11-2020

Data Quality Problems Troubling Business and Financial Researchers: A Literature Review and Synthetic Analysis

Grace Liu
West Chester University of Pennsylvania, yliu@wcupa.edu

Recommended Citation

Liu, G. (2020). Data Quality Problems Troubling Business and Financial Researchers: A Literature Review and Synthetic Analysis. *Journal of Business & Finance Librarianship*, 1-47. http://dx.doi.org/10.1080/08963568.2020.1847555

The Version of Record of this manuscript has been published and is available in Journal of Business & Finance Librarianship, Nov 19, 2020, DOI:10.1080/08963568.2020.1847555

Data Quality Problems Troubling Business and Financial Researchers: A Literature Review and Synthetic Analysis

*Grace Liu, West Chester University, PA**

Address Correspondence to Grace Liu, Assistant Professor, Business and Faculty Services Librarian, West Chester University, FHG Library, 25 W Rosedale Ave, West Chester, PA 19382, USA. E-mail: yliu@wcupa.edu

** Acknowledgment - I would like to express my special thanks to Jim Kelly, Instruction and Liaison Librarian - Business at the University of Northern Iowa, for the early discussions with me on this topic in 2017-2018. I would also like to show my gratitude to Todd Hines, the Manager of Research & Discovery and Research Subject Librarian at the Stanford GSB library for sharing his pearls of wisdom with me about this topic. I thank guest editors and the anonymous reviewer for their insights and comments on this paper. Lastly, I would like to thank my family for their support that allows me to spend the whole summer on this project during the pandemic.*

Abstract

The data quality of commercial business and financial databases greatly affects research quality and reliability. The presence of data quality problems can not only distort research results, destroy a research effort but also seriously damage management decisions based upon such research. Although library literature rarely discusses data quality problems, business literature reports a wide range of data quality issues, many of which have been systematically tested with statistical methods. This article reviews a collection of the business literature that provides a critical analysis on the data quality of the most frequently used business and finance databases including the Center for Research in Security Prices (CRSP), Compustat, S&P Capital IQ, I/B/E/S, Datastream, Worldscope, Securities Data Company (SDC) Platinum, and Bureau Van Dijk (BvD) Orbis and identifies 11 categories of common data quality problems, including missing values, data errors, discrepancies, biases, inconsistencies, static header data, standardization, changes in historic data, lack of transparency, reporting time issues and misuse of data. Finally, the article provides some practical advice for librarians to facilitate their scholarly communications with researchers on data quality problems.

Introduction

Business and finance databases are crucial for academic business research. Each year, thousands of empirical research articles are published based on the data from these databases.[1] The quality of the data can have a great impact on research quality and reliability. The presence of data quality problems can not only distort research results, destroy a research effort but also seriously damage management decisions based upon such research (Rosenberg & Houglet, 1974). Data quality issues are also relevant to business librarianship since evaluating information quality is an integral part of business information literacy. Anecdotally, many business librarians

[1] It is hard to estimate the volume of research conducted using business databases, but a conservative estimation with Google Scholar shows that there are over six thousand articles published in 2019 mentioned Compustat alone.

have encountered data quality problems, but very few have documented or discussed data problems in library literature. The only article found in the business library literature that tested and discussed data quality problems is titled "An issue of trust: are commercial databases really reliable?" by Cook et al. (2012), which commented on the ReferenceUSA's New Businesses database. Comparatively, since the early 1970s, business literature has reported a wide range of data quality problems and many of these problems have been systematically tested with statistical methods. Business librarians can greatly benefit from these studies. But the sparse coverage of business literature on databases and data quality problems makes it difficult to gain a holistic perspective. This article makes the first effort to provide a literature review and synthetic analysis over nearly a half-century of business research, trying to identify the general data quality problems. Hopefully, this research will help business librarians understand data quality problems more thoroughly and further inspire discussions on the role that librarians can play in improving data quality and safeguarding research integrity and public trust in business knowledge.

Research Questions

This article reviews a collection of the business literature that provides a critical analysis of the data quality of the most frequently used business and finance databases including the Center for Research in Security Prices (CRSP), Compustat, S&P Capital IQ, I/B/E/S, Datastream, Worldscope, Securities Data Company (SDC) Platinum, and Bureau Van Dijk (BvD) Orbis. This article attempts to find answers to the following questions:
a) what are the most common data quality problems?
b) how prevalent are these problems?
c) how are these problems identified?
d) what causes these problems?
e) what are the consequences of these problems? and
f) how to potentially solve these problems?

Methodology

In order to identify the articles that provide critical analysis on the data quality of the most frequently used business and finance databases, we searched the titles of these databases in Business Source Complete, Web of Science, and Google Scholar.[2] The specific search terms and the number of articles retrieved are listed in Table 1.[3] Most of the articles covered in this study were retrieved from Google Scholar.[4] A recent informetric study

[2] This search assumes that every article providing a critical analysis on a specific business database would mention the name of the database in its index, probably in the article title, abstract, or keywords. It also assumes that the more frequently the name of the database appears in an article, the more likely the article offers a discussion on the data quality of the database.
[3] Different search terms have different effects on the precision of the search results. "Compustat" and "I/B/E/S" as search terms are very effective in retrieving the relevant articles that mention these databases. Comparatively, "CRSP" and "Datastream" are less effective and retrieved many results in other areas. In these cases, we combined the database title with other search terms including "data", "database" or the publisher to increase the precision of the search results.
[4] Although Google Scholar doesn't disclose their search algorithm, our searches found that in some instances, Google Scholar may be able to search the full-text of an article and its "relevance" ranking criterion considers this factor. In Google Scholar, an article is

The Version of Record of this manuscript has been published and is available in Journal of Business & Finance Librarianship, Nov 19, 2020, DOI:10.1080/08963568.2020.1847555

on the major index services, which investigated over two million citations to over two thousand highly-cited documents, showed that Google Scholar consistently found the largest percentage of the citations across all areas (93%–96%), far ahead of Scopus (35%–77%) and Web of Science (27%–73%) (Martín-Martín et al., 2018). As Table 1 indicated, our searches confirmed this finding. Besides the highly-cited journal articles, Google scholar searches also retrieved more working papers, conference papers, dissertations, book chapters, and unpublished manuscripts than Business Source Complete and Web of Science. Since the discussion on data quality is often a byproduct of empirical research, many researchers choose to disclose the data quality issues in the forms of working papers, technical notes, manuscripts, or worksheets. Google scholar better meets the research needs to discover this grey literature.

TABLE 1: Search Terms and the Number of Articles Retrieved for Each Search Term

Reviewed Database	Business Source Complete (Title Search)	Business Source Complete (Index Search)	Business Source Complete (Full-text search)	Web of Science Core Collection Index (Advanced Search in title, abstract and author keywords)	Google Scholar (Title Search)	Google Scholar (Basic Search)
			Search Terms (Number of Articles Retrieved)			
CRSP	CRSP (13) "Center for Research in Security Price" (0)	CRSP (173) "CRSP data" (15) "CRSP database" (13) "Center for Research in Security Price" (0)	CRSP (9,813) "CRSP data" (1,260) "CRSP database" (1,576) "Center for Research in Security Price" (31)	TI=(CRSP) or AB=(CRSP) or AK=(CRSP) (290) TI=(CRSP data) or AB=(CRSP data) or AK=(CRSP data) (104) TI=(CRSP database) or AB=(CRSP database) or AK=(CRSP database) (34) TI=(Center for Research in Security Price) or AB=(Center for Research in Security Price) or AK=(Center for Research in Security Price) (97)	allintitle:"CRSP" (399)	"CRSP" (79,100) "CRSP data" (8,860) "CRSP database" (11,300) "Center for Research in Security Price" (237)
Compustat	Compustat (18)	Compustat (457)	Compustat (11,842)	TI=(Compustat) or AB=(Compustat) or AK=(Compustat) (377)	allintitle:"Compustat" (55)	"Compustat" (71,000)
Capital IQ	"Capital IQ" (2)	"Capital IQ" (29)	"Capital IQ" (458)	TI=(Capital IQ) or AB=(Capital IQ) or AK=(Capital IQ) (100)	allintitle:"Capital IQ" (8)	"Capital IQ" (6,680) S&P "Capital IQ" (4,200) Standard & Poor's "Capital IQ" (3,110)
I/B/E/S	I/B/E/S (6)	I/B/E/S (82)	I/B/E/S (1,495)	TI=(I/B/E/S) or AB=(I/B/E/S) or AK=(I/B/E/S) (53)	allintitle:"I/B/E/S" (16)	"I/B/E/S" (14,700)
Datastream	Datastream (6)	Datastream (106)	Datastream (7,040) Datastream AND Thomson (1,726)	TI=(Datastream) or AB=(Datastream) or AK=(Datastream) (277)	allintitle:"Datastream" (50)	"Datastream" (64,800) "Datastream database" (5,510) "Datastream" AND Thomson (20,700)

considered more relevant when the search term appears in the title of the article or when the search term appears in the abstract or the text of the article more frequently.

Worldscope	Worldscope (0)	Worldscope (18)	Worldscope (1,126) Worldscope AND Thomson (417)	TI=(Worldscope) or AB=(Worldscope) or AK=(Worldscope) (16)	allintitle:"Worldscope" (5)	"Worldscope" (9,830) "Worldscope database" (3,160) "Worldscope" AND Thomson (4,100)
SDC Platinum	"SDC Platinum" (0) SDC database (1)	"SDC Platinum" (12) SDC database (18)	"SDC Platinum" (747) "SDC database" (663) SDC Mergers and Acquisitions Database (326)	TI=(SDC Platinum) or AB=(SDC Platinum) or AK=(SDC Platinum) (27) TI=(SDC Mergers and Acquisitions Database) or AB=(SDC Mergers and Acquisitions Database) or AK=(SDC Mergers and Acquisitions Database) (10)	allintitle:"SDC Platinum" (0) allintitle: SDC database (9)	"SDC Platinum" (6,540) "SDC database" (5,140) "SDC Mergers and Acquisitions Database" (852)
BvD Orbis	BvD Orbis (0) Bureau Van Dijk Orbis (0)	BvD Orbis (0) Bureau Van Dijk Orbis (6)	BvD Orbis (12) Bureau Van Dijk Orbis (13)	TI=(BvD Orbis) or AB=(BvD Orbis) or AK=(BvD Orbis) (3) TI=(Bureau Van Dijk Orbis) or AB=(Bureau Van Dijk Orbis) or AK=(Bureau Van Dijk Orbis) (15)	allintitle: BvD Orbis (0) allintitle: Bureau Van Dijk Orbis (0)	"BvD Orbis"(316) Bureau Van Dijk Orbis (519)

Based on the initial searches, we conservatively estimated that at least 185,000 published articles mentioned the reviewed business and finance databases.[5] To further narrow the results, we used citation tracking, keyword search, and skimming techniques. For citation tracking, we closely examined the references section of the relevant articles found from initial searches and used the "cited by" function from Google Scholar to identify related articles published more recently. Also, we incorporated the keywords that describe data quality (including accuracy, reliability, quality, integrity, trust, consistency, discrepancy, differences, error, omission, credibility, and evaluation) and the keywords that describe general data issues (including challenges, problems, issues, biases, weakness, misinformation, and disinformation) to narrow the search results. Since Google Scholar uses automatic stemming and doesn't recognize truncation, the searches were mostly done with the noun forms. We also used adjective forms (including reliable, unreliable, accurate, inaccurate, inconsistent, credible, and trustworthy) to double-check the search results. Finally, we skimmed the article title and the abstract of roughly the first 500 results from each search to evaluate and identify related articles and further read the full-text to verify the findings.

Using these search methods, we repeatedly searched Google Scholar and referred to Business Source Complete and Web of Science from March through June 2020. In total, 98 articles published between 1974 to 2020 were identified and included in this study (see the summary list in Table 2 and the detailed list in Appendix I). The literature search indicated that CRSP and Compustat were the most frequently used databases in academic business research and nearly half of the publications included in this study reviewed these two databases. Most of the articles were published in the last 10 years in accounting and finance journals. Many of these articles were published as working papers and shared via the Social Science Research Network (SSRN).

[5] The calculation of this number considers the highest number of articles retrieved using one of the search terms for each database, except for the CRSP. Since the search term "CSRP" retrieved many irrelevant results, we used the number for the search term "CRSP database" instead.

Despite our effort in providing wide coverage of business literature, we may omit some related articles, due to the large volume of the publication and the fact that the data quality issues were often discussed in footnotes. Several databases including the Thomson Mutual Fund Holdings Dataset, VentureXpert (via Eikon or Thomson One), ReferenceUSA, Value Line were mentioned but didn't become the focus of this study, so the review on these databases was not extensive. Several potentially related databases such as Bloomberg, Global Financial Data, and BvD Osiris were not covered in this study. As a business librarian, the author has professional knowledge for understanding data quality problems, however, the author may not fully present the problems and their implications described in the business literature. The next section will provide the literature review of data quality problems based on the 98 reviewed articles.

TABLE 2 The Summary List of the Number of Articles Included in the Literature Review by Specific Databases, Periods, and Journals

TABLE 2 The Summary List of the Number of Articles Included in the Literature Review by Specific Databases, Periods, and Journals

Database: Number of Articles*	Period Coverage: Number of Articles	Journal Coverage: Number of Articles
CRSP: 21 Compustat: 28 I/B/E/S: 6 Datastream: 10 Worldscope: 7 SDC Platinum: 10 BvD Orbis: 5 Capital IQ: 3 VentureXpert data (via Eikon or Thomson One): 3 Thomson Mutual Fund Holdings Dataset: 2 Mergent Online: 4 ReferenceUSA: 1 Value Line: 5 Database from Foreign Countries: 3 * *There are10 duplicate records because one article reviews more than one database.*	1970-1979: 4 1980-1989: 5 1990-1999: 11 2000-2009: 23 2010-2019: 52 2020 - : 3	SSRN: 16 The Journal of Finance: 13 Other Working Paper: 6 The Accounting Review: 5 Manuscript: 5 The Review of Financial Studies: 4 Journal of Corporate Finance: 3 Journal of Financial and Quantitative Analysis: 3 Journal of Financial Reporting: 2 The Financial Review: 2 The Journal of the American Taxation Association: 2 Finance Research Letters: 2 Dissertation: 2 Accounting Horizons: 2 Others (single publication): 31
	Total: 98	

Literature Review

The literature review section is organized by database providers and then databases. This section covers the Center for Research in Security Prices, LLC product (i.e. CRSP), S&P Global Market Intelligence products (i.e. Compustat and Capital IQ), Thomson Reuters and Refinitiv products (i.e. I/B/E/S, Datastream, Worldscope, SDC Platinum and others), Bureau Van Dijk (BvD) product (i.e. Orbis) and other sources encountered during the research, including Mergent Online, Value Line and ReferenceUSA. The literature review generally follows the timeline of the reviewed articles. Each section offers a short introduction to the company and product history as a background for understanding the time frame of the reviewed articles.

I. The Center for Research in Security Prices LLC[6] Product: CRSP

The Center for Research in Security Prices ("CRSP"), as a part of the University of Chicago Booth School of Business, was established in 1960 with a grant from the Merrill Lynch, Pierce, Fenner & Smith (CRSP, LLC., n.d.a). With its focus to serve the academic and research communities, CRSP data is used widely by academic researchers in accounting, finance, economics, math, and statistics for empirical research related to stock, index, mutual fund, treasury, and REIT market (CRSP, LLC., n.d.b). The database is also used by the commercial market for backtesting and modeling calculations and by government agencies for financial and economic analysis (CRSP, LLC., n.d.c). CRSP provides several data products, including CRSP US Stock Databases, CRSP Historical Indexes, CRSP US Index History Files, CRSP US Treasury Database, CRSP Survivor-Bias-Free US Mutual Funds, CRSP/Ziman Real Estate Database, CRSP Cap-Based Portfolio Index and the CRSP/Compustat Merged Database (CRSP, LLC., n.d.b). The CRSP/Compustat Merged Database provides the historical matching of the CRSP market and corporate action data with Compustat fundamental data (CRSP LLC. 2020a).

