | 0              | 6              | 8              |

Note: Statistics have been based on the *long\_form1* spreadsheet. The total includes missing and zero observations. The total figure reported for the IDBR reference number is lower than the respective figures for the other variables. This is due to the IDBR reference number not being present in the *long\_form1* spreadsheet for 2004, when new data for 2004 was obtained from the ONS, after the original data provided for 2004 was found to incorrectly contain only data for 2003. The figures reported for the minimum and maximum length describe the length of the variables, once missing or zero values have been excluded, and include embedded blanks—blank spaces between words.

Sources: Oxera’s calculations and ONS’ BERD database.

With respect to the CRNs from BERD, around 3% and 5% of observations from the long form and the short form contain non-numeric data at the start. Non-numeric characters, present at the start of the CRNs, from the long form include:

- FC;
- RC;
- SC.

The respective non-numeric characters from the short form include:

- FC;
- IP;
- LP;
- NF;
- NI—one CRN (starting with NI) appears in each year between 2001 and 2004;
- OC;
- RC;

- SC;
- SF.

### 3.1.3 Summary

To match observations from the CT600 extract with those from BERD, a variable that uniquely identifies all observations in each database, but is also common across databases, needs to be identified. The CRN, company name and postcode are the only variables that are common across both the CT600 and BERD datasets. However, as shown by the above tables, there are differences in the format of these variables across both databases. In particular, the non-numeric characters at the start of the CRNs, as well as their length, differ across the BERD dataset and the CT600 extract. The company name and postcode also differ in length across both databases.

## 3.2 The selection of the unique identifier

To undertake any robust econometric analysis, data from the CT600 extract will need to be matched with data for the *same* company *and* year from BERD. As such, matching will need to be based on a unique identifier, comprising variables common to both databases, such as the CRN or company name, in combination with a time variable.

The CT600 data is arranged by HMRC according to the financial year that includes the accounting period end date. While the BERD data—in the format in which it was received from the ONS—is arranged by the ONS according to a time variable that represents the *start* of the financial year that includes the accounting period *end* date for the majority of observations. However, for approximately 4% of observations, the time variable is defined (in the BERD database) by the ONS as the *start* of the financial year that includes the *start* of the accounting period. As an illustrative example, for the *majority* of observations from the BERD database, an accounting period that *ends* on June 30th 2005 would fall into the financial year 2005/06,<sup>6</sup> and would be labelled as 2005.

The start and end dates of the accounting period differ across companies and across databases. The majority of accounting periods in both the CT600 extract and BERD are around 12 months—the mean length of the accounting period is 357 days and 364 days in the CT600 extract and the BERD database respectively. However, these *average* statistics conceal minor differences in the length of the accounting periods across both databases. In the CT600 extract, *no* company has an accounting period that exceeds 365 days. While in the BERD database, the accounting period exceeds 365 days for around 3% of CRNs. For these observations, the mean length of the accounting period is 422 days, while the maximum length of the accounting period is slightly over two years. As the length of the accounting periods only varies slightly across databases, any differences are unlikely to have affected the success of the matching.

Attempts have been made to identify a matching procedure that ensures the greatest degree of overlap between the CT600 and BERD databases. As a result of the minor differences in the length of the accounting periods across databases, Oxera has experimented with alternative definitions of the time variable, to ensure that the matching procedure is as robust as possible. Alternative definitions of the time variable across both databases have been assessed, which are outlined below.

- *The definitions of the time variable provided in the CT600 and BERD databases.*

<sup>6</sup> The financial year 2005/06 begins on April 1st 2005 and ends on April 30th 2006.

- *The start of the financial year that includes the accounting period end date*—this would involve altering the definition of the time variable for the 4% of observations (outlined above) in the BERD database. For example, if the return covered the period between April 2nd 2001 and April 1st 2002, the time variable would be defined as the calendar year, 2002. This is the year of the start of the financial year—between April 1st 2002 and March 31st 2003—that includes the accounting period end date. If the period of the return exceeds two years, such as the period between April 1st 2001 and June 1st 2003, the time variable would be defined as the calendar year, 2003. This represents the start of the financial year—between April 1st 2003 and March 31st 2004—that includes the accounting period end date.
- *The year that includes the majority of the data*—for example, if the start of the return was October 1st 2001 and the end of the return was September 30th 2002, the time variable would be defined as the calendar year, 2002, as this captures the majority of the return. If the accounting period is divided exactly between two years, the time variable is defined as the year of the start of the accounting period. If the accounting period exceeds two years, such as the period between April 1st 2001 and June 1st 2003, the time variable would be defined as the calendar year—2002—that spans the majority of the data.

### 3.3 The potential for matching

Before any matching was undertaken, Oxera assessed the potential degree of overlap between the CT600 and BERD databases. Table 3.4 and Figure 3.1 show the number of unique CRNs in both the CT600 and BERD databases over time, according to the databases' own definition of the time variable.<sup>7</sup> To be able to match the data, the *same* company needs to appear in both databases in the *same* year. As data for 2005 was *not* available in BERD, any data from the CT600 extract for this year will not be able to be matched, and hence it has been excluded from Table 3.4 and Figure 3.1.

**Table 3.4 Availability of data over time, as measured by the CRN**

| Year | Short-form BERD | Long-form BERD | Total BERD | CT600 | Total BERD as % of CT600 |
|------|-----------------|----------------|------------|-------|--------------------------|
| 2000 | 1,760           | 291            | 2,051      | 1,497 | 137%                     |
| 2001 | 1,743           | 320            | 2,063      | 2,903 | 71%                      |
| 2002 | 1,810           | 331            | 2,141      | 4,730 | 45%                      |
| 2003 | 1,845           | 309            | 2,154      | 5,554 | 39%                      |
| 2004 | 1,901           | 299            | 2,200      | 4,827 | 46%                      |

Note: These results are based on the cleaned data. The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date.

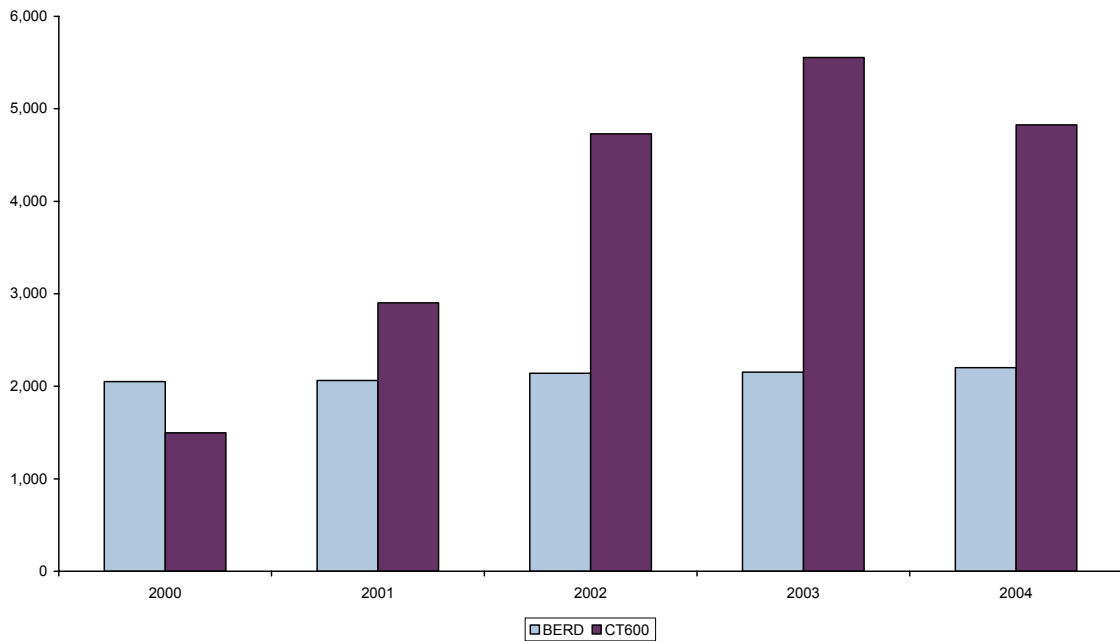
Sources: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

If all companies from the CT600 extract were present in the BERD database in the *same* year, the final column in the above table—total BERD as % of CT600—would illustrate the matching rate that would be obtained. For example, if all CRNs from the CT600 extract in 2002 also appeared in the BERD database in 2002, slightly less than half of all

<sup>7</sup> The year variable in the CT600 extract defines the start of the financial year that includes the accounting period end date, while the year variable from the BERD database closely approximates the start of the financial year that includes the end date of the return.

