ACCOUNTING INFORMATION FREE OF SELECTION BIAS: A NEW UK DATABASE 1953-1999[♣]

Stefan Nagel[*]
London Business School

October 2001

First draft: December 2000

[*] PhD student, London Business School, Sussex Place, Regent's Park, London NW1 4SA, United Kingdom, Phone +44 (0)20 7262 5050, Fax +44 (0)20 7724 7875, snagel@london.edu, http://phd.london.edu/snagel.

[♣] I thank the Sloan School of Management at MIT and Jon Lewellen for hosting me as a visiting PhD student while I was working on parts of this project. I am grateful to my dissertation advisor Elroy Dimson, S.P. Kothari, Jan Mahrt-Smith, Garrett Quigley, Ilya Strebulaev, and workshop participants at London Business School for many helpful suggestions, and to Bernd Hanke, Chris Higson, and David Stolin for valuable contributions to this project. Special thanks go to Pawan Jagetia, Daphne Lui, Sanjay Pareek, and Hongjun Yan, who shared with me the pleasures of handcollecting historical balance sheet information. This project has been generously supported by Dimensional Fund Advisors. I gratefully acknowledge financial support from the Lloyd's Tercentenary Foundation, the Economic and Social Research Council (ESRC), and Edward Jones.

ACCOUNTING INFORMATION FREE OF SELECTION BIAS: A NEW UK DATABASE 1953-1999

ABSTRACT. Accounting information is indispensable in empirical finance. Unfortunately, existing databases, particularly for non-US markets, suffer from selection bias. I construct a new UK data set by supplementing existing databases with handcollected balance sheets. It contains about 100,000 firm-years of data, covers virtually all UK firms listed on the London Stock Exchange between 1953 and 1999, is fully cross-referenced to the London Share Price Database and free of selection bias. Use of previously existing UK databases is shown to introduce biases against small and high book-to-market stocks, and downward biases in sample moments of returns, a symptom of truncation of extreme observations. The new data set is particularly useful in contexts where inference is sensitive to selection bias or repeated mining of the US data is detrimental, and for any study that needs accounting measures such as book-to-market ratios for UK firms.

Keywords: Accounting data, selection bias, survivor bias, book-to-market

JEL Classification: G14, G15, G10

1. Introduction

Accounting information linked to share price data is an essential ingredient of empirical research in finance. For example, many stylized facts in the literature on cross-sectional return predictability involve accounting information, such as book-to-market or earnings-to-price ratios. In response to this empirical evidence on predictability researchers develop new theories and models of pricing in equity markets (e.g. Fama and French 1993, Daniel and Titman 1997, Barberis, Shleifer, and Vishny, 1998, Daniel, Hirshleifer, and Subrahmanyam 1998, Hong and Stein 1999).

How can we test these empirically motivated models? There is certainly a problem when the same data set is used to motivate and to test a model (Merton 1987, Lo and MacKinlay 1990). One way to at least partially address this problem is to test models with non-US data, as in Fama and French (1998), or Daniel, Titman, and Wei (2001). However, international data usually comes with its own problems. First, accounting data for non-US markets is usually available only for a relatively short time-series, or only for a relatively small cross-section of firms. Studies on stock return predictability, for example, are likely to suffer from this problem. Second, the quality of non-US data sets tends to be inferior. While there have been some concerns about selection and survivor biases in the US COMPUSTAT database,[1] this problem is likely to be worse in non-US datasets.

This paper introduces a new database that addresses some of these problems. It provides essential balance sheet information for all UK firms ever listed on the London Stock Exchange from 1953–1999 at annual frequency. I construct the data set by merging two previously existing databases, Datastream and the Cambridge/DTI database, and by handcollecting balance sheets for all remaining firms not covered in

[1] See Kothari, Shanken, and Sloan (1995) and references therein.

these sources. In total, the data set covers close to 100,000 firm-years. It is fully cross-referenced to the London Share Price Database (LSPD)[2], which covers the period since 1955. Hence, this is a rich new data source that can be used to test theories on non-US data. Furthermore, since it covers virtually the whole population of listed firms at any point in time between 1953 and 1999, selection bias is absent by construction. This is a distinguishing feature of this new data set compared to other international data sources.

The purpose of this paper is twofold. The first objective is to introduce this database to a wider audience and to provide information on its characteristics. I show that the new database misses only about 7% of the underlying population of listed stocks. Much of it can be traced to new listings. Commonly, accounting data is not available for the time before the initial listing. In some cases the database also misses firms that are about to delist. However, a look-ahead bias can be avoided by substituting lagged accounting information for missing years of data. Hence, except for the tendency to miss newly listed firms, the database is essentially free of selection bias. With respect to selection bias, the new database is superior to the US COMPUSTAT database, which misses a far greater share of firms in earlier decades.

Second, given that the new database covers virtually the whole underlying population, I also evaluate the selection biases inherent in previously existing databases. I find that the existing ones systematically miss small firms and high book-to-market stocks. Furthermore, returns, for example on an equal-weighted index, are biased downwards. Sample standard deviations, skewness, and kurtosis of returns are understated, which closely mirrors the simulation evidence in Kothari, Sabino and Zach (2001), where extreme observations are truncated deliberately. This suggests

[2] The LSPD is maintained at London Business School, and it is forthcoming on Wharton Research Data Services (WRDS) at wrds.wharton.upenn.edu.

that existing databases miss extreme performers in both tails of the return distribution. The effect is strongest in the 1970's when Datastream exhibits survivor bias, suggesting that at least part of it is due to this form of ex-post selection bias. Kothari, Sabino, and Zach (2001) demonstrate that there are research designs in which such truncation of extreme observations can create spurious evidence of predictability. The evidence in this paper complements their findings by showing that this truncation is a real feature of existing accounting data sets.