Scholars have paid attention to the data quality of CRSP for decades. Rosenberg and Houglet (1974) discussed the data quality problems of CRSP in the article, "Error Rates in CRSP and Compustat Data Bases and their Implications." Their research compared monthly price relatives between 1963-1968 in CRSP and Compustat and found nearly 3% discrepancies for industrial price relatives and nearly 2.4% discrepancies for utility price relatives. They concluded that large errors were relatively infrequent, but data errors could lead to serious consequences: (1) "The few extreme price relatives can influence some properties of the sample to a degree out of all proportion to their small number;" (2) "large errors in the price relatives are to introduce an upward bias in any arithmetic index of monthly return;"(3) "the erroneous price relatives is to pollute statistical analyses of the individual security" (Rosenberg & Houglet, 1974, pp. 1306-1308). Beedles and Simkowitz (1978) followed Rosenberg and Houglet's study and after making appropriate corrections for data errors, they replicated prior research regarding the return behavior of high-risk common stocks and found different results (Beedles & Simkowitz, 1978, pp. 290-291).

Bennin (1980) updated Rosenberg and Houglet's study. In the article, "Error rates in CRSP and Compustat: A second look," Bennin compared the monthly return (including all distributions) data between 1962-1978 in Compustat and CRSP databases and found the overall error rate was only one-third of the rate reported in the Rosenberg and Houglet's study. In general, Bennin described "the cross-checking technique reveals a Compustat error rate of 1/1000, and a CRSP error rate of 1/10000 [on monthly return data] over the years 1962-1978" (Bennin, 1980, p. 1271). Grinblatt et al. (1984) reported the discrepancies of the announcements on proposed splits and stock dividends for the years 1967-1976 between CRSP and the Wall Street Journal Index. Sarig and Warga (1989) compared the CRSP Government Bond Price Dataset with the independently collected Shearson Lehman Brothers (SLB) Bond Data. They found the discrepancies between the two datasets were not random and were largely due to liquidity-driven price errors. They found these

[6] On January 1, 2020, CRSP spun off from Chicago Booth and became its affiliates CRSP, LLC (CRSP LLC, 2020b).

discrepancies were systematically related to certain bond characteristics and proposed some filters to reduce the noise in price records (Sarig & Warga, 1989, p. 367).

Guenther and Rosman (1994) examined the differences between SIC codes assigned to companies by Compustat and CRSP. They found large differences at two-, three-, and four-digit levels. They replicated a prior study and found using Compustat and CRSP codes yielded different results (Guenther & Rosman, 1994). Kahle and Walkling (1996) investigated approximately 10,000 firms jointly covered by Compustat and CRSP from 1974 to 1993 and found substantial differences in the SIC codes designated by the two databases. More than 36% of the classifications disagree at the two-digit level, 50% disagree at the three-digit level and nearly 80% disagree at the four-digit level and "the classification of utilities, financial firms, and conglomerate acquisitions are affected by the choice of CRSP vs. Compustat SIC codes" (Kahle & Walkling, 1996, P. 309).

Courtenay and Keller (1994) examined the distributions (i.e. stock dividends or stock splits) reported by CRSP during the calendar year 1989 against the verified Moody's Dividend Record (MDR). Among 718 observations, they found 142 discrepancies, including "91 coding differences, 20 ex-date differences, eight instances of late updates, five occurrences of arithmetic errors, one case in which an option dividend was improperly treated as a stock dividend instead of a cash dividend, and 17 reporting differences between CRSP and MDR" (Courtenay & Keller, 1994, p. 287). The 91 coding differences included 13 cases where the CRSP coding was incorrect and 64 instances where CRSP used its coding definition that was different from the annual reports (Courtenay & Keller, 1994, p. 287). The researchers further concluded that "the probability of randomly selecting a company reporting a stock distribution improperly administered by CRSP in 1989 is approximately three percent" (Courtenay & Keller, 1994, p. 290).

Loughran and Ritter (1995) alerted that the upward bias in the daily equally weighted index returns in CRSP was substantial. Canina et al. (1998) also warned researchers that compounding daily returns of the CRSP equal-weighted index could lead to surprisingly large biases. "The differences between the monthly returns compounded from the daily tapes and the monthly CRSP equal-weighted indices are almost 0.43% per month or 6% per year. This difference amounts to one-third of the average monthly return and is large enough to reverse the conclusions of a paper using the daily tape to compute the return on the benchmark portfolio" (Canina et al., 1998, p. 403). Yan (2007) offered a new method to "generate an unbiased CRSP daily equal-weighted return with dividend, which is free of the problems associated with the microstructure and consistent with the CRSP monthly index" (p. 1). Yan confirmed that the CRSP daily equal-weighted return is systemically upward biased, and the bias will "lead to a systematically undervalued intercept, which might make a Jensen's alpha more positive (attractive) than it should be. For the beta, the results are mixed and their significant levels depend on individual stocks" (Yan, 2007, pp. 7-8). Yan further elaborated that if a firm has positive excess returns during the estimation period, by using an upward biased daily index, a positive event will be exaggerated (Yan, 2007, p. 8).

Shumway (1997) and Shumway and Warther (1999) cautioned researchers against the delisting bias in CRSP. Shumway (1997) found CRSP files were missing thousands of delisting returns. "Omitted delisting returns introduce a bias into studies that use the CRSP data" (Shumway, 1997, p. 328). Without delisting returns, "it is not possible to accurately calculate the returns to a feasible portfolio" and overlooking the delisting bias may result in "other unknown data biases confounding empirical results" (Shumway, 1997, pp. 328, 340). Shumway and Warther (1999) further investigated the delisting bias in CRSP's Nasdaq data. They found many delisting returns were not collected in CRSP and some categories were missed more often than others: "delisting returns associated with poor firm performance (e.g., bankruptcy or failure to meet capital requirements) are missed much more often than returns associated with neutral or good firm performance (e.g., merger, acquisition, or migration to another exchange)" (Shumway & Warther, 1999, p. 2361). Tobek and Hronec (2018) found the quality of the delisting data improved since Shumway's study and identified 2,742 out of 20,680 (13%) delistings in CRSP were missing as of 2017 (Tobek & Hronec, 2018, p. 12).

Elton et al. (2001) examined the accuracy of the CRSP Survivor-Bias-Free US Mutual Fund Database and identified the omission bias in the database. They pointed out that "although all mutual funds are listed in CRSP, return data is missing for many and the characteristics of these funds differ from the populations" (Elton et al., 2001, p. 2415). Thus, even though the CRSP database "does not have traditional survivorship bias, it does have a form of survivorship bias called omission bias that causes the same type of problems as does traditional survivorship bias" (Elton et al., 2001, p. 2416). They also identified the upward bias in CRSP's monthly returns: "the returns in the CRSP database are upward biased in any month where there are multiple distributions on the same day" (Elton et al., 2001, p. 2416). They analyzed the accuracy of CRSP's merger data and found "the CRSP data on merger dates are inaccurate enough to require that all merger dates be independently validated" (Elton et al., 2001, p. 2425). Finally, they compared CRSP with the Morningstar database and found serious differences in alpha and returns, particularly for older data and small funds (Elton et al., 2001, p. 2429).

Wisen (2002) found the CRSP Survivor-Bias-Free US Mutual Fund Database is not bias-free and it has another type of selection bias called incubation bias. "Incubation causes a selection bias when new funds with poor performance are not added to databases as promptly as new funds with superior performance are added to databases" (Wisen, 2002, p. 3). Wisen explained, "incubation bias differs from the more widely studied problem of survivorship bias because incubation bias is due to a systematic exclusion of some new funds from databases, whereas survivorship bias is caused by the removal of terminated funds from databases" (Wisen, 2002, p. 3). Wisen also argued that the practice of CRSP in excluding the returns of new funds with less than $15 million in assets created a subtle form of survivorship bias (Wisen, 2002, p. 7). Their research found "approximately one-third of the terminated funds in [their] study were missing their initial returns in CRSP" and on average the first 15 months of returns were not recorded for these funds (Wisen, 2002, p. 8).

In terms of incubation bias in CRSP, Evans (2007) documented that "for a sample of domestic equity funds, 39.4% of funds are incubated and the incubation bias is estimated to be 4.7% in raw returns and between

1.9% and 3.3%, risk-adjusted" (p. 1). Evans (2010) further documented both public incubation bias and private incubation bias in CRSP. CRSP (2020c) admits the existence of duplication bias and selection bias in their data in the Mutual Fund Data User Guide. It mentions, "the returns histories are sometimes duplicated in the database. For example, if a fund started in 1962 and split into four share classes in 1993, each new share class of the fund is permitted to inherit the entire return/performance history. This can create a bias when averaging returns across mutual funds"; and "a selection bias favoring the historical data files of the best past performing private funds that became public does exist" (CRSP LLC., 2020c, p. 4). Jorion and Schwarz (2017) explained the backfill bias (or 'instant-history' bias) associated with incubation: "the backfill bias arises when the fund's performance is not made public during some incubation period but then is added to the database presumably following the good performance (p. 1). They believed that the listing decision generated a bias because the fund manager's decision to include the fund or not was most likely correlated with past performance (Jorion & Schwarz, 2017).

Schwarz and Potter (2016) discovered that CRSP and the Thomson Mutual Fund Database contained many voluntarily reported mutual fund portfolios, however, the two databases were missing many mandated portfolios that were available in the SEC filings (p. 3520). Their research also found that CRSP portfolios' positions before the fourth quarter of 2007 were inaccurate when the data were acquired from Morningstar, and during this period, "one in five CRSP fund portfolios has 25% or more of their positions reported inaccurately" (Schwarz & Potter, 2016, p. 3520). Schwarz and Potter didn't suggest researchers using the CRSP portfolio data before the fourth quarter of 2007 (Schwarz & Potter, 2016, pp. 3521-3522).

Francis et al. (2016) pointed out that "despite the precision of CRSP data, researchers may inadvertently generate imprecise measurements when modifying and adjusting CRSP variables" (p. 13). They reminded researchers that "stand-alone share prices adjusted with CRSP adjustment factors are inaccurate in the presence of property dividend, spin-off and rights offering events" and "ignoring covertly missing stock returns may create misleading test results" (Francis et al., 2016, p. 2).

2. S&P Global Market Intelligence Products

1) Compustat

Compustat, developed by Standards & Poor's (S&P) around 1964, was one of the earliest services that collected data on public companies (New Research Center, 1965). After S&P acquired Capital IQ in 2004, some portion of Compustat's fundamental data was available via the S&P Capital IQ platform (Zuckerman, 2004). Compustat is introduced by S&P as a comprehensive dataset with standardized, historical, and point-in-time data (S&P Global Market Intelligence, 2017). Besides time-series fundamental data, the company believes what differentiates Compustat from its competitors is their extensive research of management discussion, footnotes, and analysis of detailed supplemental data items such as historical industry classifications, segment data (including operating and geographic segment, customer and product data), debt, options, pension, and industry-specific data (S&P Global Market Intelligence, 2017). Compustat is the most

frequently used database for business and financial empirical research. Consequently, its data quality problems are widely noticed and discussed by academic researchers.

San Miguel (1977) compared the Research and Development (R&D) expense data in Compustat with the original data in the 10-K reports. For the sample data of 256 firms in 1972 retrieved from Compustat, they found 78 (30%) differences, most of which resulted from the incorrect inclusion of contract research into R&D expenses (San Miguel, 1977. P. 640). They notified Compustat about their findings and the company reviewed the data and found approximately 125 companies had the same data problem (San Miguel, 1977, p. 639).

Ball and Watts (1979) questioned the process that Compustat used to construct their data files and provided additional evidence of survival biases in Compustat. They argued that the data files in Compustat were constructed retrospectively to meet security analysts' interest and researchers "ended up analyzing data on an unrepresentative sample of firms, with a lower-than-average expected frequency of earnings decline" (Ball & Watts, 1979, p. 197). McElreath and Wiggins (1984) captured four types of data problems regarding Compustat data files, including (a) incorrect data; (b) inconsistent use of definitions, which created comparability problems; (c) survivorship bias; and (d) potential timing problems relating to whether the data were available to the public at the time a study assumed they are (McElreath & Wiggins, 1984, p. 71). The research proposed some solutions to tackle data problems, including using statistical methods and dataset comparison to detect data errors (McElreath & Wiggins, 1984, p. 73).

Banz and Breen (1986) examined the effect of the ex-post-selection bias and the look-ahead bias in the Compustat datasets. They noticed that biases in business databases had been long aware of by empirical researchers; however, since there had been no practical way of measuring the size of the biases introduced, some studies had ignored the problems, others had used various measures designed to reduce the biases, while some had claimed that the biases are of a negligible magnitude (Banz & Breen, 1986, p. 780). After comparing the results from the standard Compustat data with those from a bias-free dataset they collected over the years, they found that the portfolio rates of return from two datasets differed significantly and could result in different conclusions in hypothesis testing (Banz & Breen, 1986, p. 779). Kinney and Swanson (1993) evaluated 19 tax field data in Compustat and discovered that the "error rate varies widely," and "it is generally higher for items reported in the footnotes than for items reported on the income statement and balance sheet" (p. 121).

Several studies found significant discrepancies between Compustat and other databases. Kern and Morris (1994) compared Compustat with the expanded Value Line databases and found significant differences in commonly used financial data items such as sales and total assets. They also found the differences in the two databases can materially affect inferences about the population of firms and the outcomes of empirical research (Kern & Morris, 1994, pp. 274, 284). They noticed that the differences were primarily attributed to the differences in "data assimilation policies concerning mergers and acquisitions, accounting changes, and discontinued operation" (Kern & Morris, 1994, p. 275). Yang et al. (2003) examined the accuracy of seven

frequently used accounting variables in Compustat and Value Line databases and found "substantial data differences," resulting from definitional discrepancies and direct measurement error (p. 1). Ulbricht and Weiner (2005) examined more than 650 data items from 1985 to 2003 for the US and partly Canadian firms in Worldscope and Compustat. They found the two databases could lead to comparable results, but "if e.g. a size bias is not treated with care, the quality of results may differ [considerably]" (Ulbricht & Weiner, 2005, p. 1). Tallapally (2009) showed that different bond rating models responded differently to the choices of Compustat versus Mergent data.

Several studies disclosed that researchers improperly used Compustat data as proxies. Mills et al. (2003) warned researchers to take extra care when using Compustat net operating loss data as an indicator of a firm's US tax-loss positions, particularly when the research involves firms with foreign operations or acquisitions activity. Ali et al. (2008) criticized the industry concentration measures calculated with Compustat data. They pointed out that because Compustat only covers public firms in an industry, "they are poor proxies for actual industry concentration" and "these measures have correlations of only 13% with the corresponding US Census measures, which are based on all public and private firms in an industry" (Ali et al., 2008, p. 3839). These measures could lead to incorrect research conclusions (Ali et al., 2008, pp. 3843). Keil (2017) reiterated this problem and mentioned that popular approximations of the Herfindahl Index based on Compustat dataset "have a vanishingly low correlation with the more comprehensive Census measure" and consequently, major financial variables of interest show different correlations with these concentration indicators, which can "lead to a breakdown of regression results" (Keil, 2017, p. 467). Banyi et al. (2008) questioned earlier studies that used the data from Compustat and CRSP as proxies to estimate the number of shares repurchased.

Shi and Zhang (2011) found the differences between the two measures of accruals (one uses a balance sheet; the other uses a cash flow statement) calculated with Compustat data. They elaborated that the "non-articulation in working capital accounts and depreciation expenses between the cash flow statement and other financial statements is surprisingly prevalent and economically significant, and it can be attributed to special events, errors made by Compustat, firms' inconsistent definitions, and nonstandard classifications of assets and liabilities" (Shi & Zhang, 2011, p. 811).