CRNs from the CT600 extract would be matched. This represents the maximum possible matching rate that could be achieved. However, in practice, the actual matching rate is likely to be substantially lower. Not all R&D performers will have submitted claims for tax credits or enhanced expenditure; therefore, the CT600 database may fail to record all claims from the company's tax return for R&D tax incentives. However, the series provides an interesting benchmark against which the actual matched results can be compared.

**Figure 3.1 Availability of data over time, as measured by the CRN**



Note: The chart is based on the cleaned datasets. The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date. Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

Table 3.4 and Figure 3.1 show the general trend in the number of CRNs in both the CT600 and the BERD databases over time. The number of companies—as measured by the CRN—recorded in the CT600 extract has increased over time, as R&D companies have become more familiar with the tax incentives schemes. As companies have six years in which to submit their claims for R&D tax credits, the full set of claims for 2004 may not have been recorded in the CT600 extract, as at January 16th 2006. As a result, it is likely that the matched results may be slightly lower in 2004.

In an attempt to ascertain the *current* degree of overlap between the CT600 and BERD databases, ignoring the time element, all duplicated CRNs from each database have been removed. To analyse the *potential* for future matching, CRNs appearing in the CT600 extract in *any* year have been matched with CRNs from the BERD database in *any* year from 1999 onwards until the end of 2005.<sup>8</sup> Table 3.5 below shows the results—approximately 19% of the CRNs from the CT600 extract are also present in the BERD database in *any* given year from 1999 onwards until the end of 2005.

<sup>8</sup> This period includes the years 1999 and 2005.

**Table 3.5 Results of the matching based on the CRN**

| CT600                    | Matched observations | Unmatched observations | Total observations |
|--------------------------|----------------------|------------------------|--------------------|
| Number of observations   | 2,001                | 8,750                  | 10,751             |
| <b>% of observations</b> | <b>18.6</b>          | <b>81.4</b>            | <b>100.0</b>       |

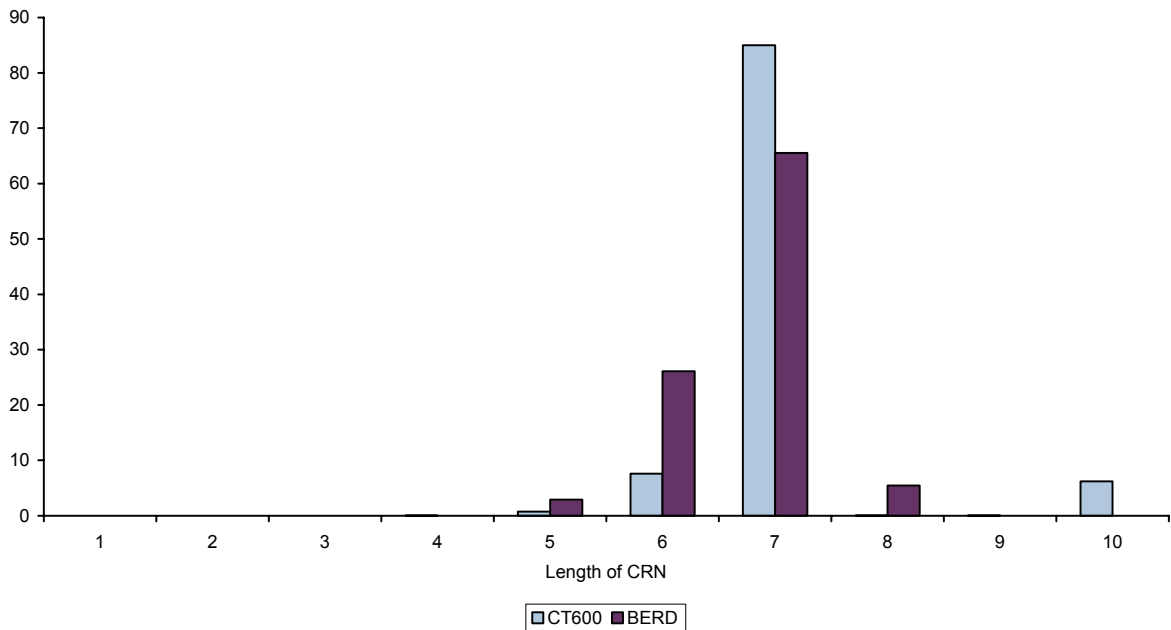
Sources: Oxera’s calculations, HMRC’s CT600 extract and ONS’ BERD database.

Further analysis, undertaken by Oxera, has shown that approximately 48% of the long-form observations from the BERD database in *any* year also appear in the CT600 database in *any* year between 1999 and 2005. If all CRNs from BERD *each* year also appeared in the CT600 database in the *same* year, around 1,300 observations from BERD would be matched each year (on average) between 2000 and 2004. As a result, it is expected that at most, no more than around a quarter of observations from the CT600 database in a *particular* year will match the CRNs in the BERD database in the *same* year between 2000 and 2004.

### 3.4 Matching by CRN

Currently, Oxera has used the unique combination of the CRN and the time variable, as the basis for the matching. Before any matching was undertaken, the CRNs in both databases were analysed by Oxera. Figure 3.2 below shows that the length of the raw (as yet, unmodified) CRNs differed across the CT600 and BERD databases.

**Figure 3.2 Length of CRNs in CT600 and BERD, as a percentage of the total (%)**



Note: The CRNs have not yet been modified, and as such, contain non-numeric characters and embedded blanks.

Source: Oxera’s calculations, HMRC’s CT600 extract and ONS’ BERD database.

To maximise the number of observations that are matched, Oxera cleaned the CRNs in both the CT600 and BERD databases, and care has been taken to ensure the comparability of the CRNs across both databases. To test the sensitivity of the results to the matching procedure, the CRNs have been cleaned in the ways outlined below. The CRNs in the following examples do not represent real cases, and are for illustrative purposes only.

- *Non-numeric characters and all leading zeros have been removed from the start of the CRNs—as an illustrative example, the raw (unmodified) CRN, SC001234, would become 1234, and the unmodified CRN, 0023456, would become 23456.*
- *Non-numeric characters have been removed from the start of the CRN, and all CRNs have been padded with leading zeros to ensure that all CRNs are the same length—10 characters—across both databases—as an illustrative example, the unmodified CRN, SC001234, would become 0000001234, and the unmodified CRN, 0023456, would become 0000023456.*
- *Non-numeric characters have been removed from the start of the CRN. Only those CRNs that contain non-numeric characters have been replaced with leading zeros, to ensure that those CRNs that previously contained non-numeric characters are the same length across databases—as an illustrative example, the unmodified CRN, SC001234, would become 0000001234, while the unmodified CRN, 0023456, would be remain as 0023456.*

To check the accuracy of the matching procedure based on the combination of the modified CRN and the time variable, Oxera employed an algorithm to measure the degree of similarity between the company names across databases.<sup>9</sup> If the modified CRN is the same across both databases, but the algorithm identifies that the company names are substantially different, this may indicate a problem with the matching process. However, this difference may also be due to the same company being identified by different names across databases. As an illustrative example, Henry Spencer’s algorithm identifies *ALPHA PLC (INC ALL SUBSIDIARIES)* to be the same company as *ALPHA PLC*. However, the algorithm may not recognise that *ALPHA & BETA & CO* is the same company as *BETA’S PAINTS*. Any differences have been investigated and have been noted in the matched dataset.