It is important to note that these problems are unlikely to be unique to the UK data. Not much is known about the firms missed by COMPUSTAT in the 60's and 70's, and in other non-US data sets these problems are probably more severe rather than less. Hence, the new UK data set fills an important gap. It will be useful for researchers who want to (1) investigate research questions that are sensitive to truncation of extreme observations and other forms of selection bias, (2) test theories motivated by empirical findings on US data with a credible non-US data set, or (3) examine UK data with standard methodology, such as book-to-market matching in long-horizon price performance event studies.[3]

The rest of the paper proceeds as follows. Section 2 presents a brief discussion of potential biases inherent in accounting databases, and it provides information on the construction of the new database. Section 3 analyzes the coverage of the new database and compares it to previously available data. Section 4 investigates the reasons for missing data in the new database. Section 5 examines sample selection biases in previously existing databases. Section 6 concludes.

[3] See e.g. Barber and Lyon (1997), Kothari and Warner (1997), and Mitchell and Stafford (2000) for references to the extensive literature on long-horizon event study methodology.

2. Sample Selection Issues

This section discusses some sample selection problems germane to databases of historical accounting information. Some of them are inherent in the data collection process. In constructing the new database I attempt to avoid them as much as possible by an appropriate set-up of the data collection process and a suitable definition of the target sample.

2.1 Selection Bias in Accounting Databases

A data set suffers from selection bias when a selection rule other than random sampling determines whether a given observation on a firm enters the database. An econometrician who ignores the inherent selection rule is implicitly conditioning on the outcome of the selection process. This may lead to biased inference.[4]

From the viewpoint of financial markets research it is useful to distinguish ex-post and ex-ante selection biases. In the case of ex-post selection bias the selection of sample firms is based on information that was not available at the time the observation was generated. Survivor bias is a prominent form of ex-post selection bias. It exists in data sets that exclude a disproportionate share of non-surviving firms, and it arises, for example, when data is backfilled, or when data is missing prior to delisting events. By using such biased accounting information, the econometrician is implicitly conditioning on the firm's survival. Ex-post selection bias is a particularly serious concern in predictability studies, since portfolios formed on accounting data with inherent ex-post selection bias do not represent trading strategies that are replicable ex-ante. Banz and Breen (1986), Kothari, Shanken, and Sloan (1995),

[4] For a general discussion of selection bias in panel data sets see e.g. Verbeek and Nijman (1995).

Davis (1996), and Kothari, Sabino, and Zach (2001) show that the results of empirical studies can be sensitive to this problem.

In the case of ex-ante selection bias the statistical problems may be less severe, as there is no conditioning on ex-post information. Nevertheless, it can give rise to systematic differences in characteristics between sample and population. For instance, firms in a database may be concentrated in particular size groups or industries, or data may be missing prior to initial public offerings. However, if the selection rule is known, then, despite ex-ante selection bias, any trading strategy based on accounting information that can be constructed in the data is, in principle, feasible in reality, too. This is not true for ex-post selection bias, and it makes ex-ante selection biases easier to deal with.

2.2 Sample Definition for the new UK Database

Complete coverage of the whole population is the best safeguard against selection bias. I use the master index of the London Share Price Database (LSPD) to identify all UK firms with a listing on the LSE betweeen 1955 and 1999.[5] Unfortunately, at present the LSPD does not offer price information for all firms prior to 1975. The pre-1975 price data consists of four overlapping samples, one of which is a stratified random sample covering 33% of the population at the start of the database in 1955 and 33% of all new listings each year. Despite this limitation, I collect accounting information for the full population, anticipating a potential future expansion of LSPD coverage. Considering that in many research designs share price information is matched with lagged accounting information, I set the collection start date to fiscal year 1953 to utilize the full length of LSPD coverage. When a firm has

[5] The master index of the LSPD contains records of all listed firms on the LSE, and it tracks name changes, listing suspensions, mergers etc. For more information on the LSPD see e.g. Dimson and Marsh (2001).

multiple share classes, they are linked to the same unit by assigning a unique issuer code to each firm. I exclude firms without official listing (Unlisted Securities Market, Alternative Investment Market etc.) and investment trusts (closed-end funds). In the following "all firms" or "all listed firms" refers to the sample just described.

3. Key properties of the new database

I obtain accounting data from three different sources. Data is available in electronic form from Datastream and from the Cambridge/DTI database. To supplement these sources I manually collect balance sheets from hardcopy issues of the Stock Exchange Yearbooks. This section highlights some key features and problems of each data source and describes the coverage of the new database.

3.1 Data sources

Datastream is the prime source of electronically available accounting information on UK firms. Data is available from the end of the 1960's. In order to merge the data with the LSPD I link the two databases as follows. Cross-reference lists of LSPD company numbers (the LSPD equivalent to the CRSP PERMNO) and Datastream codes to SEDOL codes (security identifiers assigned by the Stock Exchange) provide a starting point. Unfortunately, these SEDOL codes are unstable, and codes of delisted get reused. I clean up the link by comparing all LSPD-Datastream matches by name, market capitalization, and the end point of available histories in each database. In case of deviations I manually search for the correct match. In some cases I also have to link a LSPD company number to two different Datastream codes in different time periods. This case can arise, for example, if Datastream and the LSPD do not agree on the company that "survives" in a merger.

Preliminary inspection of the data reveals that until the end of the 70's the coverage of Datastream is very poor. Furthermore, the data for the first decade seems to have been backfilled in the mid-70's, as there is no information on dead companies in the database until this point. In addition, Datastream systematically misses small companies and certain industries (e.g. financials, property companies, colonial industries) in the 60's and 70's. This suggests that there is a serious danger of selection bias in the Datastream sample.

Next, I merge the LSPD with the Cambridge/DTI Databank of Company Accounts. I look up data for every firm-year that is not covered by Datastream. The Cambridge/DTI database is based on data from published company accounts of UK registered companies, starting in the 1940's, originally stored by the Department of Trade and Industry, and currently maintained at the University of Cambridge (see Meeks, Wheeler, and Whittington 1998, 1999). The accounting information in this database was collected year by year, which means that firms were not excluded with hindsight. Hence, the data should not suffer from survivor bias. However, the data covers only a limited sample of firms, mainly in the manufacturing sector. I construct a link to this database by matching companies manually by name. In the case of duplicate codes for different firms, or Cambridge/DTI codes changing over time, I reclassify these companies by assigning my own codes.