Since the mandatory requirement for all public US GAAP companies to file their financial reports using the XBRL (eXtensible Business Reporting Language) effective on June 15, 2011,[7] more studies have been conducted to compare Compustat's data with 10-K filings and sometimes at a large scale. Boritz and No (2013) retrieved the XBRL-tagged interactive data from SEC's EDGAR for a sample of 150 XBRL filings of 75 firms for the period 2009-2011 and compared the data with the corresponding data in Compustat. They found that 6279 (44.3%) financial facts in Compustat matched with the interactive data, 677 (4.8%) financial facts did not match and Compustat had 7,207 (50.9%) omissions (Boritz & No, 2013, p. 38). In a

[7] On June 28, 2018, the SEC adopted the amendments that require the use, on a phased in basis, of Inline XBRL for operating company financial statement information and fund risk/return summary information. See more at https://www.sec.gov/structureddata/osd-inline-xbrl.html

relatively small-scale study, Tallapally et al. (2011) compared EDGAR (normalized data) with Compustat (standardized data) of the "cost of goods sold" item for 26 manufacturing companies included in DOW 30 companies for the fiscal year 2009. Out of the 26 companies compared, only one company's "cost of goods sold" data matched between Compustat and EDGAR. Comparatively, another study by the researcher compared differences in sales (or revenues) of 27 non-financial companies included in the DOW 30 between Compustat and EDGAR. Seven discrepancies among the 27 companies were observed (Tallapally et al., 2012). Chychyla and Kogan (2014) leveraged XBRL to automatically extract thousands of 10-K numbers to create a data sample and investigate the effects of Compustat's standardized data (via Capital IQ) versus original 10-K data on bankruptcy prediction models. They concluded that "Compustat's data standardization not only yields no improvements for bankruptcy prediction models but also has a significant (up to 8.56%) negative impact on the predictive accuracy of Altman's model" (Chychyla & Kogan, 2014, p. 1).

Chychyla and Kogan (2015) extended the research from Boritz and No and conducted the first large-scale study over the Compustat North America Fundamentals Annual Filings data and the 10-K data by comparing 30 accounting variables for approximately 5,000 companies from 2011 to 2012. The research showed that the values of 17 out of 30 analyzed variables in Compustat significantly differ from the values reported in the 10-K filings. The researchers found that the discrepancies were more likely to occur to complex financial concepts such as "cost of goods sold" or "gross profit" as opposed to simple concepts such as "total assets," "total liabilities," or "net income" (Chychyla & Kogan, 2015, p. 70). They summarized four reasons for the differences, including standardization, erroneous data due to typos or rounding, not-up-to-date data due to restatements, and missing data (Chychyla & Kogan, 2015, p. 43). Bratten et al. (2016) also contributed to this discussion and estimated that "Compustat data error entry rate of 13 percent" overall for footnote entries and believed that "this is likely due to the difficulty of collecting detailed data from non-standard financial statement footnote disclosures" (Bratten et al., 2016, p. 40).

Williams (2015) explored the usefulness of the XBRL company fillings in his doctoral dissertation. Williams's research found that the original data from XBRL filings cannot be used to create earning prediction models, due to a large number of missing values; but using the functionality directly built into XBRL taxonomy, the fully populated XBRL company filings can be used to create earning predictions. Williams tested two earning prediction models using populated XBRL data and Compustat. The test found in one prediction model, "fully populated XBRL company filings predicted future earnings with a higher level of accuracy than Compustat did" and for the other model, there was no significant difference between the two datasets (Williams, 2015, pp. 89-90).

McGuire et al. (2016) described the concept of "database effect" as they compared the data from Compustat Global, Osiris, and Worldscope. Using multiple statistical estimation techniques and replication studies, they confirmed that "researchers would likely come to a different conclusion based on the database used" (McGuire et al., 2016, p. 186). Although each database claimed to be relatively comprehensive in terms of global firm coverage, the actual coverage was far from identical. The researchers believed that "country-specific differences in firm and country coverage may lead to biased conclusions, and inconsistent yearly

samples may pose challenges for researchers using panel data" (McGuire et al., 2016, p. 187). The database effect was also examined in an earlier study by Lara et al. (2006), who compared seven widely used databases for international accounting research and concluded that the database choice can affect the results and findings of international accounting research (p. 449).

Casey et al. (2016) disclosed the prevalence of missing values and data errors in Compustat. They proposed a Modified Financial Statement Balancing Model to solve the problems of missing values or erroneous entries and restore them into usable data points such as zeros or summary amounts. Using Compustat data of US nonfinancial firms for 1988-2011, their model identified 560,684 (30%) exceptions out of 1,847,444 firm-year equation observations. They followed a three-step process to resolve the exceptions: 1) replacing null variable values with zero when applicable, 2) replacing zero or missing values in total assets, total liabilities, total current assets, and total current liabilities with the sum of their respective components, and 3) making changes based on generally accepted accounting principles (Casey et al., 2016, p. 38). Hribar (2016) generally believed that this approach is "sensible" and the formalized procedure is a "logical approach to deal with the fact that the missing value could be zero or non-zero" (p. 63). Casey et al. (2019) further discussed this issue in the article "Measuring Reporting Quality."

Bostwick et al. (2016) questioned the standardization and the adjustment process of Compustat in treating depreciation, depletion, and amortization allocated to the "cost of goods sold" variable. Using a sample of 10,758 firm-years across all industries from 2008–2011, they found that Compustat "cost of goods sold" understates the 10-K "cost of goods sold" by an average of 7.5%. Since the Compustat "gross margin" is computed using the "cost of goods sold" variable, it results in the overstatement of the Compustat "gross margin" by 14.3% (Bostwick et al., 2016, p. 191). They communicated such issues with Compustat and offered some treatments for researchers to reconcile the differences (Bostwick et al., 2016).

Utke (2018) disclosed the miscoding of Compustat's auditor variables and explained that the miscoding in the auditor variable can "affect studies of Big N effects, industry specialization, auditor tenure, and auditor changes, among others" (p. 57). Utke found that many miscodings resulted from an auditor change - "following an auditor change, the previous auditor's report remains in a firm's 10-K, and Compustat occasionally codes the previous auditor as the current auditor" (Utke, 2018, p. 56). The study identified 230 (0.35%) miscodings out of the 66,365 observations (Utke, 2018, p. 57).

Heitzman and Lester (2020) quantified the coverage and measurement errors of the "net operating loss" variable in Compustat. They found that Compustat failed to identify the existence of "net operating loss" in 25% of large firms and for firms that Compustat correctly identified a tax loss, it significantly understated the balance of the loss. They also pointed out that "Compustat does not distinguish the cash tax value of losses generated across differing jurisdictions, treating one dollar of state tax loss the same as that of federal or foreign losses" (Heitzman & Lester, 2020, p. 2).

2) S&P Capital IQ

The Capital IQ platform was founded in 1998 by two investment bankers and was commercially launched in 2000. The platform provided information on public and private companies, private capital firms, transactions, and executives. It was acquired by the Standard & Poor's in 2004. The acquisition extended its content to covering "fixed income, equities, indices, and mutual funds as well as select portions of fundamental data from the Compustat unit" (Zuckerman, 2004). Since 2007, S&P Capital IQ updated its platform with debt and credit data, which provided the information about capital structure of public companies including senior and subordinated debt, secured debt, commercial paper, and bank facilities as well as fixed payment schedules and credit ratios (Heires, 2007). The product was originally designed to provide services for the investment banking community, but it gradually entered the academic market (Phillips, 2012).

The Capital IQ's debt structure and credit lines data are widely used by researchers (Kahle & Stulz, 2013; Acharya et al., 2014; Mathers & Giacomini, 2016; Chang & Shim, 2017; Choi et al., 2018). Acharya et al. (2014) utilized the information on the drawn and undrawn lines of credit from Capital IQ to conduct empirical tests and justified the "credit lines as monitored liquidity insurance" theory. Choi et al. (2018) used corporate debt structure data such as bonds, notes, and maturities for loans (revolving credit and term loans) from Capital IQ to analyze corporate debt maturity profiles. Kahle and Stulz (2013) relied on the information of bank loans and revolvers to create a sample of small, bank-dependent firms. The competitor data from Capital IQ is also mentioned frequently in business literature. Rauh and Sufi (2012) drew the competitor data from Capital IQ to study the corporate capital structure. Benedettini et al. (2013) used Capital IQ to compile a broad set of potential competitors. Röhm et al. (2019) referred to the business descriptions and corporate tree function in the database to identify investor's parent companies. The credit rating, executive remuneration data, stock listing, and management forecast data in Capital IQ are also used by researchers (Dhaliwal et al., 2014; Silva, 2017).

Despite the wide use of Capital IQ, only a small number of articles investigated its data quality. In a research note on Capital IQ's credit line data, Mathers and Giacomini (2016) compared Capital IQ's data with carefully hand-collected data and found Capital IQ often reported missing values when there was data available on credit line in the company's 10-K filings. As described in the article, three prior research found 85%, 79%, and 85% of sampled firms have a credit line respectively; however, only 18.9% of the sample drawn from Capital IQ had credit line data (Mathers & Giacomini, 2016, p. 440). As they explained, the disparity between Capital IQ credit line data and hand-collected data were significant. Due to the missing values, only 27.3 % of Capital IQ's reported "drawn amount outstanding" and 8.9% of Capital IQ's reported "undrawn revolving credit" exactly matched the hand-collected observations (Mathers & Giacomini, 2016, p. 441). If not counting the missing values, the percentage of matches for drawn and undrawn credit was 43.2% and 38.7% respectively (Mathers & Giacomini, 2016, P. 441).

Lee (2017) compared the outstanding debt data in Compustat with the same data in Capital IQ. The research found Capital IQ had a significant number of missing values: "about 51% of these samples (32,356 firm-year observations) have missing values in total term loans, 95% of commercial paper outstanding is missing, 81.4%

of subordinated debt is missing, and 77.2% of convertible debt information is missing, which makes it harder to analyze a firm's debt structure using the Capital IQ database" (p. 40).

Benedettini et al. (2013) studied a set of potential competitors and mentioned that Capital IQ provided more relevant competitor information than several other databases including Mergent Online, Hoovers, and Factiva. They attributed it to the fact that Capital IQ acquired the competitor's information from the SEC filings, press releases, and direct company contacts, while the other databases identified the competitors by industry categories and locations. They also mentioned that even though Thomson ONE Banker and Bloomberg databases also provided relevant competitor information, only a small number of the firms in their bankruptcy sample (12 and 14 respectively) were covered by these two databases. Comparatively, 54 of the 75 target firms were found in Capital IQ (Benedettini et al., 2013).

Mathers and Giacomini (2016) also cautioned researchers to notice the fiscal year coding difference between Capital IQ and Compustat. Since Capital IQ is often used together with Compustat, it is crucial to aware that "Capital IQ's fiscal year reporting doesn't match Compustat when comparing the firms with a fiscal year-end month prior to June. For example, for a firm with its fiscal year-end date of March 2003, Capital IQ reports the data as the fiscal year 2003 while Compustat reports it as the fiscal year 2002" (Mathers & Giacomini, 2016, p. 437). In the sample collected by the researchers, 9% of firms were found mismatched due to these differences in the coding of the fiscal year (Mathers & Giacomini, 2016, p. 437).

3. Thomson Reuters-Refinitiv Products

A series of data products from Thomson Reuters and Refinitiv were widely used by academic researchers. In the last 15 years, the company has gone through phases of acquisition, reorganization, and alliances.[8] Despite the company's structural changes over time, their legacy products are enduring.

1) I/B/E/S

The Institutional Brokers Estimate System (I/B/E/S) was firstly founded by a New York brokerage firm and began collecting earnings estimates for US companies around 1976. Through several transactions, the company was sold to Primark in 1995 (Bloomberg Business News, 1995). In 2000, Thomson Financial acquired Primark (Collings, 2000) and I/B/E/S became one of its major modules in Thomson One Investment Management products (Thomson Reuters, 2006). This system compiles the forecasted earnings and analysis of publicly traded companies and is recognized as one of the important security and portfolio analytical tools.

[8] Thomson Corporation acquired Reuters during 2007-2008 and formed Thomson Reuters. Thomson Financial Services Inc, was combined with Reuters to create the Markets Division, which later became Financial & Risk Division under company's restructuring during 2011-2012 (Thomson Reuters, 2008, 2011, 2012). In 2017- 2018, Thomson Reuters sold 55% of its Financial & Risk business to private equity funds managed by Blackstone and retained a 45% interest in the new company, which is now known as Refinitiv (Thomson Reuters, 2018, 2019). The London Stock Exchange committed its takeover of Refinitiv and expects to complete the deal by early 2021 (Jones, 2020; CNBC, 2020).

The I/B/E/S historical earnings forecast data and services are used widely by academic researchers to validate their investment theories and observations. Concerns over the I/B/E/S data have been documented by many researchers. Payne and Thomas (2003) found that "[I/B/E/S] adjusting for stock splits and rounding to the nearest penny can cause a loss of information" and using the actual (unadjusted) earnings and forecast data from I/B/E/S can overturn prior research results based on split-adjusted data (p. 1049). They also pointed out that because "researchers are prohibited in many cases from determining the amounts actually reported in prior years, leading to misclassified observations" (Payne & Thomas, 2003, p. 1049). Roger (2017) investigated analysts' earnings forecasts of UK companies and revealed that over 10% of the analyst codes in the database were subject to reporting errors. These reporting errors affected the evaluation of analysts' characteristics and could bias empirical studies that rely on tracking analysts (Roger, 2017).

In the article "Rewriting History," Ljungqvist et al. (2008) documented widespread ex-post changes to the historical contents of the I/B/E/S analyst stock recommendations. Across a sequence of seven downloads between 2000 and 2007, they found that between 6,594 (1.6%) and 97,579 (21.7%) of matched observations were different from one download to the next. They found the changes including the alterations of recommendation levels, the additions and deletions of records, and the removal of analyst names were non-random. The findings attracted public attention as the Financial Times disclosed the issue and raised the question about the integrity of the database (Brown-Humes, 2006; FT Alphaville, 2007). The database provider partly addressed the concerns and responded: "the names of the individual analysts remain in the database. However, they were not visible on the files seen by the academics due to an incomplete data feed;"… and "[the company is] working to rectify the problem with feeds" (Brown-Humes, 2006).

Acker and Duck (2009) added to Ljungqvist's research and reported the concerns over the I/B/E/S and Worldscope data on final earnings announcement dates of UK companies retrieved from the Thomson ONE Banker Package. They identified three major problems: (1) "year-end earnings announcement dates were frequently misreported in the I/B/E/S database." Compared with 1,874 of hand-collected data, 24% of the I/B/E/S data were incorrect, 97% of which was later than the true date (Acker & Duck, 2009, P. 4); (2) There are about 22% discrepancies between the announcement dates reported in I/B/E/S and those reported in Worldscope; and (3) "When the I/B/E/S announcement date is later than the true report date, it is possible for the forecasts also to be dated after the true report date. Analysts can, therefore, appear to be forecasting earnings per share after the actual figure has been made public" (Acker & Duck, 2009, p. 5). As they believed, such inaccuracies can in many ways distort the results of related studies. They communicated such problems with the I/B/E/S and made the vendor to review approximately 2 million records in the I/B/E/S database and identified 50,000 errors in European announcement dates. The data provider also initiated a project to review the data for US firms (Acker & Duck, 2009, p. 6).

Brown and Larocque (2011, 2013) examined the prevalence and the consequences of data discrepancy between the I/B/E/S actual earnings-per-share data (EPS) and the analysts' inferred EPS. They found that the I/B/E/S actual EPS differs from the analyst's inferred actual EPS 39% of the time. Thus, the data discrepancy is prevalent in the I/B/E/S earnings database. They also found the data discrepancy was systematic and associated with analyst, firm, industry, and year. They stressed four adverse consequences of the data discrepancy: "(1) less accurate earnings forecasts by analysts; (2) smaller forecast revision coefficients by

analysts; (3) more disperse earnings forecasts among analysts following a firm; and (4) lower market reactions to firms' I/B/E/S-based earnings surprises" (Brown & Larocque, 2011, p. 26). They also found that the discrepancy would affect the result of prior research (Brown & Larocque, 2011).