### 3.5 Matching by company name

Attempts have been made by Oxera to match the previously unmatched observations, through the combination of the company name and the time variable. The company name has been cleaned to ensure the similarity of the company names across databases, with phrases such as Limited, Ltd and PLC removed. Table 3.6 illustrates the cleaning of the company name.

**Table 3.6 Cleaning of the company name—illustrative examples**

| Phrases removed                           | Raw company name      | Modified company name |
|---|-----------------------|-----------------------|
| Limited, LIMITED, LTD, ltd, LTD. and ltd. | ALPHA BETA LIMITED    | alpha beta            |
| PLC, PLC., plc and plc.                   | ALPHA BETA PLC        | alpha beta            |
| United Kingdom, UK and U.K.               | ALPHA BETA UK         | alpha beta            |
| Full stops                                | A.L.P.H.A. B.E.T.A UK | alpha beta            |
| Brackets                                  | (ALPHA BETA)          | alpha beta            |
| Hyphens                                   | ALPHA—BETA            | alpha beta            |
| &   | ALPHA & BETA          | alpha beta            |

Note: Additional embedded blanks have been removed.  
Source: Oxera, HMRC’s CT600 extract and ONS’ BERD database.

<sup>9</sup> Henry Spencer’s algorithm performs a match of two expressions, and evaluates to one if the match is satisfied, and zero otherwise.

The differences (illustrated in the above table) in the company names across databases have been investigated. Care has been taken to ensure that all instances of the above phrases have been removed from the company names in both databases. However, there may be some instances where this procedure can be refined further by Oxera. Other differences in the company name, both across and between databases, may also be present, and may be limiting the current level of matching, based on the company name. The following list identifies some potential (illustrative) differences in the company names that may still exist across the CT600 and BERD databases:

- Beta or Beta International;
- Beta or Beta Group;
- Beta or Beta Serv or Beta Service;
- Beta or Beta Associates;
- Beta or Beta Technology or Beta Technologies;
- Beta or The Beta Group.

These differences in company name, both within and across both databases, could be investigated further, and other variations in the company name that may be limiting the current rate of matching, based on the company name, could be identified.

### 3.6 Caveat of the matching procedure

A remaining issue that could affect the robustness of the matching procedure is whether companies are represented at the same reporting level in the CT600 extract and the BERD database. For example, it may be possible that a company is included in the CT600 database at the enterprise level—the smallest group of legal units within a firm. However, the *same* company may be defined in the BERD database by the parent company that owns the enterprise. The enterprise and the parent company may have different CRNs and names, which would lead to a failure to match the observations.

Table A2.1—in Appendix 2—which shows the observations that are matched by company name, but *not* by CRN, suggests that some companies may be represented at different reporting levels across both databases.<sup>10</sup> As an illustrative example, Alpha Holdings Ltd may appear in the CT600 database, while Alpha PLC International may be included in BERD. Alpha PLC International is likely to be the parent company of Alpha Holdings Ltd, and as a result of different CRNs and company names, the observations will not have been matched. Care has been taken by Oxera to ensure that the matching process is as robust as possible, through using Henry Spencer’s algorithm to identify any differences across companies’ names in both the CT600 extract and BERD. A thorough analysis of the data, before any econometric assessment is undertaken, would be able to identify instances where companies may have been represented at different reporting levels across the databases. However, some degree of manual checking would be required, to enhance the robustness of this procedure.

Knowledge of the reporting level of the companies would enhance the robustness of any future econometric analysis. If data for the parent company—Alpha PLC International—was matched with data for the subsidiary—Alpha Holdings Ltd—it is possible that the data contained in one of the databases would need to be transformed. For example, the turnover of the subsidiary—Alpha Holdings Ltd—may be only a fraction of the turnover of its parent company—Alpha PLC International. If these companies were matched, the turnover, and the other variables, associated with either the parent company or the subsidiary would need to be transformed, so that the data series could be accurately compared across databases.

<sup>10</sup> Appendix 2 contains confidential data, and therefore cannot be circulated.

## 4 Results of the matching

The results from the matching that has been undertaken by Oxera are outlined and discussed below, before the matched observations are analysed in the subsequent section.

### 4.1 Matching based on the modified CRN

Table 4.1 below shows the results of the matching procedure that is based on the unique combination of the modified CRN and the time variable. The results (reported in Table 4.1) are obtained from CRNs that have been cleaned, such that non-numeric characters are removed from the CRN, and all CRNs are the same length across both the CT600 and BERD databases. The time variable is represented by the CT600 and BERD's own definition of the year, which closely approximates the start of the financial year that includes the accounting period end date.<sup>11</sup> The results shown in Tables 4.1 and 4.2 only include observations from 2000 until the end of 2004, and *not* 1999 or 2005, as the R&D schemes were only introduced from 2000 and 2002 onwards, and no BERD data is available for 2005.

**Table 4.1 Results of matching based on the CRN**

| <b>CT600</b>             | <b>Matched observations</b> | <b>Unmatched observations</b> | <b>Total observations</b> |
|--------------------------|-----------------------------|-------------------------------|---------------------------|
| Number of observations   | 1,887                       | 17,624                        | 19,511                    |
| <b>% of observations</b> | <b>9.7</b>                  | <b>90.3</b>                   | <b>100.0</b>              |

Note: The results only include observations from 2000 until the end of 2004, and not 1999 or 2005. Sources: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

The table above shows that 1,887 observations have been matched—the *same* CRN in the *same* year appears in both databases—representing approximately 10% of all observations from the CT600 database. The highest level of matching has been achieved between 2002 and 2004, with a greater number of observations being matched from the short form.

Henry Spencer's algorithm has identified that there are 73 instances where the company name differs across databases, despite the CRN being the same. As an illustrative example, the algorithm identifies that *ABC MEDICAL* (from one database) may not be the same company as *ABC HEALTH* (from the other database) despite the raw (unmodified) CRN being the same. These differences have been noted and remain in the database, and could be due to mergers or takeovers or other changes in the company name.

#### 4.1.1 Checks on the matching procedure

To ensure the robustness and the accuracy of the matching procedure, alternative definitions of the time variable and variations in the cleaning of the CRN (outlined in section 3) have been undertaken. However, it has been found that altering the definition of the time variable or the cleaning of the CRNs does not lead to a significant change in the number of matched

<sup>11</sup> In the CT600 database, the year variable is the start of the financial year that includes the accounting period end date. In the BERD database, the year variable closely approximates the start of the financial year that includes the accounting period end date.

observations. If the time variable is re-defined, so that it is based on the year that captures the majority of the data, 1,855 observations are matched. If the definition of time is altered in the BERD database, so that for all observations, the time variable represents the start of the financial year that includes the accounting period end date, this leads to 1,846 observations being matched. Alternative cleaning of the CRNs does not significantly alter the number of matched observations.

## 4.2 Matching based on the company name

Oxera has sought to match the previously unmatched observations (from matching based on the combination of the modified CRN and the time variable) using the unique combination of the modified company name and the time variable. However, this has only led to an additional 64 observations being matched (see Table 4.2).

The raw (unmodified) and modified CRNs associated with these 64 observations have been analysed, to identify why these observations were not matched, when the matching procedure was based on the modified CRN and the time variable. Table A2.1 (in Appendix 2) shows how the raw CRNs associated with these observations differ across both the CT600 and BERD databases.<sup>12</sup> Duplicate observations—when the same company appears in more than one year, and the raw company name and the raw CRN are the same across these years—are *not* reported in Table A2.1. Reasons why the observations were not matched, when the matching procedure was based on the company name, are explained below.