For all listed firms in a given year from 1953-1999 without Datastream or Cambridge/DTI data I handcollect balance sheets from hardcopy issues of the Stock Exchange Yearbook (the "Yearbook"). The Yearbook is an official publication by the London Stock Exchange. It contains essential information on each issuing firm or institution with securities listed on the LSE. For firms with officially listed shares the Yearbook reports an outline of the most recent balance sheet.

3.2 Handcollected data

Datastream and Cambridge/DTI offer very comprehensive accounting information, but for many firms the balance sheet provided in the Yearbook only contains a few items. However, the balance sheets are always complete in the sense that the items add up to total assets or liabilities. There are also firms for which data is provided at a greater level of detail. To keep the task manageable I categorize the data based on the minimum level of detail that is typically available for industrial companies. This classification is presented in table I, panel A. For banks I use a slightly simpler categorization, as current and non-current liabilities and assets cannot be distinguished properly based on the information given in the Yearbook. Panel B provides the outline used for banks. Appendix A discusses in more detail some of the accounting issues. Appendix B describes the checks for consistency and accuracy that were performed on the data during and following the collection process.

3.3 Coverage

This subsection examines the coverage of previously available data sets and of the new merged and supplemented database. To do this, I need to make an assumption on how the accounting information will be used. In the following I assume that book value of equity is used to form portfolios, like, for example, in Fama and French (1993). In particular, at the end of June each year *t* I check whether book value of equity is available for the fiscal year ending in year *t-1*, where book value of equity is defined as ordinary share capital plus reserves plus total deferred and future taxation. If data is not available in a particular year due to a fiscal year that is longer than 12 months I use the year *t-2* book value.

Figure 1 presents the coverage of the whole database and different subsamples. The total area represents the universe of all listed firms as defined above. The chart shows that the sample of firms with accounting data on Datastream is quite small until the mid-70s. The merged Cambridge/DTI-Datastream sample still misses many firms. However, when the handcollected sample is added, the coverage is very close to being complete. The number of firms for which I could not find data in the Yearbooks is very small, especially if I allow the use of year *t-2* accounting information in case year *t-1* information is not available.[6] Missing data means that the data was not found in the handcollection process. Therefore, whether a firm-year is missing in the full database only depends on the characteristics of the handcollected sample. In general, the coverage is remarkably high compared to COMPUSTAT, where about half of CRSP firms do not have accounting data in the 1960's (Kim 1997).

Figure 1 also reveals an interesting difference to US data sets. The Davis, Fama, and French (2000) COMPUSTAT/Moody's data set, which is the most extensive one available for the US, provides accounting information for 834 NYSE firms in June 1956, and the sample grows to 4,562 NYSE/AMEX/NASDAQ firms by 1996. In contrast, the new UK data set covers more than 3,000 firms in 1955, but it shrinks to less than 1,500 firms in 1999. As pointed out above, the LSPD currently does not provide share price information for all these firms in the pre-1975 period. However, when the LSPD random sample is used, there are still about 1,000 firms in 1955 for which accounting and share price information is available.

Panel A of table 2 provides information on coverage in terms of firm-years. It shows that the addition of the Cambridge/DTI and handcollected data extends the

[6] The spike in the number of missing firms in 1969 is discussed in the next section.

coverage considerably in the first two subperiods. The percentage of missing firm-years in the full database is low. Ranging from 7% to 9%, it is quite similar across subperiods. It would be very different without the handcollected data. Panel B reports the share of aggregate market capitalization that is captured by each of the subsamples. In later periods Datastream covers a relatively high share, but even in the latest subperiod the handcollected sample makes a contribution in terms of aggregate market capitalization. The fraction of market value missed by the combined database is negligible.

4. Examination of Missing Data

The reasons for missing data in the new database are determined by the way in which accounting information is reported in the Yearbook. Basically, the Yearbook covers all firms with official listing on the London Stock Exchange in a given year. However, there are a few circumstances in which accounting information is not available. This section provides an investigation of these issues. To put the following discussion into perspective, however, it is important to remember that the proportion of missing observations in this database is very low - lower than in any other large database of accounting information that I am aware of.

4.1 Delistings

When a firm delists from the LSE, the issue of the Yearbook following the delisting still reports some information on this firm, but often no accounting information.[7] As a result, I frequently miss data from the fiscal year preceding the delisting. The delisting problem may be particularly relevant when firms are grouped

[7] The same may happen when the trading of shares is suspended.

according to size, as delistings tend to be concentrated among small firms (Kothari, Sabino, and Zach 2001, Stolin 2001).

The first row of table 3 shows that delistings are the reason for about 12% of firm-years that are missing in the database. More precisely, in 12% of the cases where data is missing the firm will be delisting within 18 months following the point in time when I check for data availability. Table 3 also shows that the extent to which delistings cause a missing data problem depends on the length of the lag that is allowed for accounting information. This suggests a remedy to avoid a delisting bias. For relatively slow moving variables like book value of equity or total assets longer lags may be feasible. The second and third rows of the table show that if the use of year *t-1* or *t-2* accounting information is allowed, the problem is alleviated. Hence, whenever feasible, allowing longer lags is recommended to avoid delisting bias. Another partial remedy is to apply delisting returns on the last trading day, which can be months or even years earlier than the delisting day. This also increases the chance that recent accounting information is available prior to each return observation.

4.2 Initial Listings

The manually collected sample usually misses data prior to the initial listing of a firm's stock. The Cambridge/DTI data has pre-IPO information in some, but not all cases. Datastream regularly misses the pre-listing information, too. Table 3 shows that new listings account for 28% of the firm-years with missing data. Hence, a research design that requires lagged accounting information would systematically exclude many newly listed stocks. However, this is a case of ex-ante selection bias. There is less reason for concern since there is no exclusion of firms with hindsight. For an investor trying to replicate strategies tested in the data it is always feasible to exclude firms in their first year of listing. Nevertheless, researchers should be aware of this feature of the data.

4.3 Other Reasons

In some cases the Yearbook published in year *t* reports the balance sheet of the fiscal year ending in year *t-2*, while the Yearbook in year *t+1* contains the balance sheet of fiscal year *t*, i.e. the year *t-1* data is not available. This problem is particularly present in 1969, and it is the reason for the spike in the number of missing firms observed before in figure 1. This is probably due to changes in the publication dates of the Yearbook or changes in reporting times that lead to a gap year for many firms. As in the delisting case above, this problem can be mitigated by using lagged information from the preceding fiscal year. Figure 1 shows that when the use of year *t-1* or *t-2* is allowed, the spike in 1969 disappears, and the number of firms for which information is missing declines substantially.