Call et al. (2020) compared annual earnings forecasts across two versions of the I/B/E/S detail files, one made available in 2009 and the other made available in 2015. They found "substantial differences in the contents of these two versions of the detail files, as well as significant differences in the attributes of the earnings forecasts available in each version" (Call et al., 2020, "abstract"). The researchers concluded that "differences related to earnings forecasts continue to occur, with potential long-term implications for researchers" (Call et al., 2020, p. 5). In private conversations with the researchers, Thomson Reuters provided some explanation for the changes, which included retroactive adjustments for stock splits or stock dividends, "default currency" adjustments, and correction of errors. The vendor also disclosed that some differences occurred because the brokerage maintained control over the distribution of these forecasts and academic subscribers often had access to only a subset of all the earnings forecasts. Despite the explanations, the lack of transparency concerned researchers, and the inconsistency of the data sources greatly affected the researcher's practices in empirical research (Call et al., 2020, pp. 6-8).

2) Datastream

Datastream was developed by the Hoare & Co in 1967. Through a series of acquisitions, the company was sold to Dun & Bradstreet in 1984. In 1992, Datastream was acquired by Primark, which was later sold to Thomson Financial in 2000 (Derasse, 2017). Datastream, as a global financial and macroeconomic data platform, provides data on equities, stock market indices, currencies, company fundamentals, fixed income securities, economic profiles, and key economic indicators for the majority of the countries in the world. I/B/E/S is also available through Datastream (Refinitiv, 2019). The database is used widely for top-down macro analysis, financial analysis, sector research, and asset allocation strategy research (Derasse, 2017). The database is recognized for its broad market coverage and long historic market data and is used widely to conduct time-series and cross-country studies (Brückner, 2013).

Bloom et al. (2004) released a technical note on using company accounts data from Datastream. The article pointed out that due to mistakes and inconsistencies, the cleaning procedure is necessary when using the database. The problems ranged from "simple typographical errors to more complicated issues such as breaks in company time series due to mergers," and "if the data are not cleaned, then outliers can have a strong influence on any subsequent regression results" (Bloom et al., 2004, p.7).

Ince and Porter (2006) reviewed the individual equity return data from Datastream and warned researchers to "handle with care!" As they described, the most troubling finding was "the inability to distinguish easily between the various types of securities traded on equity exchanges" (Ince & Porter, 2006, p. 464). For instance, many securities classified as common stocks are not such. Besides, "the full-time series of exchange classification variables often reflected only the most current value," so delisted securities would not have a track record of prior exchanges (Ince & Porter, 2006, p. 464). The classification errors induced a survivorship bias which implied that delisted firms were less likely to be included in a Datastream sample (Ince & Porter,

2006, p. 474). They also identified several issues related to calculating total returns using return variables. They questioned some of the technical practices from the database: "rounding prices to the nearest penny can cause nontrivial differences in the calculated returns when prices are small" and "the return index is reported to the nearest tenth; therefore, when the level is very small, the rounding of large absolute price level changes can have a significant effect" (Ince & Porter, 2006, p. 473). Also, they found many instances of data errors including stock splits reflected on incorrect dates. When comparing the database with CRSP, the researchers found differences in coverage and discrepancies in closing prices and dividend observations (Ince & Porter, 2006, p. 472).

A series of studies between 2007-2011 revealed several problems of the Datastream data for the UK market. Researchers continued to discover survivorship bias in Datastream especially for foreign stocks listed in the UK and small stocks on the deadstock file of the database (Andrikopoulos et al., 2007). Andrikopoulos et al. (2007) mentioned that the biggest problem of using Datastream is "the vast inclusion of non-ordinary items in the equities section of the database; the appearance of more than one record for certain ordinary stocks and the inability to provide an accurate static mirror image of the UK equities market at any given point in time as it doesn't provide accurate listing and de-listing information" (p. 17). Espenlaub et al. (2009) reported "a fundamental error in Datastream equity data for share prices and return indices relating to a failure to make any capital adjustments for UK open offers before February 2002" (p. 61). Rossi (2011) seriously criticized the deficiencies of Datastream - "the data is so bad and flawed that statistical inferences driven without a thorough review and correction exercise are, at best, totally unreliable" (p. 3). According to Rossi (2011), the situation of incorrect share information was very common and the database often failed to correctly account for the effect of stock splits, which affected market cap and turnover calculations (p. 14). Rossi described two troubling lessons: one was that "within the Datastream sample, it is not possible to determine in a recursive, rule-based manner, which constituents must be retained and which should not," and due to incorrect static classification, the researcher excluded nearly 25% of constituents from the sample; the other lesson was that more than 50% of the already filtered sample had to be excluded due to incorrect historical data on prices, shares, returns, and volume (pp. 17-18). Rossi found that the problem not only happened to small stocks but also to large and mega-caps, especially for shares and volume data (p. 18). So, Rossi reminded researchers to be aware of data problems, both "when planning a cross-sectional analysis" and "when reading results drawn from samples that have not been treated accordingly" (p. 18).

Brückner (2013) examined the equity data from Datastream to evaluate whether the database can be used as the primary data source for a German stock market study. Brückner found the coverage of the database was insufficient for equity research before 1990 and the errors in total return indices were "mainly caused by price differences and incorrect adjustments for dividends and corporate actions" (p. 3). The classification problem identified by Ince and Porter (2006) still existed. As Brückner explained, "one of the important weaknesses is that Datastream does not provide time-series information about the market segment in which a stock is listed. As a consequence, the standard procedure of using portfolio breakpoints from the top market segment cannot be followed," which has great implications for size effect related studies (p. 1). Brückner also identified incorrect classification due to the improper translation of the German term "Vorzugsaktien" ("non-voting stock") into "preferred stocks." As Brückner described, Datastream typically classified the German non-voting

stocks as preferred shares, causing the German non-voting shares incorrectly removed from data samples (p. 3). Brückner didn't recommend Datastream as the primary data source for German stock market studies before 1990 and warned researchers that equity data after 1990 should be handled with care.

Landis and Skouras (2018) reiterated the problem caused by the improper classifications of the database - only including the current exchanges for socks and excluding secondary exchanges. They explained that this problem would "induce a sample selection bias because it will lead to the exclusion of stocks that are in secondary exchanges because they have been demoted from a primary exchange due to poor performance" (Landis & Skouras, 2018, p. 8).

Tobek and Hronec (2018) compared Compustat with Datastream and found systematic differences in the raw financial statements. They pointed out that "different stock coverage across the databases can lead to large statistically and economically significant disparities in the returns" (Tobek & Hronec, 2018, p. 1).

Nobes and Stadler (2018) mentioned another relatively complex problem of the "number of shares free float" data (data field NOSHFF) in Datastream. The data field represented the percentage of total shares available to ordinary investors. They noticed that for many large Chinese state-owned enterprises who had shares traded in both Mainland China (A Shares) and Hong Kong (H Shares), the NOSHFF was misleading. "Because Datastream collects NOSHFF for each of the two types of shares individually, even though the share capital of the firm comprises A and H Shares. For example, PetroChina has a NOSHFF of 90% at 31 December 2013 for H Shares (Datastream Code 280366), suggesting that it is not government-controlled. However, taking A and H Shares together, 86.51% of the total share capital is in government hands. This is relevant because PetroChina uses IFRS for its Hong Kong listing and, when the firm is included in an IFRS study, information related to H Shares is probably used" (Nobes & Stadler, 2018, P. 608).

3) Worldscope

The Worldscope database was created by the Wright Investors' Service, a US-based global money management firm. The company identified the need for such resources in their international investment management activities. In 1990, the Wright Investors' Service and the Disclosure Inc. (a division of Primark Corporation) formed as a joint venture Worldscope/Disclosure Partners. Primark acquired the remaining interest in Worldscope in 1999. In 2000, Thomson Corporation acquired Primark Corporation and Worldscope became one of its featured products ever since (Thomson Reuters, 2013). Worldscope is featured with detailed standardized financials, analysis, and stock performance information on the world's leading public corporations as well as many key private companies (Thomson Reuters, 2008, 2016). The database is used frequently in cross-country equity research with many instances used for equity research in emerging markets (Lins & Servaes, 2002; Alfaro et al., 2019; Esqueda & O'Connor, 2020). The database is sometimes used together with I/B/E/S and BvD Orbis (Acker & Duck, 2009; Daske et al., 2013; Gupta, & Krishnamurti, 2016; Alfaro et al., 2019). It is also frequently used together with Datastream (Daske et al., 2013; Weiß & Mühlnickel, 2014; Lu et al., 2017; Landis & Skouras, 2018; Zhang et al., 2019; Jacobs & Müller, 2020).

Firat (2002) noticed the puzzling differences in reported financial data in Worldscope and Datastream. Worldscope explained what contributed to the differences was the different reporting standards (namely data item definitions and representations) used by the different databases to meet different user preferences (Firat, 2002). As Firat explained, based on those preferences, database providers usually provide financial data in one or more of the following ways: "As Presented (data provided by the SEC and similar foreign agencies); As Reported (data modified to fit a standard attribute naming convention); In Local Format (data fits local accounting practices); and Standardized (data modified based on the knowledge of the industry and extensive research in order to allow meaningful performance analysis)" (p. 2). Firat believed that these preferences and local adaptations of data contributed to data-level, ontological and temporal heterogeneity in financial data sources (pp. 2-3).

Ulbricht and Weiner (2005) systematically compared Worldscope and Compustat to investigate whether the choice of the data sources has an impact on the outcome of empirical research. They found that the two databases lead to comparable results, but if the size bias is not treated with care, the outcome may differ considerably (Ulbricht & Weiner, 2005, p. 1). Their research found that the coverage of Worldscope data is 25% broader than Compustat since 1998, but the overall quality of Compustat is higher than Worldscope (Ulbricht & Weiner, 2005, p. 12). The statistics for key accounting items such as net sales showed the mean and median values differ significantly between the two databases (Ulbricht & Weiner, 2005, p. 12). Appling two datasets to the multiple valuation procedure, which compared the estimated enterprise value and the observed enterprise value, they concluded that Worldscope had significantly lower valuation error than Compustat (Ulbricht & Weiner, 2005, P. 27).

Acker and Duck (2009) compared hand-collected earnings announcement dates for UK companies with their counterparts in Worldscope and found 8% of the Worldscope dates were incorrect (p. 5). As they reviewed the reliability of I/B/E/S earnings announcement dates, they found 22% discrepancies between Worldscope and I/B/E/S and expressed their concerns over the internal consistency of Thomson Reuters (Acker & Duck, 2009).

Daske et al. (2013) utilized Worldscope and Compustat Global Vantage to study the liquidity and cost of capital effects around the voluntary and mandatory International Accounting Standard (IAS) and International Financial Reporting Standards (IFRS) adoption. As a byproduct of the research, they used massive hand-collected data to assess the suitability of commercial databases for their research question and gauge the effect of potential misclassifications. They found "the two data sources provide contradictory information on about every third IAS firm-year observations" (Daske et al., 2013, p. 536). They also noticed that both databases exhibited "substantial classification differences" compared to their hand-collected data from the annual reports – "hand-coding disagrees in about 25% of the cases with Worldscope or Global Vantage" (Daske et al., 2013, p. 536). They also revealed that the two databases have substantial differences in the proportion of IAS adopters at the individual country level. One example was that the percentage of IAS adopters in Italy was as high as 78.7% according to Worldscope, but only 0.2% based on Compustat Global Vantage; while the hand-coding indicated that the percentage was 25.1%. The researchers finally created an

"augmented" Worldscope data, for which they used hand-coded data and Compustat Global Vantage data to correct the initial Worldscope coding. This practice led to 2,202 cases, for which their hand-coded data overrode the conflicting Worldscope data (Daske et al., 2013, p. 537). They reminded other researchers: "because of the large number of inconsistencies, commonly used accounting standards classifications in Worldscope and Global Vantage have to be used judiciously" (Daske et al., 2013, p. 500).

Weiß and Mühlnickel (2014) had to exclude 25 insurers from a 154-insurer sample due to incomplete balance-sheet variables in Worldscope. They noted that "excluding insurers with incomplete data from the analysis could lead to a selection bias in the results because the incompleteness of an insurer's data could be the result of the insurer's opacity. The sample could thus be biased because (presumably) systemically riskier insurance firms are systematically omitted" (Weiß & Mühlnickel, 2014, p. 33). They had to take great effort to manually check other sources to mitigate this bias (Weiß & Mühlnickel, 2014).

McGuire et al. (2016) compared the data from Compustat Global, Osiris, and Worldscope and confirmed that researchers would likely come to a different conclusion based on the database used – particularly for developing country studies. They found a significant number of missing values for reported employee data - only 17.4% of observations reported employee data in 2010 (McGuire et al., 2016). They raised another concern over the treatment of delisted or unlisted firms in Worldscope: "data on firms not currently covered in Worldscope (due to delisting, bankruptcy, merger, or deletion from the database) are not available for prior years even though the firm might have been traded, and data [has been] reported during an earlier period. Firms included in later years but not previously covered were listed as having missing data for earlier years. Therefore, the 'raw' number of Worldscope firms (not considering missing data) were identical for each year. Similar issues did not appear in the other two databases" (McGuire et al., 2016, p. 190).

Nobes and Stadler (2018), as they studied the international differences in financial reporting, addressed four data problems of Worldscope: 1) Current data. For some data fields, such as "stock exchange listed" or "industry classification," only the current data was available. This created problems for time-series research and sampling. 2) Misleading data. For example, the Worldscope data of pension discount rates for Italy's Eni was 8.35% in 2010; however, Eni's 2010 Annual Report disclosed the discount rates for obligations under "TFR" (an Italian defined benefit pension obligation) was 4.8% and for the obligations under "Foreign pension plans" was 2.7-14.0%. The data value reported in Worldscope didn't reflect any single value or sum, so the number was misleading. 3) Missing data. For instance, Worldscope systematically showed an empty field for actual return on plan assets, however, according to the pension accounting requirements, the actual return on plan assets was the sum of (i) expected return on plan assets and (ii) actuarial gains/losses related to the plan assets, which were required disclosures under the IFRS. 4) Erroneous data. They compared the "projected benefit obligation" data of the German HDAX equity index firms in the period 1998 to 2006 in Worldscope with their hand-collected data. Among 433 firm-year observations that both datasets had, they found 47 observations that Worldscope deviated more than 5% from the hand-collected data, which indicated that more than 10% of the investigated data in Worldscope were erroneous by a significant amount.

They also spotted errors resulted from the status changes of companies such as delisting from stock exchange (Nobes and Stadler, 2018).

4) SDC Platinum

SDC Platinum (hereafter, "SDC") was a premium product of the Securities Data Co, a New Jersey-based company. The company was acquired by Thomson Financial in 1988 (Dalton, 1996). In 1999, the joining of the Investext Group, Securities Data Co., and CDA/Spectrum created the Thomson Financial Securities Data (TFSD), which is a part of the Thomson Financial (Library of Congress, n.d.). The featured content of SDC includes global mergers and acquisitions (M&A) transactions, bond deals, equity capital market new issues, and global corporate loan transactions (Refinitiv, 2020). The database is used widely by scholars for empirical research on M&A and strategic alliances (Croson et al., 2004; Netter et al., 2011; Barnes et al., 2014; Keasler & Denning, 2009; Yan et al., 2020.). The database is often compared with the BvD Zephyr database (Ma et al., 2009; Bollaert & Delanghe, 2015). Since SDC provided access to the VentureXpert database, some researchers also used the database to research private equity and venture capital transactions (Rogers, 2020). The database is often used together with Worldscope, Datastream, CRSP, and Compustat (Faccio & Masulis, 2005; Netter et al., 2011; Mulherin & Aziz Simsir, 2015; Betton et al., 2018).