- *Large data entry errors*—the majority of observations have not been matched as a result of major differences among the raw CRNs across databases. As an illustrative example, the raw CRN for Alpha Beta Ltd may be 1234567 in the CT600 extract, while the CRN for the same company in BERD may be reported as 9000000.
- *Minor data entry errors*—around 15% of the observations could not be matched by CRN as a result of small differences in the CRNs across databases. As an illustrative example, the CRN for Gamma Ltd may be reported as 12345678 in the CT600 extract, while the CRN for Gamma Ltd is 12346578 in BERD.
- *Differences in reporting units*—companies may have the same name, but actually represent different reporting units within a larger corporate group. Around 10% of observations could not be matched when the matching procedure was based on the CRN, as the companies appeared to be a UK subsidiary of a larger corporate group or an international firm. As an illustrative example, Sigma UK Ltd may appear in the CT600 extract, while Sigma PLC is present in BERD. As the CRNs of these companies differ across databases, the observations will not have been matched when the matching procedure was based on the CRN. Another 10% of companies could not be matched by using the CRN, as these companies appear to represent the regional subsidiary of a larger group. As an illustrative example, the CRN for Gamma Ltd may be defined as SC0012345 in the CT600 extract and 999988889 in BERD. This suggests that the company contained in the CT600 extract is the Scottish subsidiary of a larger group that is included in BERD.

This analysis has highlighted two important issues that need to be addressed, to ensure the accuracy of any future econometric analysis. Ideally, data entry errors in the CRNs across both the CT600 and BERD databases, and differences in the reporting levels of companies across both databases need to be identified. Identifying differences in the CRNs of the same company across databases would require a degree of manual checking. However, as the

<sup>12</sup> Appendix 2 contains confidential data, and therefore cannot be circulated.

matching procedure has been based on the company name, as well as the CRN, those companies not matched as a result of errors in the recording of CRNs in both databases may be matched, when the procedure is based on the company name. An extensive analysis of the data, before any econometric analysis is undertaken, should be able to identify instances where the reporting level of companies may differ across databases.

Table 4.2 illustrates the results of the matching based on the company name. Overall, 1,951 observations have been matched, if the matching procedure is based on both the CRN and the company name. The table shows that around 10% of all observations from the CT600 extract in a particular year also appear in BERD in that same year.

**Table 4.2 Results of matching based on the company name**

| CT600                    | Matched observations | Unmatched observations | Total observations |
|--------------------------|----------------------|------------------------|--------------------|
| Number of observations   | 1,951                | 17,560                 | 19,551             |
| <b>% of observations</b> | <b>10.0</b>          | <b>90.0</b>            | <b>100.0</b>       |

Note: The results only include observations from 2000 until the end of 2004, and not 1999 or 2005. Sources: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

### 4.3 Results of the matching by year

This section describes the pattern of the matched observations over time. Table 4.3 shows the results of the matching between 2000 and 2004.

**Table 4.3 Results of matching, by year**

|              | Matched observations | Total observations in the CT600 | Matched observations as % of CT600 | Completed BERD forms as % of BERD population | Matching ratio |
|--------------|----------------------|---------------------------------|------------------------------------|--|----------------|
| 2000         | 87                   | 1,497                           | 5.81                               | 36.69  | 15.84          |
| 2001         | 195                  | 2,903                           | 6.72                               | 36.24  | 18.54          |
| 2002         | 528                  | 4,730                           | 11.16                              | 38.44  | 29.04          |
| 2003         | 638                  | 5,554                           | 11.49                              | 37.20  | 30.88          |
| 2004         | 503                  | 4,827                           | 10.42                              | 43.54  | 23.93          |
| <b>Total</b> | <b>1,951</b>         | <b>19,511</b>                   | <b>10.00</b>                       | <b>n.a.</b>                                  | <b>n.a.</b>    |

Note: The matching ratio is defined as the percentage of the matched observations from the CT600 extract divided by the ratio of the number of completed BERD forms to the total BERD population. For example, the matching ratio for 2000 is calculated as  $(5.81/36.69) \times 100$ . However, care should be taken when interpreting this ratio. Meaningful total figures cannot be calculated for the final two columns.

Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

Table 4.3 shows that the percentage of matched observations from the CT600 extract increases over time, up to 2003. Only a small number of observations were matched in 2000 and 2001, as a result of tax relief for R&D companies only being introduced from 2000. This suggests that the dataset, on which any future econometric analysis would be based, is likely to be particularly unbalanced in the first two years of the sample. Only a small number of companies will be present in the matched database for the full years of the sample, with the number of companies rising substantially in 2002. The penultimate column in Table 4.3 shows that either the short- or the long-form version of the BERD survey was completed by around 38% of the BERD population between 2000 and 2004.

As requested by HMRC, Oxera has calculated the *matching ratio*, according to the definition below.

% of matched observations from CT600 / completed BERD forms as % of the total BERD population

The matching ratio is useful in providing an insight into the results of the matching, although care should be taken when interpreting the statistic. For example, comparing the 11% of observations that were matched from the CT600 extract in 2003 with the 37% of the BERD population that completed the survey implies an effective matching rate of around 31%. This is slightly lower than the take-up reported by Clemens et al (2005) in an evaluation survey of R&D tax credits.<sup>13</sup>

Of companies which had not claimed but were aware of R&D tax credits, over half (55 per cent) did not think that they would be eligible for relief or that they would not be able to access value from the relief.

#### 4.3.1 Additional investigations

To increase the level of matching, Oxera has analysed the CRNs associated with large R&D spenders, as measured by total intramural expenditure (from the BERD database). To ensure that companies from the BERD database would qualify for the tax credit, and would therefore appear in the CT600 extract, only those companies with levels of total intramural expenditure on R&D in excess of £25,000 per year have been analysed. It would be expected that companies spending large amounts on R&D are most likely to appear in both the BERD and the CT600 databases.

Table A2.2 (in Appendix 2) shows those large R&D spenders from the BERD database and their availability (and, as such, the instances where the observations have been matched) in the CT600 extract.<sup>14</sup> Any companies that are likely to have merged or have been taken over since the large companies' scheme was introduced in 2002 are *not* included. Table A2.2 shows that a quarter of companies from the sample of the large R&D spenders from BERD are *not* present in the CT600 extract. Of the remaining observations, over 40% of all observations are included in the CT600 database for the three years since the large companies' scheme was introduced in 2002. The analysis shown in Table A2.2 suggests that despite the large companies' scheme not being introduced until 2002, a high rate of matching among the larger R&D companies has been achieved, particularly as a result of the following factors.

- HRMC expects that not all large companies, particularly if they are loss-making, will have submitted claims for R&D tax credits for this period, by January 16th 2006, as companies have six years in which to submit claims for R&D tax incentives.
- HMRC expects that the CT600 dataset will only be broadly complete for two years—the financial years 2002/03 and 2003/04—and over the period, take-up will have been increasing.
- Take-up by large companies may not have been 100% in the first year of the new scheme in 2002 (although, HRMC expects that take-up would have been higher than for the SME scheme).

<sup>13</sup> Clemens, S., Savage, B. and Malicka, D. (2005), 'Research and Development Tax Credits Final Report: Prepared for HMRC', British Market Research Bureau, December, p.5.

<sup>14</sup> Appendix 2 contains confidential data, and therefore cannot be circulated.

## 5 Analysis of the matched data

This section analyses the observations that have been matched by Oxera, and describes the availability of the matched companies through time as well as their characteristics, such as their SIC grouping and level of turnover and R&D expenditure. The matched dataset has been constructed from the results of the matching, shown in Tables 4.1 and 4.2 above. The matching has predominantly been based on the unique combination of the modified CRN and the time variable, which have been defined so that:

- all non-numeric characters have been removed from the CRNs and the CRNs have been replaced with leading zeros, to ensure that they are the same length across databases;
- the time variable has been represented by the definitions provided in the CT600 and BERD databases, which closely approximate the start of the financial year that includes the accounting period end date.<sup>15</sup>

### 5.1 Availability of matched observations

Table 5.1 and Figures 5.1 and 5.2 show the distribution of matched observations over time, and their split between the short and long form. The number of matched observations has varied over time, with a greater number of observations being matched from the short form, compared with the long form. HMRC expects that as companies have six years in which to submit their claims for R&D tax credits, not all large companies will have yet submitted their claims for 2003 and 2004. As a result, it is expected that the number of matched observations from both the short and long form in 2003 and 2004 will increase with time.