In a few cases I may also have failed to find the information on a particular company, for example due to a failure to correctly track name changes. The occurrence of these problems is likely to be random and should have no effect on the quality of the data. Finally, in some instances available balance sheet information was not collected on purpose, for example, when denominated in foreign currency. To get some more information on how much these different reasons contribute to the missing data problem, table 4 presents a detailed analysis of missing information in the portfolio formation year 1998 (i.e. fiscal years ending in 1997). I track the reasons for missing information for each firm with missing data. In total data is missing for 79 firms. As can be seen in the table, the bulk of it can be attributed to new listings in the preceding 18 months. The rest of it is accounted for by delistings and the other reasons I discussed above. The findings in table 4 suggest that, except for the new listings problem, there is not much reason to be concerned about selection bias problems in this data.

5. Evidence on selection biases

The new data set with accounting data for virtually the whole population of listed firms on the LSE presents an opportunity to investigate selection bias effects in previously existing databases. It is important to note that the direction of the effects of selection biases are not clear a priori. Survivor biases are usually belived to cause an upward bias in returns, but there may also be industry, size and other biases. The total effect of these can only be determined by comparing the biased sample with the underlying population.

5.1 Sample selection effects on average firm characteristics

Table 5 compares the characteristics of the subsamples that make up the new merged database along some important dimensions. Panel A reports the average relative market capitalization of stocks in each subsample for various time periods. Each year at the end of June I compute the equal-weighted average market capitalization for the whole population of listed stocks (L) (pre-1975 I use the LSPD random sample). This average market cap is then used to standardize each firm's market cap, which makes it more comparable over time than simple market cap. Averaging these standardized market caps within each subsample and time period yields the results reported in Panel A. It is evident that the Datastream sample is tilted towards larger stocks, particularly in the 1966-1977 period. In contrast, the average market cap of firms in the Cambridge/DTI database is only half of the population average. Firms from the handcollected sample tend to be relatively small, albeit larger than the average Cambridge/DTI firm. The few firms for which I did not find accounting information tend to be very small firms, which is consistent with the finding from above that many of these are newly listed firms.

Panel B presents average book-to-market ratios, standardized in similar manner as market capitalizations in Panel A. It shows that both Datastream and the Cambridge/DTI database tend to miss firms with relatively high book-to-market ratios. Firms in the handcollected sample have the highest book-to-market in every subperiod. Hence, the results in table 5 suggest that when only Datastream is used, or Datastream combined with the Cambridge/DTI database, the sample of firms tends to be biased against small firms and high book-to-market stocks. Since it is a well documented fact that these firm characteristics are associated with substantial cross-sectional variation in average returns, there is reason to suspect that this selection bias could affect returns on certain portfolios and investment strategies. Furthermore, small and high book-to-market stocks are more likely to delist than other firms (Kothari, Sabino, and Zach 2001). A bias against these firms could therefore be a symptom of survivor bias.

5.2 Sample selection effects on estimated moments of returns

Kothari, Sabino, and Zach (2001) find that non-surviving firms tend to be either extremely bad or extremely good performers. Survivor bias implies truncation of such extreme observations. The authors show that even a small degree of such non-random trunction can have a strong impact on sample moments of stock returns. In table 6 I therefore examine the effects of sample selection on some important summary statistics of returns.[8] Each year these summary statistics are computed from monthly stock returns of all firms in a given sample during that year. The numbers in the table are time-series averages of these annual statistics.[9] Panel A compares the population of listed stocks (L) with the full sample for which accounting data is available (D+C+H) and the subsample without the handcollected data (D+C). As one would expect, given the almost complete overlap, the statistics for the D+C+H sample are almost identical to those of the population L. Restricting the sample to D+C however produces much lower standard deviation, skewness, and kurtosis. These are precisely the symptoms of truncation of extreme observations suggested by the simulation results in Kothari, Sabino, and Zach (2001). Hence, previously existing databases appear to be biased against extreme performers.

Panel B investigates the effect of restricting the sample to firms on Datastream (D) for the period 1969-2000.[10] It shows that the underestimation of standard deviation, skewness, and kurtosis are not artifacts of the data from the 50's and 60's. It also applies to the widely used Datastream sample. Panel C indicates that these problems originate mainly from the period 1969-1979 when Datastream coverage is relatively poor and data has been backfilled. Summary statistics for later periods (not shown) are very similar to the full sample. All three Panels also point to an apparent

[8] In return calculations I pay particular attention to delisting returns. I follow Shumway (1997) and approximate correct delisting returns by adjusting the return on the last day of trading. Shumway suggests a particular correction for CRSP NYSE stocks, based on CRSP delisting codes. Of course, I cannot apply this correction straightforwardly to the LSPD. Hence, I use my own approximation. For all delistings for which the delisting code in the LSPD suggests that the stock delisted valueless I assume a delisting return of –100%. I also check the LSPD delisting codes for a small random sample of stocks against the UK Capital Gains Tax Book, which contains information on final distributions. Based on this investigation I believe that my approximation does a good job of capturing the wealth effects experienced by an investor holding delisted stocks. Qualitatively, the results presented below are not sensitive to the approximation method.

[9] Without averaging statistics across years, differences in statistics could arise from a different distribution of observations across year, coupled with statistics changing over time.

[10] Datastream is compared to the full sample only for the 1969-2000 period since Datastream does not offer coverage before 1968. The first possible portfolio formation date therefore is June 1969.

underestimation of mean returns when previously existing databases are used (C and/or D). The next subsection takes a closer look at this.