Faccio and Masulis (2005) collected the M&A partners' identities, country, and industry (3 digit SIC code) data from SDC. They also gathered the initial announcement date, dollar value, method of payment, and legal form of the deals. The researchers reported that due to the inconsistent entries in SDC, they had to collect the "method of payment" information from the description section, rather than the "method of payment" data field (Faccio & Masulis, 2005, p. 12). They realized that the bidder and target information for European bidders is "often missing" (Faccio & Masulis, 2005, p. 12). They doublechecked the outliers in "deal value" using LexisNexis and found many mistakes in the SDC database (Faccio & Masulis, 2005, p. 55).

Boone and Mulherin (2007) showed that the inaccurate data in SDC caused "incorrect inferences [of prior research] on the association of termination provisions with judicial decisions, bidder toeholds, and deal size" (p. 485). The inaccuracy stemmed from "the incomplete reporting of termination provisions in the SDC database" (Boone & Mulherin, 2007, p. 469). For 400 takeovers included in their sample, the researchers found "the SEC filings indicated that 91% of the sample takeovers had a termination provision. By contrast, the SDC data report termination provisions for only 66% of the takeovers, a difference of 25%" (Boone & Mulherin, 2007, p. 469). To find out if the under-reporting was just for their sample, they conducted a random sampling at a larger scale. For the 73 deals selected from the random sample, SDC classified 23 (31.5%) as having a termination fee, a stock option agreement, or both; however, the classification based on SEC documents found 57 (78%) of the 73 deals had a termination fee, a stock option agreement or both (Boone & Mulherin, 2007, p. 469). They summarized that "the difference between the SEC filings and the SDC data is especially noticeable in the early years of the sample. In 1989 and 1990, the difference is 50% or more. The differences are also 40% or more in 1993 and 1996. In 1997 and later, the differences are not as large" (Boone & Mulherin, 2007, p. 469). They also found that the accuracy of the SDC data was related to firm size (Boone & Mulherin, 2007, 472).

Banyi et al. (2008) identified several problems of SDC data in estimating share repurchases: "overall, SDC announcements of repurchases do not include all repurchase programs; [they] are poor predictors of the number of shares that will be repurchased by a firm, and are not inclusive of all repurchase authorizations" (p. 463). Chapman and Klein (2009) searched the deals of the Kohlberg Kravis Roberts & Co. in SDC and found "its first important deal, an acquisition of Houdaille Industries, was missing, and of more than 150 transactions that the company is the named buyer, only 30 appear in SDC" (p. 4). They also disclosed that in syndicated transactions, it is often difficult to distinguish the firms coded as a "buyer" from the firms coded as an "investor" (Chapman & Klein, 2009, p. 4).

Netter et al. (2011) analyzed a comprehensive set of SDC M&A data between 1992 and 2009. They found that the number of domestic deals data before 1988 in SDC was considerably less than the dataset reported by the W. T. Grimm & Co. (Netter et al., 2011, p. 2320). They believed that the subjective nature of defining the different types of M&As was a challenge - although researchers established eight types of classifications based on different characteristics of M&A transactions, "the SDC classifications are more general and are broadly based on the amount of the firm acquired" (Netter et al., 2011, p. 2321). They also expressed their concerns over the transparency of the SDC data: "SDC provides very little guidance as to how the data are collected or how the variables that classify the data are defined. This lack of guidance leaves the researcher with little help in determining if classifications regarding M&As are correct or appropriate for his or her research.… [Also, due to the lack of proper sources to compare], there is also little certainty on the degree to which the SDC database is complete, even when one of the parties in the transaction is public" (Netter et al., 2011, p. 2323).

Barnes et al. (2014) compared 20 years of data from SDC with their hand-collected dataset. They found that their hand-collected data was generally more accurate than SDC data, but the accuracy and coverage of SDC improved over time. "SDC also appeared to be fairly complete from 1984 onward; however, coverage before 1984 appears to be poor to moderate compared with our hand-collected data set" (Barnes et al., 2014, p. 795). Their investigation of the discrepancies between the two datasets found that "SDC is more prone to errors on smaller, high book-to-market acquirers with weak announcement period market responses. Preliminary analyses suggest that this potential bias is not significant, but could affect inferences when examining smaller, high book-to-market firms" (Barnes, et al., 2014, P. 793).

Bollaert and Delanghe (2015) carried out an in-depth analysis between SDC and Zephyr and tried to assess the information quality and suitability for different types of research. Their research found that SDC is more suitable than Zephyr for most M&A research. However, for the research related to the acquisitions that involve multiple acquirers and targets, Zephyr seemed more suitable than SDC (Bollaert & Delanghe, 2015, p. 97).

Mulherin and Aziz Simsir (2015) investigated the accuracy of the "original date announced" (ODA) data field in SDC. Using news articles from the LexisNexis database, they found "the actual frequency of ODA

events is more than double the frequency of the events that are reported in the SDC database, …even though the SDC fails to record a significant portion of merger-related events of all types, its accuracy is lowest for the search for buyer announcements" (Mulherin & Aziz Simsir, 2015, pp. 2, 6). Betton et al. (2018) used the "takeover announcement dates" data in SDC to study takeover rumor rationales. To verify if the associated announcement dates were accurate, they conducted a manual search of both Factiva and Google for the announcement dates and corrected "52 errors and 143 omissions found within SDC" (Betton et al., 2018, p. 274).

Professor Ritter (2019) has extensively used SDC for initial public offering research and has curated a list of errors and mistakes from the SDC database. Ritter noted that the mistakes include a wide range of misclassifications regarding the unit offers, REIT, industry classification (i.e. SIC), and errors in financial data such as sales, assets, offer prices, and market prices. The document highlighted the following mistakes in SDC: (1) In general, SDC has a high error rate on the post-issue shares outstanding. The database sometimes adds the shares issued to the post-issue number of shares outstanding, double-counting the shares issued. (2) The number of overallotment shares exercised is frequently wrong. Many IPOs are listed as having 0 shares exercised, when in fact some or all of the overallotment option was exercised. (3) SDC has some mistakes in the number of managing underwriters. In a few cases, they list the total number of underwriters in the syndicate rather than the number of managers. (4) IPOs that had financial sponsors (venture capitalists or buyouts) sometimes are not classified as such. (5) Some offerings classified by SDC as an IPO was already trading on the pink sheets or bulletin board and could be classified as a follow-on offer. Also, the database lists a lot of foreign companies that issued American Depository Receipts (ADR) or American Depository Shares and listed on the NYSE or NASDAQ as IPOs when in fact they were follow-on offerings (Ritter, 2019).

Ritter (2016) also supplemented missing or questionable data from SDC with the data from prospectuses retrieved from EDGAR and other sources. In the updates on IPO statistics, Ritter excluded ADRs in most cases because, "among other reasons, the accounting data is not always reliable (SDC sometimes makes translation mistakes)" (Ritter, 2020, p. 42)

5) Other Thomson Reuters -Refinitiv Products

a. The Thomson Mutual Fund Holdings Dataset (s12 data file)

Schwarz and Potter (2016) disclosed significant discrepancies between the SEC filings and the Thomson Mutual Fund Holdings dataset and found 20% of their sampled portfolios (77,555) included in the SEC filings were not available in Thomson's data. Zhu (2020) found the coverage of Thomson's dataset "drops significantly in 2008 and continues to deteriorate" (p. 1201). "58% of newly founded US equity mutual fund share classes in the CRSP mutual fund database from 2008 to 2015 cannot be matched to the Thomson Reuters database" (Zhu, 2020, p. 1193).

b. VentureXpertdata (via Eikon or Thomson ONE)

Gornall and Strebulaev (2015) recorded many missing and miscoded venture capital data in Thomson ONE. Kaplan and Lerner (2016) commented "there are large inconsistencies in VentureXpert and VentureSource[9] databases and a general problem of incompleteness.… Qualitatively, both show deterioration in data quality over the past decade… [VentureXpert's exit status] coverage has dropped dramatically in recent years, suggesting a lack of investment in collecting new data" (pp. 5-6). Röhm et al. (2019) pointed out that the data problems caused by the different definitions of "corporate venture capital" across venture capital databases were prevalent.

4. Bureau Van Dijk (BvD) Product: Orbis

The Bureau Van Dijk was firstly established under the name Bureau Marcel van Dijk with one of its offices in Brussels in the 1970s (World Heritage Encyclopedia, 1991). The company spun off from the Bureau Marcel van Dijk in 1991 (Bureau Van Dijk, n.d.). In 2017, the company was acquired by Moody's Analytics. Orbis is BvD's flagship database that provides public and private company data. The private company data, particularly company financials are sourced from official registers, reporting companies, and third-party data providers (Franchina, & Sergiani, 2019, p. 396). Orbis features a standardized format, which allows users to compare companies across regions and countries. The database offers a 10-year history function for academic research. Orbis Historical provides access to historical data going back 15-20 years (Bureau Van Dijk, 2020). The financial and ownership data in Orbis are widely used by researchers to study companies in European countries (Jaraitė, et al., 2013; Monasterolo et al., 2017; Katz, 2019; Succurro & Costanzo, 2019; Kalemli-Ozcan, 2019). Other BvD research products used frequently for academic research include Amadeus (focusing on companies across Europe), Osiris (information on listed, and major unlisted/delisted, companies across the globe), and Zephyr (information on M&A, IPO, private equity and venture capital deals).

Jaraitė et al. (2013) matched the Operator Holding Accounts and the Person Holding Accounts from the European Union Emissions Trading System to their parent companies using the Orbis database. The research found that over 25% of the past "global ultimate owner" data and 15% of the current "global ultimate owner" data were either not available or BvD IDs were not traceable in Orbis. Hintermann and Ludwig (2019) confirmed that they were not able to associate all accounts in EU Transactions Log with the Orbis database, especially for Person Holding Accounts. Didier et al. (2015) matched the security issuances data in SDC with the balance sheet information from the Worldscope and Orbis. They found that the matched SDC-Worldscope dataset covers a longer period (including the 1990s), but the "matched Worldscope dataset contains a smaller set of firms than the matched Orbis dataset" (Didier et al., 2015, p. 10). Monasterolo et al. (2017) identified several issues regarding the classification of entities and shareholders in Orbis. Kalemli-Ozcan et al. (2019) proposed strategies to construct nationally representative firm-level longitudinal data for 27 European countries, because they noticed that the data samples downloaded from Orbis were often not nationally representative.

[9] Dow Jones' VentureSource is often compared with VentureXpert for venture capital research. The Dow Jones discontinued its VentureSource database and services as of March 31, 2020 (Dow Jones, 2020).

Kalemli-Ozcan et al. (2019) further explained the problems of the financial module and the ownership module in Orbis and Amadeus. They found that for historical financial information, researchers may encounter the following problems: (1) download speed and cap issues. Download speed is generally slow since the product is not designed for bulk data downloads. The downloaded files may have missing information due to the data download cap. (2) survivorship bias. Amadeus will delete a company from the database if the company did not report anything in the last 5 years. Comparatively, Orbis will keep this company as long as the company is active in the business register. (3) reporting lag of about 2 years, on average. There are differences in the coverage of particular variables and the lag varies by country and by data product. (4) presentation format issue. "Certain variables, such as employment, will not be on the balance sheet, but rather in memorandum item" (Kalemli-Ozcan et al., 2019, p. 22). (5) merging issues. The unique company BvD ID number may change over time due to "changes of address, legal form, or M&A activity" or maybe changed by BvD to "harmonize the IDs across databases using a set of priority rules" (Kalemli-Ozcan et al., 2019, p. 22). For the ownership module, they mentioned that researchers may encounter problems including (1) Vintage issue. BvD browser online only contains the latest available ownership information (static or "as of date"). Historic (time-series) ownership information is only available in the company's standard report, which cannot facilitate large dataset download for academic research. (2) Merging issues. The BvD ID changes also cause problems in tracking ownership changes (Kalemli-Ozcan et al., 2019, p. 23).

Regarding why "Orbis web browser interface displays a large number of unique firm identifiers, but the actual financial or real variables, when downloaded, turn out to be missing, especially going back in time," Kalemli-Ozcan et al. (2019) explained that it may result from reporting lag, deletion of company records due to no report for a certain period, a download cap or some variables were not covered in specific data platforms such as the Wharton Research Data Services (Kalemli-Ozcan et al., 2019, p. 4). Previously, responding to the questions from the Business Information Review Survey on the inadequate coverage of the business information in Orbis, Green (2003), the Head of Marketing and Communication from BvD, explained that the confusion often came from the lack of understanding of different reporting obligations in different countries. "Some gaps in private European company information are better known than others. An example often cited is that of Germany, where large numbers of private companies flout filing obligations. Conversely, the UK and Belgium have high levels of compliance. In the Netherlands, accounts are filed but companies are often slow to submit" (Green, 2003, p. 69).

5. Other Sources

Except for the databases mentioned above, this research encountered many articles that examined other data sources. Here we provide a summary of these sources.

1) Mergent Online

Tallapally (2009) reported that hundreds of bond issue data were excluded from their sample because the data was not available in Mergent. When mining competitor relationships using Mergent Online, Ma et al. (2011) estimated that the company profiles from Mergent cover only 24.9% of all competitor pairs. Berrios (2013)

used the database to draw samples from state commercial banks and used the ownership (percent of shares held by bank insiders), compensation, tenure, and financial data to analyze the relationship between bank credit risk, profitability, and liquidity. The sampling frame included 793 public companies generated with SIC 6022 State Commercial Banks. Among the 79 US banks selected from random sampling, 39 (49%) banks were excluded from their analysis because the financial data were incomplete (Berrios, 2013). Lu and Shang (2017) obtained a sample of 1,113 companies from Mergent Online to study their supply chain structure. 246 (22%) companies were excluded from the sample due to the lack of relationship or product tree data.

2) Value Line

Anderson and Lee (1997) found that the discrepancies between the Value Line and the Spectrum databases could affect economic inferences drawn from regressions using their "ownership" data. Discrepancies were also reported between Value Line and Compustat (Kern & Morris, 1994; Yang et al., 2003). Ramnath et al. (2005) compared Value Line with I/B/E/S and found that "I/B/E/S earnings forecasts outperform Value Line significantly in terms of accuracy and as proxies for market expectations" (p. 185). However, Zhang and Alexander (2016) reviewed half a century of research on Value Line and found "the evidence on Value Line enigma is less than conclusive, … and despite some mixed results, the evidence seems to suggest that Value Line EPS data are accurate and reliable relative to those of [I/B/E/S]. Moreover, evidence strongly suggests that reporting discrepancies of financial statement data between [Value Line] and other databases exist, and the selection of database could materially affect the results of the study" (p. 812).

3) ReferenceUSA

Cook et al. (2012) reviewed the data quality of ReferenceUSA's New Businesses database and found "in the one-month sample, almost 40% of the firms had invalid phone numbers" and "the number of employees was not checked prior to its release," so they believed that the database was not vetted as they promised (p. 309). Their communication with the vendor confirmed that the database is "comprised of records gleaned from many sources and is not verified" (Cook et al., 2012, p. 310). They also mentioned that after the release of their article, the ReferenceUSA invited several librarians to visit their research center and introduced a new feature to their records: verified versus non-verified businesses (Cook et al., 2012, p. 300).

Besides, researchers also disclosed the data quality issues of regional financial data aggregators for Malaysia (Suret et al., 1998), Korea (Nam et al., 2017), and European countries (Olbrys and Majewska, 2014).

Discussion

After reviewing the literature on individual databases, we have identified several categories of data quality problems, including missing values, data errors, discrepancies, biases, inconsistencies, static header data, standardization, changes in historic data, lack of transparency, reporting time issues and misuse of data. These problems are not separate. They are intricately related to each other. Since the literature review covers nearly 50 years of publication, some of the problems identified in specific databases may no longer exist or have changed its form. This discussion is not to criticize a specific database, instead, it is to find answers to the

research questions, inform business researchers and librarians of common data problems and discuss their implications for business reference and research consultation.