**Table 5.1 Distribution of matched observations over time**

|              | Both short and long form | Short form   | Long form  | Short form as % of total matched observations | Long form as % of total matched observations |
|--------------|--------------------------|--------------|------------|---|--|
| 2000         | 87                       | 80           | 7          | 92.0  | 8.0  |
| 2001         | 195                      | 182          | 13         | 93.3  | 6.7  |
| 2002         | 528                      | 388          | 140        | 73.5  | 26.5   |
| 2003         | 638                      | 478          | 160        | 74.9  | 25.1   |
| 2004         | 503                      | 396          | 107        | 78.7  | 21.3   |
| <b>Total</b> | <b>1,951</b>             | <b>1,524</b> | <b>427</b> | <b>78.1</b>                                   | <b>21.9</b>                                  |

Note: The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date.

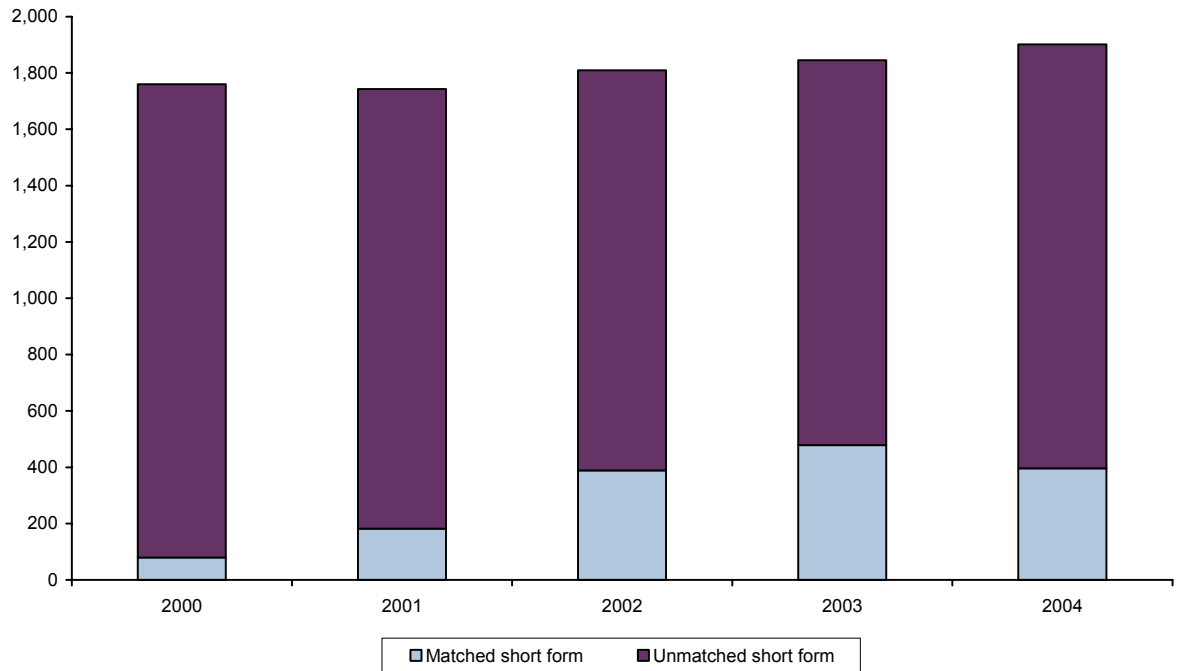
Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

Figure 5.1 shows the distribution of the short-form observations from BERD over time, split by matched and unmatched observations. The number of short-form observations has

<sup>15</sup> In the CT600 extract, the time variable is defined as the start of the financial year that includes the accounting period end date. For almost all observations in the BERD database, the time variable represents the start of the financial year that includes the accounting period end date. For example, if the return covers the period between April 1st 2000 and March 31st 2001, the time variable (for almost all observations) has been defined as the year 2000.

remained relatively constant over time, while the number of matched short-form observations has increased up until 2003, before declining in 2004.

**Figure 5.1 Distribution of matched short-form observations over time**

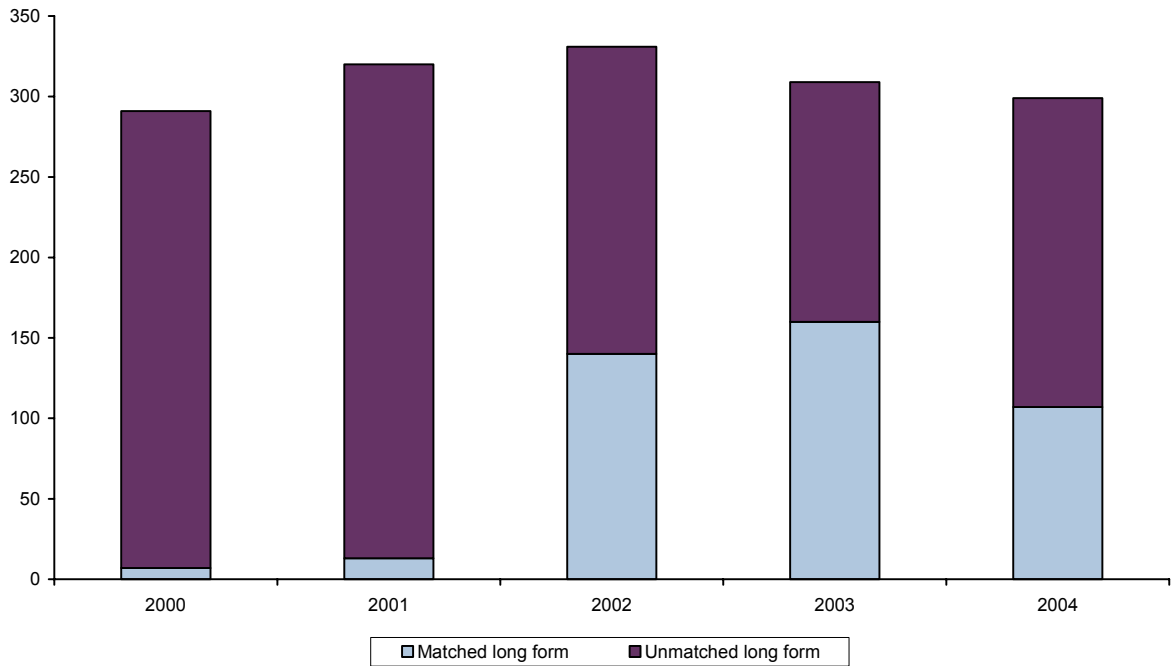


Note: The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date.

Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

Figure 5.2 below illustrates the distribution of the matched long-form observations over time. In comparison with the short form, there has been slightly greater variation in the total number of long-form observations over time. A small number of observations from the CT600 extract have been matched with the BERD's long form, before the large companies' scheme was introduced in 2002. These observations will need to be excluded from the dataset before any econometric evaluation of R&D tax incentives can be undertaken.