5.3 Sample selection effects on portfolio returns

Table 7 compares the returns on portfolios that utilize the full data set with those that result when the sample is restricted to firms covered in previously existing databases. I compute equal-weighted returns on various portfolios based on the entire database (D+C+H) and on subsamples (D+C) or (D). The raw returns, as well as the mean difference in these portfolio returns and associated t-statistics are reported in table 7. In Panel A I form a simple equal-weighted index. Panels B and C report returns on size and book-to-market portfolios, respectively. The latter portfolios are formed based on quintile breakpoints derived from the full sample (D+C+H). These full sample breakpoints are also applied to the subsample that is to be compared with the full sample, which follows the Banz and Breen (1986) tests for selection bias in the COMPUSTAT database. The portfolios are rebalanced annually at the end of June.

The results in Panel A show that without inclusion of the handcollected sample the equal-weighted index return is significantly understated. The firms missed by Datastream and the Cambridge/DTI database tend to have higher returns. The difference is smaller and only marginally significant, however, when only Datastream is compared to the full database for the period 1969-2000, as shown in the seond set of columns. In general, there is a preponderance of negative signs in all three panels of the table. Forming size portfolios in panel B does not eliminate return differences. Hence, it can not be attributed to a mere size bias. Interestingly, the return differences appear to originate mainly from the smallest size group. The differences related to

Datastream-only size portfolios in the second set of columns all have negative signs, but they are not statistically significant. Panel C examines return differences for book-to-market portfolios. Here a clear pattern emerges. Both Datastream and the Cambridge/DTI database tend to miss some medium to high book-to-market stocks with high returns. The addition of the handcollected data significantly increases returns for book-to-market portfolios 3 and 4.

Thus, a researcher who requires accounting information for firms in his sample systematically misses some well performing stocks if she does not use the handcollected sample to supplement existing databases. Given the results above, the problem appears to be most severe for investigations that focus on very small firms, or on stocks with relatively high book-to-market ratios.

So far I compared previously existing databases (D, C) with the full database (D+C+H). The few firms still missed by the full database are not likely to cause any significant bias in returns in most research designs. However, to check this I repeat above tests, and compare the LSPD population to the sample of firms for which I have accounting information from any source (D+C+H). Differences indeed turn out to be negligible. For example, in the case of five size portfolios the differences in mean returns are all lower than half a basispoint per month. None of the differences is significant at any conventional significance level, there is no uniform pattern in signs, and they are not jointly significant.[11] To conserve space I do not report the full results.

[11] To test for joint significance I run SUR regressions as in Banz and Breen (1986), which account for covariation of return differences across size groups.

5.4 Discussion

The result that firms missed in existing databases tend to have higher mean returns may seem somewhat surprising.[12] In the case of survivor bias one might expect to find those firms underperforming instead. However, the situation is more complex. First, the return implications of survivor bias are not clear a priori. Attrition of firms from exchanges not only occurs through bad performance and distress but also, for example, through takeovers and mergers. Second, survivor bias is certainly not the only form of selection bias with possible impact on returns of firms in these databases. I have presented some evidence for size and book-to-market tilts, but there are potentially many more, including industry biases, new listing biases etc.[13]

Perhaps more important than the mean returns evidence, depending on the research design, is the finding that existing databases miss firms that exibit performance in the tails of the return distribution. The effects on sample standard deviations, skewness, and kurtosis presented here are strikingly similar to the simulation results of Kothari, Sabino, and Zach (2001) where extreme observation are truncated deliberately. The evidence presented here complements their work by showing that the effects they generate in simulations do indeed arise in widely used databases.

[12] Interestingly, Banz and Breen (1986), who test for selection bias effects in a version of COMPUSTAT database that excludes firms that have disappeared, also find a higher return for stocks that are missed by the database.

[13] An additional caveat is in order. It is well documented that in earlier decades many very small stocks on the London Stock Exchange feature thin trading and stale prices that can impact return statistics even at monthly frequencies (see e.g. Dimson and Marsh 1983). These microstructure problems could be another reason for the higher return observed for stocks in the handcollected sample. For example, consider a price sequence 1, 1.05, 1.125 for a stock where trading is frequent. An otherwise similar, but thinly traded stock may have a price sequence 1, 1, 1.125. The arithmetic average return for the latter one is 6.25%, whereas it is 5% for the first one, even though the fundamental value was identical at each point in time. For this reason I also run tests where I sort stocks on trading frequency. However, it turns out that, after controlling for trading frequency, stocks in the handcollected sample still tend to have higher returns.

6. Conclusions

This paper introduces a new data set of accounting information for UK firms. I merge two existing databases, supplement them with handcollected data, and obtain a total of about 100,000 firm-years of key balance sheet items. The data set covers virtually all firms that have been listed on the London Stock Exchange between 1953 and 1999, and it is fully cross-referenced to the London Share Price Database. Most importantly, it is free of the selection biases commonly present in databases of accounting information.

Based on the new data set I provide evidence on the effects of selection biases in previously existing UK databases. The firms missed in these data sets tend to be small and have high book-to-market. Excluding them produces a downward bias in mean returns, particularly for medium to high book-to-market portfolios. Furthermore, standard deviations, skewness, and kurtosis of monthly returns are understated. This suggests that existing databases miss extreme performers in both tails of the return distribtion. The evidence closely mirrors the simulation results in Kothari, Sabino, and Zach (2001), where extreme observations are truncated deliberately. At least a part of these effects can be attributed to the survivor bias inherent in Datastream data until the end of the 1970's.

Whether these biases in firm characteristics and sample moments should be reason for concern depends on the research design. Particular caution is warranted when inference is sensitive to biases in second and higher moments, as, for example, in long-horizon abnormal return measurement. Moreover, it seems unlikely that the selection biases documented here are confined to UK databases. Little is known about the firms missed by COMPUSTAT in the 60's and 70's. Data sets from other countries probably suffer rather more than less from these problems. Hence, the results in this paper illustrate a general problem in databases of accounting information that researchers should be aware of.

The absence of selection bias in the new data set allows researchers to avoid these problems. Of course, the number of data items offered per firm in my handcollected sample is just a fraction of what is offered on COMPUSTAT, but even so, the data is very useful, for example, to match firms or to form portfolios based on accounting characteristics such as book-to-market ratios. By enabling researchers to apply such standard methodology, the new data set also opens a way for testing of empirically motivated models with a credible and extensive non-US data set. Hence, it helps to alleviate the problem of repeated mining of the US data.