Common Data Quality Problems

1. Missing Values

Missing values are one of the most prevalent data quality problems. The CRSP files were found missing delisting returns, mutual fund returns, and mandated mutual fund portfolios (Shumway, 1997; Elton et al., 2001; Wisen, 2002; Schwarz & Potter, 2016). A large number of data omissions were identified in Compustat (Boritz & No, 2013; Casey et al., 2016; Heitzman & Lester, 2020). The Omission of the "global ultimate owner" data was reported in Orbis (Jaraitė et al., 2013). Credit line data and outstanding debt data in Capital IQ were observed having a significant number of missing values (Mathers & Giacomini, 2016; Lee, 2017). The balance-sheet data, the reported employee data, and the actual return on plan assets data were found missing in Worldscope (Weiß & Mühlnickel, 2014; McGuire et al., 2016; Nobes & Stadler, 2018). The information for bidders, termination provisions, share repurchase announcements, takeover announcements, and merger-related events were found missing in SDC (Faccio & Masulis, 2005; Boone & Mulherin, 2007; Banyi et al., 2008; Mulherin & Aziz Simsir, 2015). The Bond issue data, stock financial data, competitor data, and product tree data in Mergent Online were found incomplete (Tallapally, 2009; Ma et al., 2011; Berrios, 2013; Lu & Shang, 2017).

Researchers often identify missing values by directly comparing the data with the original data in SEC filings and more recently, with the XBRL-tagged interactive data from SEC's EDGAR (Tallapally et al., 2011; Boritz & No, 2013; Chychyla & Kogan, 2015; Schwarz & Potter, 2016). Missing values happen more common to complex accounting concepts such as "net operating loss," "credit line," "outstanding debt," "actual return on plan assets" (Mathers & Giacomini, 2016; Lee, 2017; Nobes & Stadler, 2018; Heitzman & Lester, 2020). Missing values also more frequently occur to complex transactions that need great effort to track the changes such as share repurchases and mergers and acquisitions (Banyi et al., 2008; Chapman & Klein, 2009; Netter et al., 2011). Missing values are more often observed for the data in footnotes or the data that do not directly appear on financial statements, such as "actual return on plan assets" or "undrawn revolving credit" (Mathers & Giacomini, 2016; Nobes & Stadler, 2018). In Orbis, missing value can also occur due to the cap on the amount of data allowed to be downloaded (Kalemli-Ozcan et al., 2019).

Researchers sometimes take special procedures or filters to exclude missing values from the research sample. However, this practice may inevitably create omission bias or selection biases (Elton et al., 2001; Weiß & Mühlnickel, 2014). Dropping all observations that contain missing values is a naïve strategy and can have a marked effect on the statistical power of the tests (Hribar, 2016, p. 63). Excluding missing values can create misleading results and a great number of missing values may make a database not usable for specific research (Francis et al., 2016; Lee, 2017). Missing delisting returns can result in delisting bias and other unknown data biases confounding empirical results (Shumway, 1997; Shumway & Warther, 1999). Missing values can be reduced by comparing the datasets with the SEC filings or other data sources. Casey et al. (2016) proposed a

Modified Financial Statement Balancing Model to partly solve the problems of missing values or erroneous entries and restore them into usable data points such as zeros or summary amounts.

2. Data Errors

Data errors are another common data quality problems. Many arithmetic errors, coding errors, merger date errors, portfolio position errors were found in CRSP (Courtenay & Keller, 1994; Elton et al., 2001; Schwarz & Potter, 2016). Typos or rounding errors, classification errors, calculation errors, miscoded auditor variables were found in Compustat (San Miguel, 1977; Chychyla & Kogan, 2015); misclassification of entities and shareholders were found in Orbis (Monasterolo et al., 2017). The disparity between Capital IQ credit line data and the data in 10-K filings were found significant (Mathers & Giacomini, 2016). Analyst codes and earnings announcement dates in I/B/E/S were observed subject to reporting errors (Acker & Duck, 2009; Roger, 2017). Data errors in volume, prices, shares, return and total return indices, classifications, dates, and delisting information were found in Datastream (Bloom et al., 2004; Rossi, 2011; Brückner, 2013). Errors in the unit offers, industry classifications, sales, assets, offer prices, market prices, or deal values were reported in SDC. A high error rate of post-issue shares outstanding, the number of the overallotment shares exercised, and the number of managing underwriters was reported in SDC as well (Ritter, 2019).

Many researchers identify data errors by comparing different databases or by comparing the database with their hand-collected data (Beedles & Simkowitz, 1978; Courtenay & Keller, 1994). Data errors can happen in simple forms such as typos or rounding errors. But more often, data errors happen in more complex forms. Data error may result from improperly including or excluding certain accounting items in computations, such as including contract research into R&D expenses, mistreating operating and investing activities, excluding accrued imbalances payable from accounts payable (San Miguel, 1977; Shi & Zhang, 2011). Errors are more likely to occur to complex financial concepts such as "cost of goods sold," "gross profit," or "net operating loss" than to simple concepts such as "total assets," "total liabilities," or "net income" (Chychyla & Kogan, 2015; Heitzman & Lester, 2020). Errors more often happen to complex transactions such as mergers and acquisitions, changing exchanges, delisting, or stock splits (Andrikopoulos et al., 2007; Rossi, 2011; Nobes & Stadler, 2018; Ritter, 2019). Miscoding errors in auditor variables happen more often when there are auditor changes (Utke, 2018). Error rates are generally higher for items reported in the footnotes than for the items reported on the income statement or balance sheet (Kinney & Swanson, 1993; Bratten et al., 2016). Classification errors can happen due to the misunderstanding of business and financial concepts such as common stocks or industry classifications (Ince & Porter, 2006; Ritter, 2019). In terms of foreign firms or foreign transactions, errors may result from improper translation or interpretation of foreign accounting terms or foreign firm conditions (Brückner, 2013; Nobes & Stadler, 2018). Errors also arise when databases don't promptly update the data when there are changes in exchange, industry classification, auditor, restatement, etc. (Rossi, 2011; Chychyla & Kogan, 2015; Utke, 2018).

Data errors can in many ways distort the results of related studies (Acker & Duck, 2009). Rounding prices to the nearest penny may not be a hard error, but it can cause nontrivial differences in the calculated returns when prices are small (Ince & Porter, 2006). Reporting errors in analyst codes can impact the evaluation of

analysts' characteristics and may bias empirical studies that rely on tracking analysts (Roger, 2017). Large errors can influence some properties of the sample to a degree out of proportion to their small number, introducing biases and polluting statistical analyses (Rosenberg & Houglet, 1974). A large amount of inaccurate information would need researchers to independently validate the data, which undermines the value of using commercial databases (Elton et al., 2001). Data errors can be detected and reduced by cross-checking with other sources. The Modified Financial Statement Balancing Model can be used to find erroneous entries (Casey et al., 2016). Statistical sampling methods are often used to identify data errors and outliers (CRSP LLC, 2020c).

3. Discrepancies

Many researchers found discrepancies when comparing different databases or datasets (Grinblatt et al., 1984, Sarig & Warga, 1989; Guenther & Rosman, 1994; Courtenay & Keller, 1994; Kern & Morris, 1994; Kahle & Walkling, 1996; Elton et al., 2001; Yang et al., 2003; Ulbricht & Weiner, 2005; Daske et al., 2013; Bollaert & Delanghe, 2015; McGuire et al., 2016; Tobek & Hronec, 2018). Discrepancies may result from differences in database coverage, definitions, coding policies, identifiers, classifications, calculation models, selection biases, or data errors (Sarig & Warga, 1989; Courtenay & Keller, 1994; Kern & Morris, 1994; Yang et al., 2003; Tallapally, 2009; Dreyer & Hines, 2014; Nam, et al., 2017).

Most often, discrepancies are found by comparing two different databases or datasets. However, discrepancies can also happen within one database particularly in the aggregator databases that acquire data from different sources. For example, researchers found that the I/B/E/S actual EPS differs from the analyst's inferred actual EPS 39% of the time (Brown & Larocque, 2011, 2013). Discrepancies are more common for complex transactions such as stock splits, stock dividends, mergers and acquisitions, accounting changes, or discontinued operations (Grinblatt et al., 1984; Courtenay & Keller, 1994, Kern & Morris, 1994). Discrepancies are more prevalent for data items that rely on a subjective assignment such as SIC code (Guenther & Rosman, 1994; Kahle & Walkling, 1996; Stasch, 2014); Discrepancies may happen due to the standardization procedures in databases which may not necessarily benefit empirical research (Tallapally et al., 2012; Chychyla & Kogan, 2014). A higher level of discrepancies among databases is identified for international data, particularly the data for developing countries (Lara et al., 2006; Daske et al., 2013; McGuire et al., 2016; Hines & Sharma, 2018). The great disparity in data availability between different countries is largely due to the different accounting practices and filing rules based on different historical and cultural practices (Green, 2007). Firat (2002) identified at least three types of heterogeneities (data-level, ontological and temporal heterogeneities) between databases. The differences arise when databases choose different units, scales, or formats to represent the same firm's data; or when they apply different accounting definitions, currency conversion policy; or when entity values or definitions belong to different times, or time intervals (Firat, 2002).

Discrepancies across databases imply that it would be risky to rely on any single database (McGuire, 2016). Discrepancies can further lead to the "database effect," which means researchers would come to different conclusions based on different databases. Differences in databases can affect sample selections, the inferences

about the population, and can further affect the outcome of empirical research (Kern & Morris, 1994; Kahle & Walkling, 1996; Lara et al., 2006; McGuire et al., 2016). Many researchers choose to use multiple data sources for their research; however, great effort needs to be taken to compare coding policies, matching, filtering, and cleaning data (Chakrabarty & Trzcinka, 2006).

4. Biases

The concerns about biases in databases are widely discussed. Upward bias, delisting bias, omission bias, survivorship bias, incubation bias, backfill bias (or 'instant-history' bias), and duplication bias were found in CRSP (Loughran & Ritter, 1995; Shumway, 1997; Canina et al., 1998; Shumway & Warther, 1999; Elton et al., 2001; Wisen, 2002; Yan, 2007; Evans, 2010; Jorion and Schwarz, 2017; CRSP, 2020). Survivorship bias, ex-post-selection bias, and look-ahead bias were found in Compustat (Ball, 1979; Banz & Breen, 1986). Survivorship bias and selection bias were reported in Datastream and Orbis (Ince & Porter, 2006; Andrikopoulos et al., 2007; Kalemli-Ozcan et al., 2019).

Biases can happen for various reasons. Overstatement by a statistical measure or index can result in an upward bias (Rosenberg & Houglet, 1974; Loughran & Ritter, 1995). When observations are excluded from the sample due to a selection rule other than random sampling, it can create selection bias. Survivorship bias is an example of the selection bias driven by the disproportionate exclusion of stocks that were delisted over time (Waszczuk, 2014). Incubation bias is an example of selection bias due to excluding poor-performing new funds from adding or promptly adding to a database (Wisen, 2002). Backfill bias arises when the fund's performance is not made public during some incubation period but then is added to the database presumably following the good performance (Jorion & Schwarz, 2017). Look-ahead bias occurs when data used in the study is assumed to be publicly available at a specific time while in reality, the data is only available at a later time (Andrikopoulos et al., 2007). Omitting delisting returns for a large number of companies can result in delisting bias (Shumway & Warther, 1999; Waszczuk, 2014). Biases are likely to occur to new firms, small firms, foreign firms, delisting firms, and firms with poor performance (Andrikopoulos et al., 2007; Landis & Skouras, 2018). Biases can distort empirical research results and produce unreliable conclusions. Some researchers try to remedy the survivorship bias by adding delisted firms back to their sampling. However, reducing biases greatly relies on the efforts of database vendors in improving data quality and data integrity.

5. Inconsistencies

Inconsistencies depict a situation where the data is not treated consistently with the same rule, particularly in the same database. Inconsistencies need to be further investigated for potential data errors. Inconsistency problems were reported in Compustat, Orbis, and SDC databases (McElreath & Wiggins, 1984, Faccio & Masulis, 2005; Shi & Zhang, 2011; Kalemli-Ozcan et al., 2019). One example of inconsistency is that the same accounting items are defined differently in different financial statements. It was reported that in Compustat "depreciation and amortization expenses" reported on the income statement and the cash flow statement included different items. The special events such as "selling of assets" and "inventory write-offs" were adjusted in the cash flow statement but not in the balance sheet. The restated amount reflected in the cash flow statement was not adjusted in the balance sheet. The change of classifications reflected in the

balance sheet was not updated in the cash flow statement (Shi & Zhang, 2011). Moreover, the inconsistency may occur to different periods, for example, Compustat's definition of accounts payable in 2002 and 2003 were different (Shi & Zhang, 2011). Inconsistency issues can also happen to proprietary identifiers. A proprietary identifier is often recognized as the unique ID to track a company; however, it may not be consistently used this way. Researchers disclosed that a company's BvD ID can change due to changes in address, legal form, or M&A activities, which can cause problems in tracking changes of a company (Kalemli-Ozcan et al., 2019).

Researchers may assume that databases from the same vendor have the same reporting policy. But Mathers and Giacomini (2016) reminded researchers that Compustat and Capital IQ both from the S&P Global Market Intelligence have a different coding policy on a fiscal year (Mathers & Giacomini, 2016). I/B/E/S and Worldscope were both from Thomson Reuters, while 22% discrepancies between announcement dates were reported in these two databases (Acker & Duck, 2009). Information from different sections of a database may be inconsistent as well. Researchers reported that in SDC "method of payment" data obtained from the description section and the "method of payment" variable field have frequent differences (Faccio & Masulis, 2005). Inconsistent data can create comparability problems (McElreath & Wiggins, 1984). It can cause similar problems to data errors and sometimes mislead researchers. Since it takes great effort for researchers to notice and resolve inconsistencies in databases, the problem undermines the value of commercial databases.

6. Static Header Data

The static header data issue (or vintage issue) happens when the database only provides the latest available information and lacks time-series records (Kalemli-Ozcan et al., 2019). Static header data issue happens mostly in the data field including "company name," "ticker," "address," headquarter," "ownership," "stock exchange," and "industry code" (Ince & Porter, 2006; Hines, 2016; Landis & Skouras, 2018; Nobes & Stadler, 2018; Kalemli-Ozcan et al., 2019). Such problems were observed in Compustat, Orbis, Worldscope, and Datastream (Ince & Porter, 2006; Hines, 2016; Landis & Skouras, 2018; Nobes & Stadler, 2018; Kalemli-Ozcan et al., 2019). The static header data erase the track record of prior changes, so it restricts researchers from including these data variables into time-series analysis. Since the data may not accurately represent the point in time, when using static header data fields to generate samples, it may induce incorrect samples, omissions, or selection biases (Landis & Skouras, 2018).

7. Standardization

Standardization and adjustments make the data more comparable across companies, over time, and particularly make it possible to compare companies from different countries. The obvious benefit of standardization makes it a selling point of several databases such as Compustat, Worldscope, and Orbis. However, standardization also causes data problems. The standardization and adjustment of "depreciation, depletion, and amortization" data in Compustat were found "understated 10-K cost of goods sold by 7.5%" and "overstate 10-K gross margin by 14.3%" (Bostwick et al., 2016). Researchers also found the standardization in Compustat "yields no improvements for bankruptcy prediction models and a significant negative impact on the predictive accuracy of Altman's model" (Chychyla & Kogan, 2014, P. 1). The standardization of stock-

split adjusted I/B/E/S data and rounding to the nearest penny causes loss of information (Payne & Thomas, 2003). So, researchers should be more deliberate of their choices of standardized vs. non-standardized datasets. Extra efforts are needed to understand the standardization process and changes made to the original datasets due to such a process.