**Figure 5.2 Distribution of matched long-form observations over time**

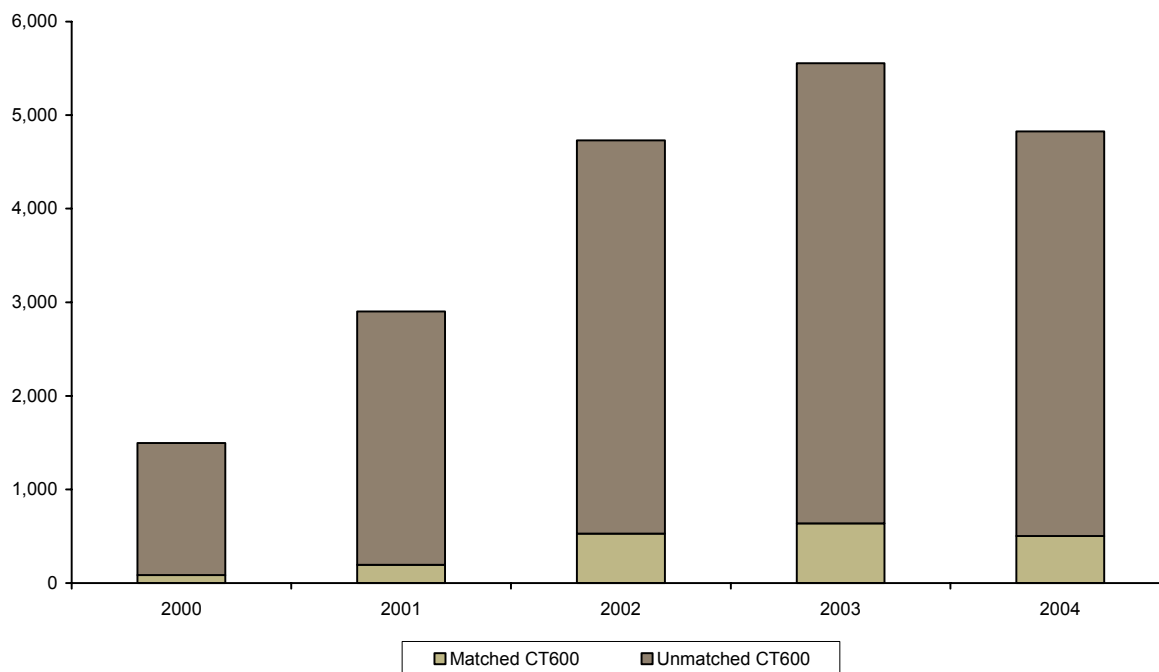


Note: The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date.

Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

Figure 5.3 shows the total number of observations from the CT600 extract each year, split by matched and unmatched observations. This shows the clear variation in the number of observations recorded in the CT600 extract, and the rise in the number of matched observations in 2002 and 2003.

**Figure 5.3 Distribution of matched observations over time**



Note: The year variable is defined according to the BERD and CT600 definition, and closely approximates the start of the financial year that includes the accounting period end date.  
 Source: Oxera’s calculations, HMRC’s CT600 extract and ONS’ BERD database.

Oxera has analysed the availability of data for the *same* company over time in the matched dataset. This will help to inform the nature of the dataset that could be expected for future econometric analysis. Table 5.2 shows the availability in the matched dataset of consecutive data over time for the same company. It is noticeable that:

- data for only one year is available for the majority of the CRNs;
- data for two years is available for 319 CRNs, although only data for two consecutive years is available for around 237 CRNs—74.29% of 319;
- data for three consecutive years is available for around 100 CRNs.

**Table 5.2 Frequency of matched observations**

| Number of years | Number of CRNs | Percentage of CRNs with data for consecutive years |
|-----------------|----------------|--|
| 1               | 910            | n.a.   |
| 2               | 319            | 74.29  |
| 3               | 108            | 92.59  |
| 4               | 16             | 93.75  |
| 5               | 3              | 100.00   |

Source: Oxera’s calculations, HMRC’s CT600 extract and ONS’ BERD database.

This suggests that the matched dataset is fairly unbalanced at present—there is a limited availability of consecutive data for the same company over time. This would affect any econometric evaluation of R&D tax incentives, as it would be difficult to identify the impact of R&D tax incentives on R&D expenditure over time.

## 5.2 Characteristics of the matched observations

Oxera has analysed the characteristics of the matched observations, according to industry sector, turnover and levels of R&D expenditure, to ascertain whether there is likely to be any bias in the matched sample.

### 5.2.1 Industry sector

Table 5.3 compares the distribution of the matched observations across SIC product groups.<sup>16</sup> The description of the broad product groups is provided in Appendix 1.<sup>17</sup> The majority of the matched companies operate in the services, manufacturing, electrical machinery and mechanical engineering sectors. This is broadly in line with the distribution of all the companies reported in BERD.

**Table 5.3 Matched observations—broad SIC (92) product groups (%)**

| Broad SIC (92) product groups | Matched observations | Total BERD    |
|-------------------------------|----------------------|---------------|
| Aerospace                     | 0.97                 | 1.07          |
| Other                         | 1.64                 | 3.11          |
| Transport equipment           | 2.87                 | 3.97          |
| Pharmaceuticals               | 3.08                 | 1.79          |
| Chemicals                     | 7.18                 | 6.70          |
| Mechanical engineering        | 11.58                | 14.89         |
| Electrical machinery          | 11.74                | 10.69         |
| Other manufacturing           | 15.33                | 21.99         |
| Services                      | 45.62                | 35.79         |
| <b>Total</b>                  | <b>100.00</b>        | <b>100.00</b> |

Note: The definition of the product groups follows the approach adopted by the IFS (2003).<sup>18</sup> The data in the above table has altered from the previous version of the matching report.

Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

A high percentage of the matched observations are from the services sector, which includes R&D companies from the software and IT services. The low proportion of matched companies from the pharmaceuticals sector may reflect a difficulty in matching R&D companies in relatively new sectors, such as biotechnology. These sectors may also contain a high number of new small companies, which may be difficult to identify across the BERD and the CT600 extract.

### 5.2.2 Turnover

To assess whether there is any link between the matching rate and the companies that have been matched, the turnover of the matched companies has been compared with the turnover of all companies appearing in the CT600 extract each year. Figure 5.4 shows the distribution of the matched companies according to turnover bands—quartiles—calculated from all companies included in the CT600 extract. As such, the observations have been divided into four quartiles, each containing 25% of the total number of observations from the CT600

<sup>16</sup> The BERD database contains SIC codes, but these codes are not included in BERD.

<sup>17</sup> Appendix 1 is included at the end of this report.

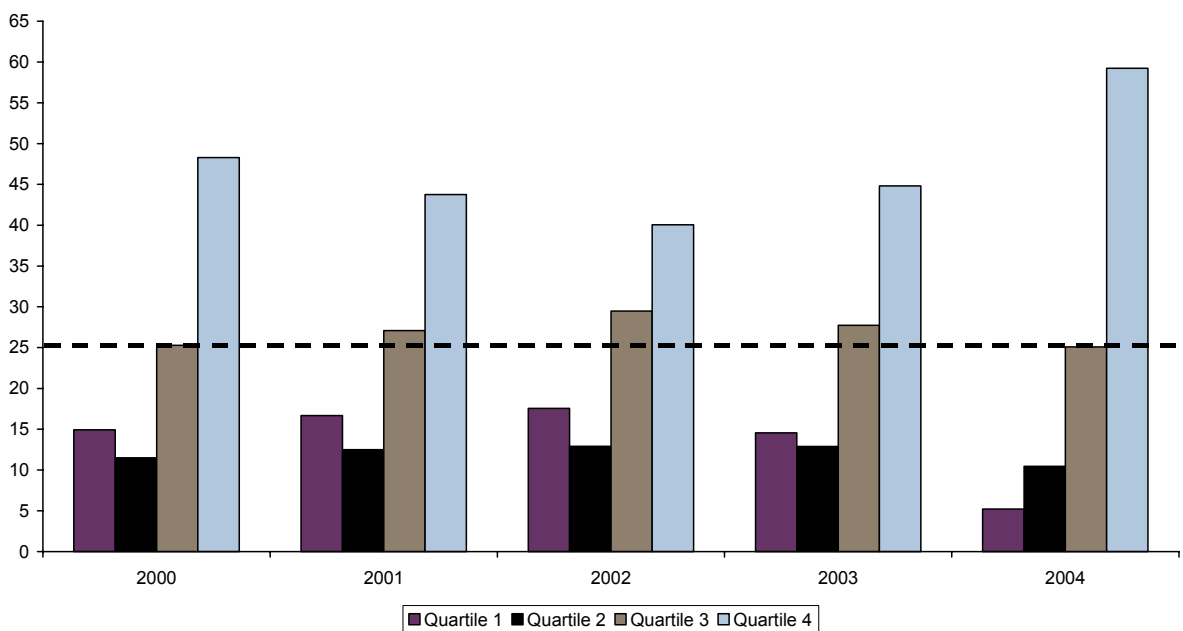
<sup>18</sup> Institute for Fiscal Studies (2003), 'Report on Estimating Private and Social Rates of Return to R&D Using Matched ARD and BERD Micro Data', January, p. 45.

extract. The matched companies' turnover has been compared according to these intervals that have been defined by the turnover of all the companies included in the CT600 extract. As such, Figure 5.4 can identify whether the matching has been biased towards companies of a particular size.