Appendix A: Accounting Issues

In the majority of cases, categorizing the balance sheet data from the Yearbook into the outline presented in table 1 is very straightforward. However, in some cases it is less trivial and involves some judgment. This usually occurs for balance sheet items that are special to certain time periods, specific industries, or particular situations. To standardize this process as much as possible I conducted a pilot study in which I collected data for a whole cross-section of firms in one year. This was the basis for the development of a classification scheme for all balance sheet items for which the appropriate classification is not self-evident.[14] However, some problem areas remain.

Usually the distinction between equity and liabilities is clear from the data. For this reason information on book value of equity, total assets, and leverage should be reliable. The distinction between current and non-current assets and liabilities however is not always clear. For example, some firms report current liabilities and provisions in one item. Frequently, small firms report current liabilities as their only liability item. It may be that this is the result of some simplification in reporting or in recording by the publishers of the Yearbook. Thus, the categorization into current and non-current should be regarded as rather tentative.

For banks the classification into current and non-current is particularly troublesome. I therefore do not make this distinction for banks and classify all financial assets as investments and all non-financial assets as fixed assets. Frequently, banks' reserves are merged with some liabilities accounts. Therefore, shareholders' equity of banks can be severely understated and should be interpreted with caution.

[14] I am grateful to Chris Higson for working with me on this classification scheme.

Insurance companies also require some special treatment. Usually, insurance companies feature lots of different financial assets on the asset side. These are classified as investments as long as they are not labeled as fixed, other, or current. On the liability side insurance companies are very special with respect to the distinction between equity and liabilities. I classify any reserve funds as reserves and insurance funds as long-term liabilities. Apparently, starting in 1974, insurance companies were allowed to hide some of their reserves in insurance funds, which makes the task of figuring out a meaningful number for reserves basically impossible. Hence, book value of equity figures for insurance companies should also be used with caution.

Waterworks, mainly in the 1950s and 1960s, only report an aggregate number for share capital and long-term debt. In this case I calculate a proxy for share capital by deducting the outstanding amounts of debentures as reported in the Yearbook from this aggregate amount. I put the total amount of debentures into long-term debt.

In general, the balance sheet provided in the Yearbook is a consolidated balance sheet. However, in some few cases foreign subsidiaries are not consolidated. In this case I take the non-consolidated parent balance sheet as the best available approximation.

Appendix B: Tests for Consistency and Accuracy

Several test have been applied to the handcollected data to check for consistency and accuracy. Identifying data entry errors is made somewhat difficult by the fact that the data in the Yearbook is also not free from errors. During the data entry process the total sums of assets and liability side of each balance sheet were checked for consistency, with double-checking in cases of divergence. In a separate step, after completion of all data entry, all differences were scaled by book equity. Differences that amounted to more than 50% of book equity were checked again against the Yearbook and the balance sheet was compared to preceding and subsequent balance sheets. In a few cases this process revealed previously undetected inconsistencies with the Yearbook. Most importantly however, the comparison in the time series allowed to identify cases where one balance sheet item was missing in the Yearbook (e.g. sometimes nominal capital was missing in a particular year). In cases where the difference appeared to be due to such a missing item with reasonable certainty I attributed the difference to the missing item. This left only very few differences over 50% of book value of equity.

Checks of consistency of balance sheets over time also helped to identify misidentified companies. During data entry companies were looked up in the Yearbook by their name, taken from of the LSPD master index. There are cases when companies have very similar names and data for the wrong company had been entered. By checking for increases in nominal capital from one period to the next, followed by a subsequent decrease, I filtered out suspicious cases. I then checked these cases again against the Yearbook. This method also allowed to filter for cases were the number of digits (i.e. thousands or millions) was not correct. Furthermore, it also allowed to check that the spreadsheets used for data entry had not been corrupted in some way that would have destroyed the link between company identifiers, names, and data.

References

Banz, R., Breen, W.J., 1986, Sample-dependent results using accounting and market data: some evidence. Journal of Finance 41, 779-793.

Barber, B.M., Lyon, J.D., 1997, Detecting long-run abnormal stock returns: the empirical power and specification of test statistics. Journal of Financial Economics 43, 341-372.

Barberis, N., Shleifer, A., Vishny, R., 1998, A model of investor sentiment. Journal of Financial Economics 49, 307-343.

Daniel, K.D., Hirshleifer, D., Subrahmanyam, A., 1998, Investor psychology and security market over- and underreaction. Journal of Finance 52, 1839-1885.

Daniel, K.D., Titman, S., 1997, Evidence on the Characteristics of Cross-Sectional Variation in Stock Returns. Journal of Finance 51, 1-33.

Daniel, K.D., Titman, S., Wei, K.C., 2001, Explaining the Cross-Section of Stock Returns in Japan: Factors or Characteristics? Journal of Finance, 55, 743-766.

Davis, J.L., 1996, The cross-section of stock returns and survivor bias: Evidence from delisted stocks. Quarterly Review of Economics and Finance 36, 365-375.

Davis, J.L., Fama, E.F., French, K.R., 2000, Characteristics, covariances, and average returns: 1929-1997. Journal of Finance 55, 389-406.

Dimson, E., Marsh, P.R., 1983, The stability of U.K. risk measures and the problem of thin trading. Journal of Finance, 38, 753-783.

Dimson, E., Marsh, P.R., 2001, U.K. financial market returns 1955-2000. Journal of Business. 74, 1-30.

Dimson, E., Nagel, S., Quigley, G., 2001, Value versus growth in the UK stock market 1955-2000. Working paper, London Business School.

Fama, E.F., French, K.R., 1993, Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3-56.

Fama, E.F., French, K.R., 1998, Value versus growth: the international evidence. Journal of Finance, 53, 1975-1999.

Hong, H., Lim, T., and Stein, J.C., 2000, Bad news travels slowly: size, analyst coverage and the profitability of momentum strategies. Journal of Finance 55, 265-295.

Kim, D, 1997, A reexamination of firm size, book-to-market, and earnings price in the cross-section of expected stock returns. Journal of Financial and Quantitative Analysis 32, 463-489.