8. Changes in Historical Data

Historical data can be revised due to the correction of errors or updates from restatement. But systematic changes in historic data can be a problem. Researchers used I/B/E/S analyst stock recommendations found date files for the same period downloaded at different times have substantial differences. The differences include recommendation levels, additions and deletions of records, removal of analyst names, and changes in the attributes of the earnings forecasts available in each version (Ljungqvist et al., 2008; Call et al., 2020). The seemingly not-random alterations raised concerns over data integrity (Brown-Humes, 2006). Changes in historical data can be very common due to the fluid nature of commercial databases, especially for data aggregators that acquire the data from other primary data sources. Changes in primary data providers, contributors, and contract terms can all lead to changes in databases. In the instances that a database gives data contributors direct controls over the databases, the changes are more likely to occur and hard to manage.

9. Lack of Transparency

Lack of transparency is a fundamental issue for many other data problems. Generally, database vendors are not transparent about their data collection and management practices, let alone warning researchers about potential data problems and biases (Annaert, et al., 2016). Researchers expressed concerns over the transparency of SDC data on how the data are collected or how the variables are defined (Netter et al., 2011). Although data issues were often not publicly explained, sometimes they are disclosed through private conversations between researchers and database providers. Being questioned about the changes of the historical data of individual analysts, I/B/E/S responded that the names of the individual analysts remain in the database, but they were not visible on the files seen by the academics due to an incomplete data feed (Brown-Humes, 2006). When asked about the changes in historic earnings forecasts, I/B/E/S explained the retroactive adjustments may occur due to stock splits, stock dividends, default currency adjustments or correction of errors and further disclosed some differences occurred because the brokerage maintains control over the distribution of these forecasts and academic subscribers often have access to only a subset of all the earnings forecasts contributed to the database (Call et al., 2020). Despite the explanations from vendors, the lack of transparency concerns researchers and greatly affects researchers' trust in commercial databases (Call et al., 2020).

10. Reporting Time Issues

Data reporting time can be a problem when database providers are not transparent about their data reporting time and update schedule. It can affect research when the dates that the data is available to the public are different from the date that a study assumes it is (McElreath & Wiggins, 1984). Improperly recorded reporting time can induce look-ahead bias (Andrikopoulos, et al., 2007). Reporting lag or delay is a different problem that sometimes is inevitable. In the case that the data are aggregated from different vendors, the

reporting lag may be caused by specific embargo contract terms. In the case that the data are sourced from different countries, reporting lag may also be caused by different reporting compliance practices in different countries (Green, 2003). The reporting lag in Orbis was found to be about 2 years on average and varied by country and by data product (Kalemli-Ozcan et al. 2019). The reporting lag due to incubation (when new funds with poor performance are not added to databases as promptly as new funds with superior performance) can result in selection bias (Wisen, 2002).

11. Misuse of Data

Data problems are not limited to data itself, researchers may sometimes improperly use the data as proxies or measurements. For example, the practice of compounding daily returns of the CRSP equal-weighted index to calculate monthly returns could lead to large biases (Canina et al., 1998). CRSP adjustment factors for share price have included the effects of property dividend, spin-off, and rights offering events, so researchers must only use the CRSP adjustment factors to accommodate events that lack such economic substance, otherwise, it may create erroneous sample observations and misleading results (Francis et al., 2016). Mills et al. (2003) warned researchers to take extra caution when using Compustat net operating loss data as an indicator of a firm's US tax-loss positions, particularly when the research setting involves the firms with foreign operations or corporate acquisitions activity. Moreover, since Compustat only covers public firms in an industry, it is a poor proxy for actual industry concentration. Constructing industry concentration measures using Compustat data can lead to incorrect conclusions (Ali et al., 2008; Keil, 2017). Announcements of repurchases in SDC are poor predictors of the number of shares that will be repurchased by a firm, because the data is not complete (Banyi et al., 2008).

Throughout the research, we found some of the data quality problems can be alleviated through the improvement in statistical testing methods, data profiling and mining techniques, and the adoption of new data reporting systems and policies. However, the improvement of overall data quality largely depends on transparent quality control practices of data providers and their open engagement with the research community.

Implications for Business Reference and Research Consultation

Helping library users and researchers understand information quality is an essential part of library reference services. Librarians have paid special attention to the information quality dimensions such as reliability, validity, accuracy, authority, timeliness, and biases (ACRL, 2000). We also developed practical approaches such as the CRAAP (Currency, Relevance, Authority, Accuracy, and Purpose) test to facilitate information literacy education (Blakeslee, 2004). However, this research finds the data quality problems can be much more complicated and troubling. Business librarians need to be aware of these data quality problems and raise researchers' awareness of these problems through business reference and consultation. Below is some practical advice that we can provide to researchers.

- Researchers should not take the quality of the reputable commercial databases for granted. They should always evaluate data quality and check its accuracy and completeness. They need to use caution with projects studying complex accounting concepts or business transactions, or projects involving new firms, small firms, foreign firms, or delisting firms.
- Researchers should not solely rely on a single data source. If possible, they should consult multiple sources, especially the original data sources. They should pay attention to the different definitions and coding policies of different databases and consider the discrepancies between databases as an opportunity to identify missing values or data errors. They should err on the side of caution for the "database effect" and test their theories through multiple data sources.
- Researchers should carefully read the data manuals and understand how the data is defined or calculated, especially for standardized or adjusted data. They should be aware of the inconsistencies of the database in treating such definitions and adjustments across financial statements, historical periods, and throughout the database.
- Researchers should be cautious of using databases as a screening tool to identify data samples, especially using the static header data field as a variable to screen data samples. They need to assess if the sample data are proper proxies or measurements. The incomplete data and incorrect classifications in databases may cause missing or incorrect records in data samples or even induce selection bias and misleading results.
- Researchers should not ignore the biases in databases. Instead, they need to seek proper procedures to mitigate the biases and clearly explain the impact of potential biases and the limitations of their research.
- Researchers should not treat data acquisition as a one-time transaction. Instead, they need to understand the fluid nature of a research database, preserving the data at different points in time, and leaving a paper trail of data access and usage.
- Researchers need to be cognizant of the reporting time, reporting lag, update schedule, or embargo period of the data sources. They may need to adjust their data retrieval or update practices accordingly.
- Researchers should keep open communications with vendors on data problems. Their communication will allow the vendors to identify similar problems or initiate projects to make large scale changes.

Conclusions

This article provides a literature review on business and financial literature that addresses data quality problems and covers the databases including CRSP, Compustat, Capital IQ, I/B/E/S, Datastream, Worldscope, SDC, and BvD Orbis. The synthetic analysis on the business literature identified 11 categories of common data quality problems, which include missing values, data errors, discrepancies, biases, inconsistencies, static header data, standardization, changes in historic data, lack of transparency, reporting time issues and misuse of data. These data problems can in many ways introduce errors and biases into empirical research, polluting statistical analysis, and distort research results. Many researchers have overturned

prior research results after correcting specific data problems. Despite the prevalence of data quality problems, these databases are widely used by academic researchers to conduct empirical studies in accounting, finance, economics, math, and statistics, by the commercial market for backtesting and modeling calculations, and by government agencies for financial and economic analysis. Academic research has long been trusted as a reliable way to create knowledge and achieve scientific and theoretical advances in related areas. Evidence-based and data-driven decision making has been widely applauded as a more reliable decision practice. The data quality problems will not only undermine the value of academic research, mislead business decisions, confound government policies but also damage public trust in knowledge. Librarians have played a crucial role in assisting the research community in accessing data sources and educating researchers. Hopefully, this article will help facilitate librarians' communication with researchers on data quality problems and will raise awareness of the research community in this regard.

Appendix I List of Reviewed Articles

Article (Author-Date)	Journal Title	Summary (provided for the database with over five cited articles.)
Center for Research in Security Prices (CRSP)		
Rosenberg and Houglet (1974)	The Journal of Finance	Period Coverage: Number of Articles
Beedles and Simkowitz (1978)	The Journal of Finance	1970-1999: 12
Bennin (1980)	The Journal of Finance	2000-2009: 4
Grinblatt et al. (1984)	Journal of Financial Economics	2010-2019: 5
Sarig and Warga (1989)	Journal of Financial and Quantitative Analysis	Journal Coverage: Number of Articles
Guenther and Rosman (1994)	Journal of Accounting and Economics	The Journal of Finance: 9
Courtenay and Keller (1994)	The Accounting Review	Social Science Research Network (SSRN): 4
Loughran and Ritter (1995)	The Journal of Finance	Journal of Financial and Quantitative Analysis: 2
Kahle and Walkling (1996)	Journal of Financial and Quantitative Analysis	Journal of Financial Economics: 1
Shumway (1997)	The Journal of Finance	The Accounting Review:1
Canina (1998)	The Journal of Finance	The Review of Financial Studies: 1
Shumway and Warther (1999)	The Journal of Finance	Journal of Accounting and Economics: 1
Elton et al. (2001)	The Journal of Finance	Working Paper: 1
Wisen (2002)	SSRN	Manuscript: 1
Evans (2007)	Manuscript	Total: 21
Yan (2007)	SSRN	
Evans (2010)	The Journal of Finance	
Schwarz and Potter (2016)	The Review of Financial Studies	
Francis et al. (2016)	SSRN	
Jorion and Schwarz (2017)	SSRN	
Tobek and Hronec (2018)	IES Working Paper	
Compustat		
San Miguel (1977)	The Accounting Review	Period Coverage: Number of Articles
Ball (1979)	The Journal of Finance	1970-1999: 6
McElreath and Wiggins (1984)	Financial Analysts Journal	2000-2009: 6

Banz and Breen (1986)	The Journal of Finance	2010-2019: 15
Kinney and Swanson (1993)	The Journal of the American Taxation Association	2020- : 1
Kern and Morris (1994)	The Accounting Review	Journal Coverage: Number of Articles
Mills et al. (2003)	The Journal of the American Taxation Association	SSRN: 5
Yang et al. (2003)	Industrial Management & Data Systems	The Journal of Finance: 2
Ulbricht and Weiner (2005)	SSRN	Accounting Horizons: 2
Ali et al. (2008)	The Review of Financial Studies	The Accounting Review: 2
Banyi et al. (2008)	Journal of Corporate Finance	Journal of Corporate Finance: 2
Tallapally (2009)	Dissertation	Journal of Financial Reporting: 2
Tallapally et al. (2011)	Review of Business Information Systems	The Journal of the American Taxation Association: 2
Shi and Zhang (2011)	Accounting Horizons	Dissertation: 2
Tallapally et al. (2012)	Manuscript	Accounting, Organizations and Society: 1
Boritz and No (2013)	SSRN	Advances in Accounting: 1
Chychyla and Kogan (2014)	SSRN	Financial Analysts Journal: 1
Williams (2015)	Dissertation	Industrial Management & Data Systems: 1
Chychyla and Kogan (2015)	Journal of Information Systems	Journal of Information Systems: 1
Bratten et al. (2016)	Accounting, Organizations and Society	Journal of International Management: 1
Bostwick et al. (2016)	Accounting Horizons	Review of Business Information Systems: 1
McGuire et al. (2016)	Journal of International Management	The Review of Financial Studies: 1
Casey et al. (2016)	Journal of Financial Reporting	Manuscript: 1
Hribar (2016)	Journal of Financial Reporting	Total: 28
Keil (2017)	Journal of Corporate Finance	
Utke (2018)	Advances in Accounting	
Casey et al. (2019)	SSRN	
Heitzman and Lester (2020)	SSRN	
I/B/E/S		
Payne and Thomas (2003)	The Accounting Review	Period Coverage: Number of Articles
Ljungqvist et al. (2008)	The Journal of Finance	2000-2009: 3; 2010-2019: 3; 2020 - : 1
Acker and Duck (2009)	SSRN	Journal Coverage: Number of Articles
Brown and Larocque (2011; 2013)	The Accounting Review	SSRN: 2
Roger (2017)	Finance Research Letters	The Accounting Review: 2
Call et al. (2020)	SSRN	The Journal of Finance: 1
		Finance Research Letters: 1
		Total: 6
Datastream		
Bloom et al. (2004)	Manuscript	Period Coverage: Number of Articles
Ince and Porter (2006)	Journal of Financial Research	2000-2009: 5; 2010-2019: 5
Lara et al. (2006)	Abacus	Journal Coverage: Number of Articles
Andrikopoulos et al. (2007)	Occasional Paper Series Paper	SSRN: 2
Espenlaub et al. (2009)	European Journal of Finance	Working Paper: 2
Rossi (2011)	Working Paper	Journal of Financial Research: 1
Brückner (2013)	SSRN	Abacus: 1
Landis and Skouras (2018)	SSRN	Occasional Paper Series Paper: 1

Tobek and Hronec (2018)*	IES Working Paper	European Journal of Finance: 1
Nobes and Stadler (2018)	The British Accounting Review	The British Accounting Review: 1
		Manuscript:1
		Total: 10
Worldscope		
Firat et al. (2002)	MIT Working Paper	Period Coverage: Number of Articles
Acker and Duck (2009)*	SSRN	2000-2009: 3; 2010-2019: 4
Ulbricht and Weiner (2005)*	SSRN	Journal Coverage: Number of Articles
Daske et al. (2013)	Journal of Accounting Research	SSRN: 2
Weiß and Mühlnickel (2014)	Journal of Financial Stability	Working Paper: 1
McGuire et al. (2016)*	Journal of International Management	Journal of Accounting Research: 1
Nobes and Stadler (2018)*	The British Accounting Review	Journal of Financial Stability: 1
		Journal of International Management: 1
		The British Accounting Review: 1
		Total: 7
SDC Platinum		
Faccio and Masulis (2005)	The Journal of Finance	Period Coverage: Number of Articles
Boone and Mulherin (2007)	The Review of Financial Studies	2000-2009: 4; 2010-2019: 6
Banyi et al. (2008)*	Journal of Corporate Finance	Journal Coverage: Number of Articles
Chapman and Klein (2009)	SSRN	Journal of Corporate Finance: 2
Netter et al. (2011)	The Review of Financial Studies	The Review of Financial Studies: 2
Barnes et al. (2014)	The Financial Review	The Financial Review: 1
Bollaert and Delanghe (2015)	Journal of Corporate Finance	The Journal of Finance: 1
Mulherin and Aziz Simsir (2015)	Financial Management	Financial Management: 1
Betton (2018)	International Review of Financial Analysis	SSRN: 1
Ritter (2016, 2019, 2020)	Manuscript	International Review of Financial Analysis
		Manuscript: 1
		Total: 10
Bureau Van Dijk (BvD) Orbis		
Jaraitė et al. (2013)	SSRN	
Didier et al. (2015)	NBER Working Paper	
Monasterolo et al. (2017)	Climatic Change	
Hintermann and Ludwig (2019)	University of Basel Working Paper	
Kalemli-Ozcan et al. (2019)	NBER Working Paper	
Capital IQ		
Benedettini et al. (2013)	Cambridge Service Alliance News	
Mathers and Giacomini (2016)	The Financial Review	
Lee (2017)	Conference Paper	
VentureXpert data (via Eikon or Thomson One)		
Gornall and Strebulaev (2015)	SSRN	
Kaplan and Lerner (2016)	NBER Working Paper	
Röhm et al. (2019)	Finance Research Letters	
Thomson Mutual Fund Holdings Dataset		

Schwarz and Potter (2016)*	The Review of Financial Studies	
Zhu (2020)	Management Science	
Mergent Online		
Tallapally (2009)*	Dissertation	
Ma et al. (2011)	Electronic Commerce Research and Applications	
Berrios (2013)	Journal of Business and Finance Research	
Lu and Shang (2017)	Journal of Operations Management	
Value Line		
Kern and Morris (1994)*	The Accounting Review	
Anderson and Lee (1997)	Journal of Financial and Quantitative analysis	
Yang et al. (2003)*	Industrial Management & Data Systems	
Ramnath et al. (2005)	International Journal of Forecasting	
Zhang and Alexander (2016)	Managerial Finance	
Reference USA		
Cook et al. (2012)	Journal of Business & Finance Librarianship	
Database from Foreign Countries		
Suret et al., (1998)	Asia-Pacific Journal of Accounting	
Olbrys and Majewska (2014)	Pensee Journal	
Nam et al. (2017)	Sustainability	

* indicates that the article reviews multiple databases and has a duplicate record.