For example, if the matched sample comprises a greater proportion of companies with particularly large turnover, compared with all companies included in the CT600 extract, this can be identified in Figure 5.4 by the quartile 4 column exceeding 25%. Conversely, if a lower proportion of companies with the highest turnover—quartile 4—have been matched compared with the distribution of turnover for all companies included in the CT600 extract, this is represented by the quartile 4 column being less than 25%.

Figure 5.4 shows that the turnover distribution of the matched observations has been fairly stable over time, apart from 2004, where a lower proportion of the smallest companies have been matched. The majority of the matched observations are those companies from the CT600 extract with the highest levels of turnover—the top 25% (quartile 4). This suggests that there is a lack of data from companies with the lowest levels of turnover from the CT600 extract—the bottom 25% (quartile 1). These results might reflect the sampling of the BERD survey, where the largest R&D spenders are asked to complete the long form every year, while the circulation of the short forms is based on a much smaller sample. This limitation biases the results in favour of larger firms and will continue to do so as long as the short form only sample companies only once every four years.

**Figure 5.4 Turnover of the matched companies, relative to the CT600 population (% of the total number of matched companies each year)**

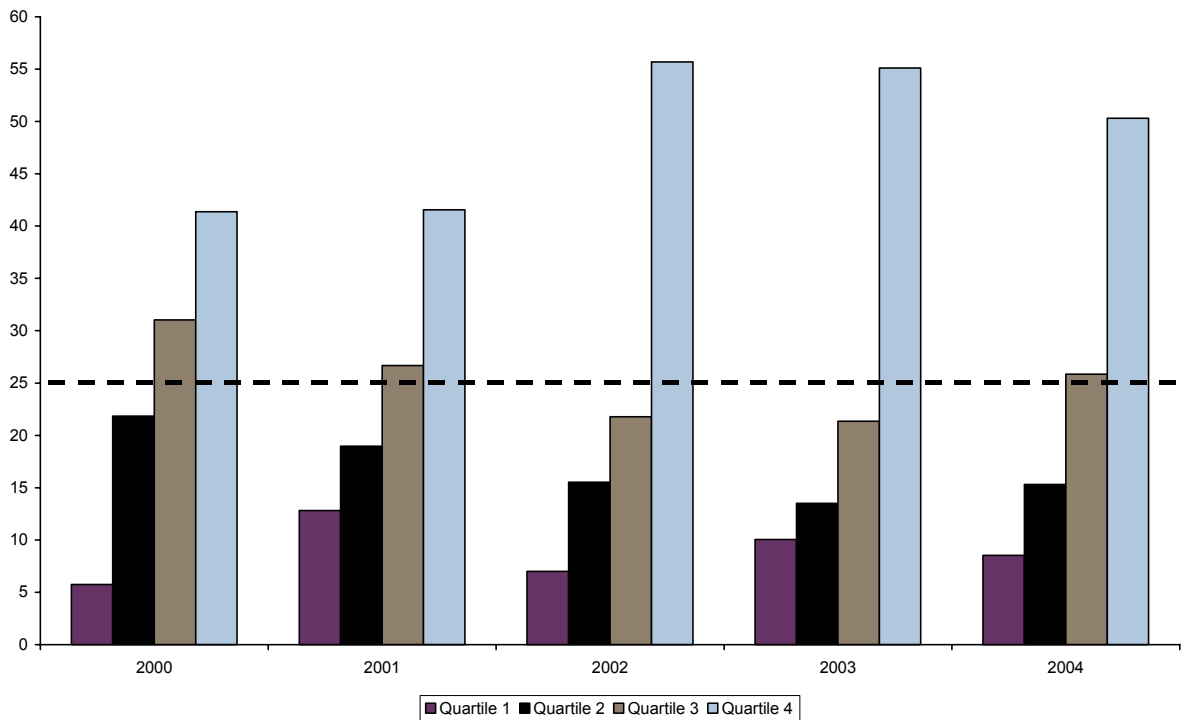


Note: Quartile 1 represents companies from the CT600 extract with the lowest levels of turnover—the bottom 25%. Quartile 2 represents companies with turnover exceeding the lowest 25%, but with turnover less than or equal to the median. Quartile 3 represents companies with turnover exceeding the median, but with turnover less than the top 25%. Quartile 4 represents companies with the highest levels of turnover—the top 25%. Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

### 5.2.3 R&D expenditure

The above analysis from section 5.2.2 has been repeated according to the level of R&D expenditure, as measured by the actual amount of R&D expenditure reported in the CT600 extract. Figure 5.5 shows the distribution of the matched companies according to R&D expenditure bands—quartiles—calculated from companies included in the CT600 extract.

**Figure 5.5 R&D expenditure of the matched companies, relative to the CT600 population (% of the total number of matched companies each year)**



Note: Quartile 1 represents companies from the CT600 extract with the lowest levels of R&D expenditure—the bottom 25%. Quartile 2 represents companies with R&D expenditure exceeding the lowest 25%, but with R&D expenditure less than or equal to the median. Quartile 3 represents companies with R&D expenditure exceeding the median, but with R&D expenditure less than the top 25%. Quartile 4 represents companies with the highest levels of R&D expenditure—the top 25%.

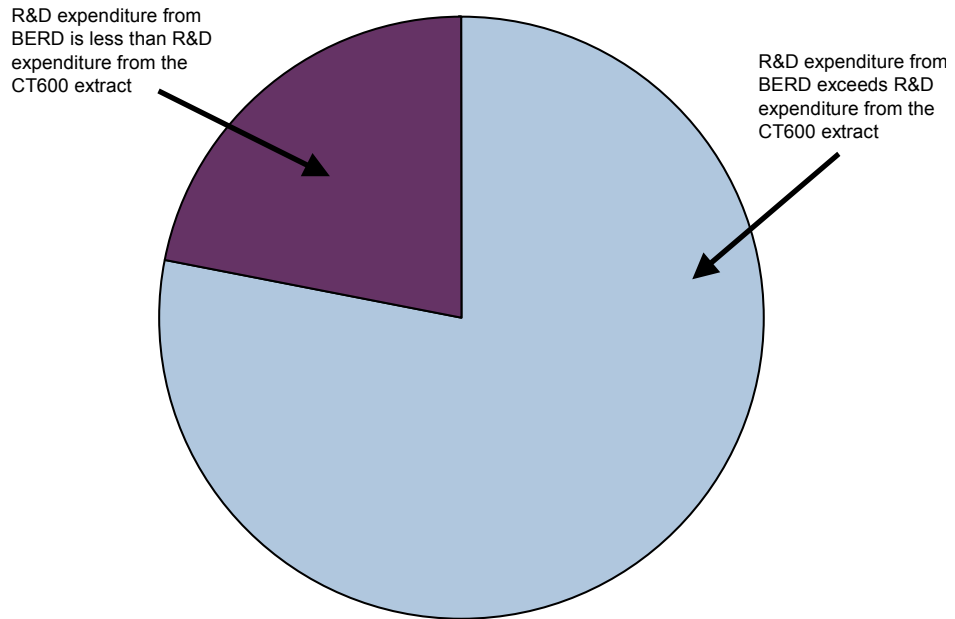
Source: Oxera’s calculations, HMRC’s CT600 extract and ONS’ BERD database.