Kothari, S.P., Shanken, J., Sloan, R.G., 1995, Another look at the cross-section of expected stock returns. Journal of Finance 50, 185-224.

Kothari, S.P., Warner, J.B., 1997, Measuring long-horizon security performance. Journal of Financial Economics 43, 301-339.

Kothari, S.P., Sabino, J.S., and Zach, T., 2001, Implications of data restrictions on performance measurement and tests of rational pricing. Working paper, MIT.

Lo, A., MacKinlay, A.C., 1990, Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431-467.

Meeks, G., Wheeler, J.M., Whittington, G., 1998, The Cambridge/DTI Databank of company accounts: an introduction for users. Manual, Department of Applied Economics, University of Cambridge.

Meeks, G., Wheeler, J.M., Whittington, G., 1999, Cambridge/DTI databank of company accounts, 1948-1990. Computer file (2nd ed.), Department of Trade and Industry (original data producer), The Data Archive (distributor), Colchester, Essex.

Merton, R., 1987, On the state of the efficient market hypothesis in financial economics. In: Dornbusch, R., Fischer, S., Bossons, J. (eds.), Macroeconomics and finance: essays in honor of Franco Modigliani. MIT Press, Cambridge, 93-124.

Mitchell, M.L., Stafford, E., 2000, Managerial decisions and long-term stock price performance. Journal of Business, 73, 287-320.

Nagel, S., 2001, Is it overreaction? The performance of value and momentum strategies at long horizons. Working paper, London Business School.

Rouwenhorst, G., 1997, International momentum strategies. Journal of Finance, 53, 267-284.

Shumway, T., 1997, The delisting bias in CRSP data. Journal of Finance, 52, 327-340.

Stolin, D., 2001, Survivorship issues in share price research. Unpublished PhD Dissertation, London Business School.

Verbeek, M.,Nijman, T., 1995, Incomplete panels and selection bias. In: Matyas, L., Sevestre, P. (eds.), The econometrics of panel data: handbook of theory with applications (2nd ed). Kluwer, Dordrecht, Netherlands, 449-490.

Table 1

Balance sheet items collected from the Yearbook

Panel A: All firms

1	Year
2	Month of fiscal year end
3	Day of fiscal year end

Panel B: All firms except banks

	Assets		Liabilities
1	Intangible Assets	1	Ordinary Share Capital
2	Fixed Assets	2	Preference Share Capital
3	Investments	3	Reserves
4	Current Assets	4	Deferred and Future Taxation
5	Other Assets	5	Minority Interest
		6	Long-term Liabilities
		7	Current Liabilities
		8	Other Liabilities
		9	Unresolved Difference

Panel C: Banks

	Assets		Liabilities
1	Intangible Assets	1	Ordinary Share Capital
2	Fixed Assets	2	Preference Share Capital
3	Investments	3	Reserves
4	Other Assets	4	Deferred and Future Taxation
		5	Minority Interest
		6	Liabilities
		7	Other Liabilities
		8	Unresolved Difference

Table 2

Coverage of the new database

L is the total population of primary shares listed on the London Stock Exchange end of June each year *t* . In Panel A I check whether book value of equity is available for the fiscal year ending in calendar year *t-1* . D denotes firms for which book values are obtained from Datastream, C refers to the Cambridge/DTI database, and H is the handcollected data. Panel B reports the share of aggregate market capitalization captured by the samples each year, averaged across years.

Panel A: Number of firm-years

	Data source									
Period	L		D		C		H		missed	
1955-2000	98,697	100%	37,164	38%	27,879	28%	26,276	27%	7,378	7%
1955-1965	35,361	100%	-	0%	19,833	56%	12,950	37%	2,578	7%
1966-1977	29,403	100%	8,980	31%	8,017	27%	9,765	33%	2,641	9%
1978-2000	33,933	100%	28,184	83%	-	0%	3,590	11%	2,159	6%

Panel B: Relative aggregate market capitalization

	Data source				
Period	L	D	C	H	missed
1955-2000	100%	82.72%	31.80%	18.90%	1.93%
1955-1965	100%	-	50.61%	46.94%	2.42%
1966-1977	100%	69.07%	11.11%	17.75%	3.08%
1978-2000	100%	92.82%	-	6.04%	1.14%

Table 3

Reasons for missing data: Delistings, new listings, and others

For each firm I check the availability of book value of equity for the fiscal year ending in calendar year *t-1* in year *t-1* or *t-2*, or in year *t-1*, *t-2*, or *t-3*. New listings are those firms for which no accounting information is available and which obtained a new listing in the 18 months prior to end of June of year *t* Delistings are those which delist in the 18 months following June year *t*.

Accounting data required	Missing		Reasons					
			New listings		Delistings		Other Reasons	
in year t-1	6,889	100%	1,950	28%	843	12%	4,096	59%
in year t-1 or t-2	4,170	100%	1,950	47%	287	7%	1,933	46%
in year t-1, t-2, or t-3	3,752	100%	1,950	52%	200	5%	1,602	43%

Table 4

Detailed inspection of missing accounting data for the fiscal year ending in 1997

Total is the number of firms for which accounting information for the fiscal year ending in 1997 is not available from any source. All of these firms should have been looked up in the Yearbook. The table shows the reasons for non-availability or non-collection, obtained from double-checking the Yearbook after data collection was finished. AIM/USM refers to the alternative investment market and unlisted securities market, respectively.

Reason why missing	number of firms
Total	79
Not available in the Yearbook:	
New listings	56
Delistings	4
Fiscal year longer than 12 months	4
Recent promotion from AIM/USM	1
Recently converted investment trust	1
Available, but not collected:	
Balance sheet denominated in foreign currency	4
Available, but failure to collect:	
Failure to collect due to name change	1
Failure to collect for unknown reason	8
	79

Table 5

Characteristics of database subsamples

L, D, C, H are defined as in table 2. Panel A reports the average relative market capitalization, which is the equal-weighted average market capitalization in each sample each year divided by the equal-weighted average market capitalization in the population L in the same year. The numbers given in the table are simple time-series averages. In Panel B the average relative book-to-market ratio is standardized and averaged in the same way.