Figure 5.5 shows that the matched sample is biased towards those companies with high levels of R&D expenditure—quartile 4—and the distribution of R&D expenditure of the matched companies has remained relatively constant over time (this is as expected due to the positive correlation between turnover and R&D expenditure).

### 5.3 R&D expenditure across both datasets

Levels of R&D expenditure recorded in BERD have been compared with R&D expenditure that is reported by companies in their tax returns, to claim tax credits or enhanced expenditure. Oxera has compared the sum of intramural and extramural R&D expenditure from BERD with the total amount of R&D expenditure from the CT600 extract. Figure 5.6 below shows that, for the majority of the matched companies, R&D expenditure reported in the tax returns is lower than levels of R&D expenditure from BERD.

**Figure 5.6 Comparisons of R&D expenditure across BERD and the CT600 extract**



Source: Oxera's calculations, HMRC's CT600 extract and ONS' BERD database.

On average, total intramural and extramural R&D expenditure reported in BERD exceeds R&D expenditure from the CT600 extract by nearly £150,000 (in constant 2000 prices).<sup>19</sup>

- If R&D expenditure from BERD exceeds the level reported in the CT600 extract, the average difference is around £400,000 (in constant 2000 prices).
- For the small number of companies where R&D expenditure from the CT600 extract exceeds the level of expenditure from BERD, the average difference is around £40,000 (in constant 2000 prices).

<sup>19</sup> The average is defined as the median (midpoint). The median is reported instead of the mean, as the mean is strongly influenced by large outlying observations.

## 6 Conclusions

To date, Oxera has undertaken matching based on the unique combination of the modified CRN and the time variable. Oxera has undertaken some matching using the modified company name and the time variable, although there is potential for refinements to be made to the matching process based on the company name. However, as only 19% of the CRNs from BERD in *any* year appear in the CT600 extract in *any* year between 1999 and 2005, it is unlikely that any refinements to the matching process (based on the current datasets) will lead to the percentage of matched observations from the CT600 extract rising significantly above this level.

At present, 1,951 observations have been matched, which represents around 10% of all observations from the CT600 extract. A higher proportion of observations have been matched from the long form (relative to the total number of long-form respondents), although a greater absolute number of observations have been matched from the short form. The matched observations are concentrated among companies with the highest turnover that operate in the services, manufacturing, electrical machinery and mechanical engineering sectors. For the majority of companies, data is available for only one year, and there are only around 100 companies with data for three consecutive years.

The limited level of matching may reflect the following factors.

- The re-sampling of companies from the short form—the same business cannot be interviewed more than once every four years.
- The large companies' scheme was not introduced until 2002, and therefore at most only three years of data will be available for the large companies. However, as the CT600 data is not complete for 2004, for the majority of observations, only one or two years will be available.
- Large companies cannot claim payable credit if they are loss-making, although they can carry forward enhanced losses—Oxera does not have exact information on the number of loss-making companies.
- The lack of Northern Ireland data from BERD—although the small number of Northern Ireland forms is unlikely to significantly affect the results of the matching.
- Around 1,000 companies did not undertake R&D during the reporting period, despite being identified as R&D performers by BERD.
- Companies may delay filing their tax returns, causing discontinuities in the availability of data over time.
- The number of claims for the SME scheme has taken several years to build up. The lower take-up in the first two years after the scheme was introduced has resulted in much lower matched observations in these years.

Taking the above factors into account, the current level of matching appears reasonable, and indicates the baseline level of matching that may be expected in the future. It is expected that the number of matched observations may rise over time, particularly as companies have six years in which to submit their claims for R&D tax credits.

At present, the lack of continuous data for the same companies over time would severely restrict the ability to undertake any econometric analysis. There are a very low number of

matched observations in 2000 and 2001. This would mean that the potential dataset, on which an econometric analysis could be carried out, would be particularly unbalanced in the early years of the sample. In addition, the concentration of the matched observations among those companies with the highest turnover (in part due to BERD sampling) and R&D expenditure would lead to a potential bias in the results from any quantitative assessment of the impact of R&D tax credits or enhanced expenditure.

There is scope to improve the number of matched observations by making refinements to the matching process, based on the company name, or by undertaking matching based on HMRC's amended dataset. As the full impact of R&D tax credits will take time to emerge, even if all companies from BERD in *each* year appeared in the CT600 extract in the *same* year, there is unlikely *at present* to be a sufficiently long time series availability of data over which robust estimates of the impact of R&D credits could be obtained. However, as more data becomes available with time, this would enable a robust econometric assessment of R&D tax credits to be undertaken, at least for large companies (given the sampling of the short form survey).

The analysis has shown that improvements to the recording of data could be implemented, while extra data becomes available over time.

- The CT600 extract included data on only those claims for R&D tax credits or enhanced expenditure that were submitted by companies. To evaluate the impact of tax incentives for R&D, using the outcome of the claims—the amended data—may represent a more robust approach. This should improve the accuracy of the dataset, as erroneous or invalid claims should be excluded, while claims not recorded in a company's tax return would be included.
- The availability of BERD data, and hence the degree of overlap with the CT600 data, could be improved by obtaining information for Northern Ireland. In addition, the substantial number of companies—around 1,000—that did not undertake R&D during the reporting period, although the companies were identified as being R&D performers, should be investigated. Only sampling smaller R&D firms once every four years significantly limits the potential for assessing the impact of R&D tax incentives on smaller R&D firms.

## Appendix 1 Definitions of product groups

**Table A1.1 Definitions of product groups**

| Sector                 | Description   | SIC code   |
|------------------------|---|--|
| Pharmaceuticals        | Pharmaceuticals, medical chemicals and botanical products   | 24.4   |
| Chemicals              | Chemicals, chemical products and man-made fibres  | 24 (excluding 24.4)  |
| Mechanical engineering | Non-metallic minerals, basic iron and steel and ferro-alloys, fabricated metal products and machinery and equipment   | 26, 27.1, 27.2, 27.3, 27.51, 27.52, 28 and 29                                |
| Electrical machinery   | Office machinery, computers, electrical machinery, radio, TV and communications equipment   | 30, 31 and 32  |
| Transport equipment    | Motor vehicles, motor parts and engines, railway locomotives and rolling stock, ships and boats and transport equipment   | 34, 35.2, 35.4 and 35.5  |
| Aerospace              | Aircraft and spacecraft   | 35.3   |
| Other manufacturing    | Food, beverages, tobacco, textiles, clothes, leather, footwear, wood and wood products, pulp, paper, publishing, printing, recorded media, refined petroleum products, nuclear fuel, rubber and plastics, precious and non-ferrous metals, medical and precision instruments, furniture, jewellery, musical instruments, sports goods, games and toys and other manufacturing | 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27.4, 27.53, 27.54, 33, 36 and 37    |
| Services               | R&D services, wholesale, retail, transport, storage, post, financial, real estate, computing and public administration  | 50, 51, 52, 55, 60, 61, 62, 63, 64, 65, 66, 67, 70, 71, 72, 73, 74 and 75–99 |
| Other                  | Agriculture, extraction, electricity, gas, water and construction   | 01, 02, 05, 10, 11, 12, 13, 14, 40, 41 and 45                                |

Source: Institute for Fiscal Studies (2003), 'Report on Estimating Private and Social Rates of Return to R&D Using Matched ARD and BERD micro data', January, p. 45.

**Oxera**

Park Central  
40/41 Park End Street  
Oxford OX1 1JD  
United Kingdom

Tel: +44 (0) 1865 253 000

Fax: +44 (0) 1865 251 172

[www.oxera.com](http://www.oxera.com)