Panel A: Average relative market capitalization

	Data source				
Period	L	D	C	H	missed
1955-2000	1	1.27	0.54	0.74	0.37
1955-1965	1	-	0.85	1.33	0.58
1966-1977	1	1.70	0.48	0.76	0.42
1978-2000	1	1.10	-	0.43	0.25

Panel B: Average relative book-to-market ratio

	Data source			
Period	L	D	C	H
1955-2000	1.00	0.99	0.93	1.13
1955-1965	1.00	-	0.79	1.27
1966-1977	1.00	0.93	0.86	1.09
1978-2000	1.00	1.01	-	1.08

Table 6
Sample selection effects on summary statistics of monthly returns

Population L and samples D, C, H are defined as in table 2. Each year summary statistics are calculated from individual stocks' monthly returns. The table reports the time-series averages of these yearly summary statistics. Panel A shows the results for the whole period, Panels B and C depict subperiod results. Mean, median, and standard deviation is given in percent per month.

Sample	Mean	Median	Standard deviation	Skewness	Kurtosis
Panel A: 1955-2000					
All stocks (L)	1.68	0.07	13.97	6.93	286.62
D+C+H	1.67	0.08	13.80	6.28	244.42
D+C	1.59	0.16	12.52	3.08	60.15
Panel B: 1969-2000					
All stocks (L)	1.70	0.09	14.47	3.74	75.33
D+C+H	1.70	0.09	14.42	3.58	71.91
D	1.63	0.19	13.40	2.76	54.01
Panel C: 1969-1979					
All stocks (L)	1.96	-0.09	14.91	4.35	82.78
D+C+H	1.94	-0.11	14.75	3.91	71.72
D	1.70	0.06	12.24	1.53	17.31

Table 7
Sample selection effects on portfolio returns

Returns from June in year t to July in year t+1 are matched to book values of the fiscal year ending in year t-1 (data from year t-2 or t-3 is substituted if not available in year t-1). Samples D,C, and H are defined as in table 2. Equal-weighted portfolio returns are computed for a restricted sample that uses only previously existing databases (D or D+C), and for the full sample that includes the handcollected data set (D+C+H). The table shows the return of the restricted sample portfolios minus the return of the full sample portfolios. Portfolio breakpoints in Panels B and C are always set at full sample (D+C+H) quintiles.

	D + C minus D + C + H				D minus D + C + H			
	1955-2000				1969-2000			
	D+C	D+C+H	difference	t-statistic	D	D+C+H	difference	t-statistic
Panel A: Monthly equal-weighted index returns (in percent)								
	1.609	1.679	-0.070	(-3.02)	1.652	1.714	-0.062	(-1.87)
Panel B: Monthly size portfolio returns (in percent)								
small	2.173	2.340	-0.167	(-2.17)	2.270	2.313	-0.042	(-0.47)
2	1.703	1.697	0.006	(0.19)	1.669	1.691	-0.022	(-0.48)
3	1.513	1.493	0.021	(0.92)	1.512	1.541	-0.029	(-0.83)
4	1.410	1.437	-0.028	(-1.31)	1.471	1.501	-0.030	(-0.95)
large	1.360	1.381	-0.021	(-0.98)	1.468	1.487	-0.020	(-0.81)
Panel C: Monthly book-to-market portfolio returns (in percent)								
low	1.191	1.205	-0.014	(-0.40)	1.239	1.226	0.013	(0.28)
2	1.381	1.418	-0.037	(-1.53)	1.441	1.453	-0.012	(-0.36)
3	1.554	1.621	-0.067	(-2.35)	1.553	1.683	-0.130	(-3.05)
4	1.782	1.908	-0.127	(-3.70)	1.803	1.963	-0.160	(-3.40)
high	2.103	2.196	-0.094	(-1.65)	2.181	2.206	-0.025	(-0.35)

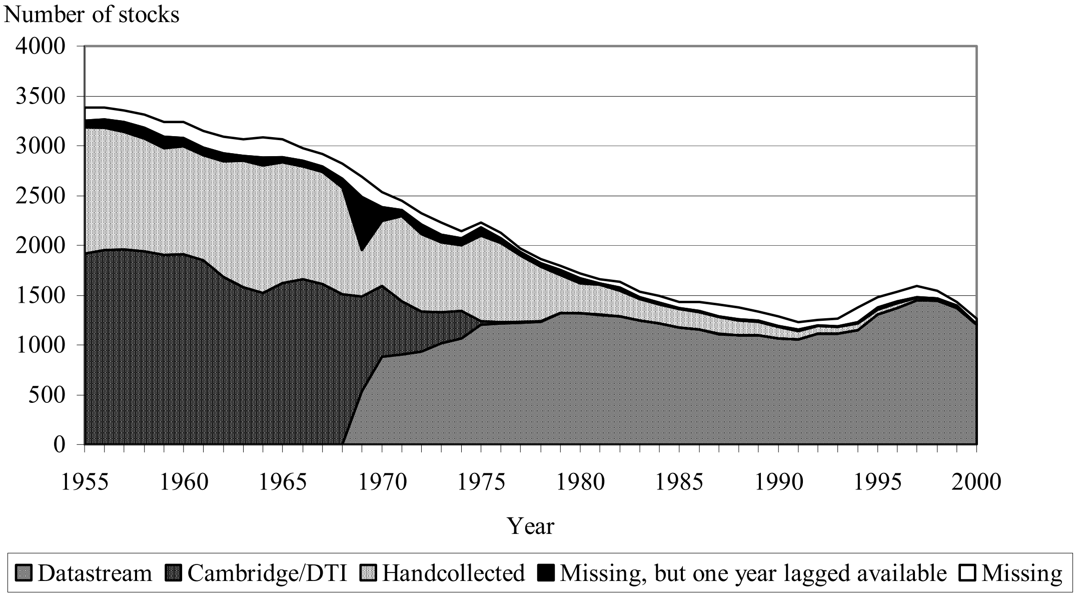

Fig. 1. Coverage of the new database and its subsamples. Definitions as explained in table 2. The total area represents the whole population of firms with stocks listed on the main market of the London Stock Exchange. For firms where data for the fiscal year ending in year *t-1* is not available I check whether data is available for year *t-2*. Firm-years for which this is the case are represented by the black area.