References

Acharya, V., Almeida, H., Ippolito, F., & Perez, A. (2014). Credit lines as monitored liquidity insurance: Theory and evidence. *Journal of Financial Economics*, *112*(3), 287-319.

Acker, D., & Duck, N. W. (2009). *On the reliability of I/B/E/S earnings announcement dates and forecasts*. Social Science Research Network. https://ssrn.com/abstract=1505360

ACRL. (2000). *Information Literacy Competency Standards for Higher Education*. https://alair.ala.org/handle/11213/7668

Alfaro, L., Asis, G., Chari, A., & Panizza, U. (2019). Corporate debt, firm size and financial fragility in emerging markets. *Journal of International Economics*, *118*, 1-19.

Ali, A., Klasa, S., & Yeung, E. (2008). The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *The Review of Financial Studies*, *22*(10), 3839-3871.

Anderson, R. C., & Lee, D. S. (1997). Ownership studies: The data source does matter. *Journal of Financial and Quantitative analysis, 32*(3), 311-329.

Andrikopoulos, P., Daynes, A., Pagas, P., & Latimer, D. (2007). UK market, financial databases and evidence of bias. *Occasional Paper Series Paper, (79). http://www.dmu.ac.uk/documents/business-and-law-documents/business/occasional-papers/paper79ukmarketfinancialdatabasesandrikopoulos.pdf.*

Annaert, J., Buelens, F., & Riva, A. (2016). Financial history databases: Old data, old issues, new insights?. In D. Chambers & E. Dimson (Eds), *Financial Market History* (pp. 44-65). CFA Institute Research Foundation.

Ball, R., & Watts, R. (1979). Some additional evidence on survival biases. *The Journal of Finance, 34*(1), 197-206. doi:10.2307/2327153

Banyi, M. L., Dyl, E. A., & Kahle, K. M. (2008). Errors in estimating share repurchases. *Journal of Corporate Finance*, *14*(4), 460-474.

Banz, R. W., & Breen, W. J. (1986). Sample-dependent results using accounting and market data: some evidence. *The Journal of Finance*, *41*(4), 779-793.

Barnes, B. G., L. Harp, N., & Oler, D. (2014). Evaluating the SDC mergers and acquisitions database. *Financial Review*, 49(4), 793-822.

Beedles, W. L., & Simkowitz, M. A. (1978). A note on skewness and data errors. *the Journal of Finance*, *33*(1), 288-292.

Benedettini, O., Swink, M., & Neely, A. (2013). Firm's characteristics and servitization performance: A bankruptcy perspective. *Cambridge Service Alliance News*, 1-11.

Bennin, R. (1980). Error rates in CRSP and COMPUSTAT: A second look. *The Journal of Finance*, *35*(5), 1267-1271.

Berrios, M. R. (2013). The Relationship between Bank Credit Risk and Profitability and Liquidity. *International Journal of Business and Finance Research, 7*(3), 105–118.

Betton, S., Davis, F., & Walker, T. (2018). Rumor rationales: The impact of message justification on article credibility. *International Review of Financial Analysis*, 58, 271-287.

Blakeslee, S. (2004).The CRAAP Test. LOEX Quarterly 31(3), 6-7. https://commons.emich.edu/cgi/viewcontent.cgi?article=1009&context=loexquarterly

Bloom, N., Klemm, A., Newton-Smith, R., & Vlieghe, J. (2004). *Technical appendix: Using company accounts data from Datastream*. https://pdfs.semanticscholar.org/ce7e/ec81bcf056b677ec526aad0a50c6829ad6ea.pdf.

Bloomberg Business News. (1995, May 30). Primark to buy VNU unit in financial information. *New York Times*. Retrieved from ProQuest Database.

Bollaert, H., & Delanghe, M. (2015). Securities Data Company and Zephyr, data sources for M&A research. *Journal of Corporate Finance*, *33*, 85-100.

Boone, A. L., & Mulherin, J. H. (2007). Do termination provisions truncate the takeover bidding process?. *The Review of Financial Studies*, *20*(2), 461-489.

Boritz, J. E., & No, W. G. (2013). *The quality of interactive data: XBRL versus Compustat, Yahoo Finance, and Google Finance*. Social Science Research Network. Retrieved from https://ssrn.com/abstract=2253638

Bostwick, E. D., Lambert, S. L., & Donelan, J. G. (2016). A wrench in the COGS: An analysis of the differences between Cost of Goods Sold as Reported in Compustat and in the Financial Statements. *Accounting Horizons*, *30*(2), 177-193.

Bratten, B., Jennings, R., & Schwab, C. M. (2016). The accuracy of disclosures for complex estimates: Evidence from reported stock option fair values. *Accounting, Organizations and Society*, *52*, 32-49.

Brown, L. D., & Larocque, S. (2011). *Discrepancy between I/B/E/S Actual EPS and Analysts' Inferred EPS*. https://www.academia.edu/24143157/Discrepancy_between_I_B_E_S_Actual_EPS_and_Analysts_Inferred_EPS .

Brown, L., D. & Larocque, S. (2013). I/B/E/S Reported Actual EPS and Analysts' Inferred Actual EPS. *The Accounting Review, 88*(3), 853-880.

Brown-Humes, C. (2006). 20,000 analyst names 'missing'. *Financial Times*. https://www.ft.com/content/ce9fb8d0-6ea8-11db-b5c4-0000779e2340

Brückner, R. (2013). *Important characteristics, weaknesses and errors in German equity data from Thomson Reuters Datastream and their implications for the size effect*. Social Science Research Network. https://ssrn.com/abstract=2243816

Bureau Van Dijk. (2020). *Orbis*. Retrieved June 23, 2020, from https://www.bvdinfo.com/en-us/our-products/data/international/orbis

Bureau Van Dijk. (n.d.). *Company history*. https://www.bvdinfo.com/en-gb/about-us#secondaryMenuAnchor2

Call, A. C., Hewitt, M., Watkins, J and Yohn, T. L. (2020), *Analysts' annual earnings forecasts and changes to the I/B/E/S Database*. Social Science Research Network. https://ssrn.com/abstract=2788140

Canina, L., Michaely, R., Thaler, R., & Womack, K. (1998). Caveat compounder: A warning about using the daily CRSP equal-weighted index to compute long-run excess returns. *The Journal of Finance*, *53*(1), 403-416.

Casey, R. J., Gao, F., Kirschenheiter, M. T., Li, S., & Pandit, S. (2016). Do Compustat financial statement data articulate?. *Journal of Financial Reporting*, *1*(1), 37-59.

Casey, R., Gao, F., Kirschenheiter, M., Li, S., & Pandit, S. (2019). *Measuring reporting quality*. Social Science Research Network. https://ssrn.com/abstract=3126504

Chakrabarty, B., & Trzcinka, C. (2006). Momentum: Does the database make a difference?. *Journal of Financial Research, 29*(4), 441-462.

Chang, K., & Shim, H. (2017). Employee treatment and the choice of liquidity: lines of credit versus cash holdings. *Applied Economics Letters*, *24*(18), 1294-1297.

Chapman, J. L., & Klein, P. G. (2009). *Value creation in middle-market buyouts: A transaction-level analysis*. Social Science Research Network. http://ssrn.com/abstract=1372381

Choi, J., Hackbarth, D., & Zechner, J. (2018). Corporate debt maturity profiles. *Journal of Financial Economics*, *130*(3), 484-502.

Chychyla, R., & Kogan, A. (2014). *Does Compustat data standardization improve bankruptcy prediction models?*. Social Science Research Network. http://ssrn.com/abstract=2406136.

Chychyla, R., & Kogan, A. (2015). Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings. *Journal of Information Systems*, *29*(1), 37-72.

CNBC. (2020, July 31). *London Stock Exchange may sell Milan bourse to secure Refinitiv deal*. https://www.cnbc.com/2020/07/31/london-stock-exchange-may-sell-milan-bourse-to-secure-refinitiv-deal.html

Collings, R. (2000). I/B/E/S launches active express to easily accessed data. *Investment Management Weekly, 13*(48), 4.

Cook, R. G., Campbell, D. K., & Kelly, C. (2012). An issue of trust: are commercial databases really reliable?. *Journal of Business & Finance Librarianship*, *17*(4), 300-312.

Courtenay, S. M., & Keller, S. B. (1994). Errors in databases revisited: An examination of the CRSP shares-outstanding data. *Accounting Review*, 285-291.

Croson, R. T., Gomes, A., McGinn, K. L., & Nöth, M. (2004). Mergers and acquisitions: An experimental analysis of synergies, externalities and dynamics. *Review of Finance*, 8(4), 481-514.

CRSP LLC. (2020a). *CRSP/Compustat Merged Database*. Retrieved June 20, 2020, from http://www.crsp.org/products/research-products/crspcompustat-merged-database

CRSP LLC. (2020b). About CRSP. Retrieved June 20, 2020, from http://www.crsp.org/about-crsp

CRSP LLC. (2020c). CRSP Survivor-Bias-Free US Mutual Fund guide for CRSPSift. Retrieved June 23, 2020, from http://www.crsp.org/products/documentation/crsp-survivor-bias-free-us-mutual-fund-guide-crspsift

CRSP LLC. (n.d.a). *CRSP*. Retrieved June 26, 2020, from http://www.crsp.org/products/documentation/crsp

CRSP LLC. (n.d.b). *Research data*. Retrieved June 26, 2020, from http://www.crsp.org/products/research-products

CRSP LLC. (n.d.c). *Why CRSP*. Retrieved June 26, 2020, from http://www.crsp.org/main-menu/why-crsp

Dalton, J. (1996). *Slaine to step down at Thomson*. Bizjournals. https://www.bizjournals.com/boston/stories/1996/12/02/story1.html

Daske, H., Hail, L., Leuz, C., & Verdi, R. (2013). Adopting a label: Heterogeneity in the economic consequences around IAS/IFRS adoptions. *Journal of Accounting Research, 51*(3), 495-547.

Derasse, V. (2017, October 3). *Datastream at 50: still the best tool for macro research*. Retrieved June 23, 2020, from https://www.refinitiv.com/perspectives/market-insights/datastream-at-50-still-the-best-tool-for-macro-research/

Dhaliwal, D., Li, O. Z., Tsang, A., & Yang, Y. G. (2014). Corporate social responsibility disclosure and the cost of equity capital: The roles of stakeholder orientation and financial transparency. *Journal of Accounting and Public Policy*, *33*(4), 328-355.

Didier, T., Levine, R., & Schmukler, S. L. (2015). *Capital market financing, firm growth, and firm size distribution*. National Bureau of Economic Research. https://www.nber.org/papers/w20336.pdf

Dow Jones. (2020, June 23). *Dow Jones VentureSource and LP Source*. Retrieved June 23, 2020, from https://www.dowjones.com/products/pevc/

Dreyer, K. & Hines, T. (2014, September 30). *Linking financial data sets: possible problems and solutions* [PowerPoint Slides]. https://files.stlouisfed.org/files/htdocs/conferences/btn2014/docs/papers/dreyer-hines.pdf

Elton, E. J., Gruber, M. J., & Blake, C. R. (2001). A first look at the accuracy of the CRSP Mutual Fund Database and a comparison of the CRSP and Morningstar Mutual Fund Databases. *Journal of Finance, 56*(6), 2415–2430.

Espenlaub, S., Iqbalb, A. & Stronga, N. (2009). Datastream returns and UK open offers, *European Journal of Finance, 15*(1), 61-69.

Esqueda, O. A., & O'Connor, T. (2020). Corporate governance and life cycles in emerging markets. *Research in International Business and Finance*, *51*, 101077.

Evans, R. B. (2007). *The incubation bias* [Unpublished Manuscript]. Darden Graduate School of Business. The University of Virginia. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.192.5338&rep=rep1&type=pdf

Evans, R. B. (2010). Mutual fund incubation. *The Journal of Finance*, *65*(4), 1581-1611.

Faccio, M., & Masulis, R. W. (2005). The choice of payment method in European mergers and acquisitions. *The Journal of Finance*, 60(3), 1345-1388.

Firat, A., Madnick, S., & Grosof, B. (2002). *Knowledge integration to overcome ontological heterogeneity: Challenges from financial information systems*. MIT Sloan School of Management. Massachusetts Institute of Technology. https://dspace.mit.edu/bitstream/handle/1721.1/3683/CS016.pdf?sequence=2

Franchina, L., & Sergiani, F. (2019, September). High quality dataset for machine learning in the business intelligence domain. In *Proceedings of SAI Intelligent Systems Conference* (pp. 391-401). Springer, Cham.

Francis, R. N., Mubako, G., & Olsen, L. (2016). *Archival research considerations for CRSP data*. Social Science Research Network. https://ssrn.com/abstract=2608273

FT Alphaville. (2007, March 7). Buy? Sell? Hold? Delete! - doctored research fight rumbles on. *Financial Times*. Retrieved June 23, 2020, from https://ftalphaville.ft.com/2007/03/07/2979/buy-sell-hold-delete-doctored-research-fight-rumbles-on/

Gornall, W., & Strebulaev, I. A. (2015). *The economic impact of venture capital: Evidence from public companie*s. Social Science Research Network. https://ssrn.com/abstract=2681841

Green, L. (2003). Not so elusive-quality information on private European companies: Bureau van Dijk's response to the business information resources survey 2003. *Business Information Review*, *20*(2), 68-73.

Green, L. (2007). Deficiencies in European company information: The challenges for vendors. *Business information review, 24*(2), 89-111.

Grinblatt, M. S., Masulis, R. W., & Titman, S. (1984). The valuation effects of stock splits and stock dividends. *Journal of Financial Economics*, *13*(4), 461-490.

Guenther, D. A., & Rosman, A. J. (1994). Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics*, *18*(1), 115-128.

Gupta, K., & Krishnamurti, C. (2016). Product market competition and corporate environmental performance. In *Handbook of Environmental and Sustainable Finance* (pp. 385-404). Academic Press.

Heires, K. (2007). Capital IQ upgrading with debt and credit data. *Securities Industry News, 19*(7), 11.

Heitzman, S., & Lester, R. (2020). *Tax loss measurement*. Social Science Research Network. https://ssrn.com/abstract=3553527

Hines, T. & Sharma S. (2018, November 8). *Working with international financial data sets: potential issues and solutions* [PowerPoint Slides]. https://research.stlouisfed.org/conferences/btn2018/presentations

Hines, T. (2016, October 6). *How to deal with some commonly encountered financial data research problems* [PowerPoint Slides]. https://files.stlouisfed.org/files/htdocs/conferences/btn2016/docs/papers/hines.pdf

Hintermann, B., & Ludwig, M. (2019). *Home country bias in international emissions trading: Evidence from the EU ETS* [Paper Presentation]. Twelfth Toulouse Conference on The Economics of Energy and Climate, Atria Mercure Compans Caffarelli, Toulouse, France. https://www.tse-fr.eu/sites/default/files/TSE/documents/conf/2019/Energy_Climate2019/hintermann.pdf

Hribar, P. (2016). Commentary on: Do Compustat financial statement data articulate?. *Journal of Financial Reporting, 1*(1), 61-63.

Ince, O. S., & Porter, R. B. (2006). Individual equity return data from Thomson Datastream: Handle with care!. *Journal of Financial Research, 29*(4), 463-479.

Jacobs, H., & Müller, S. (2020). Anomalies across the globe: Once public, no longer existent?. *Journal of Financial Economics*, *135*(1), 213-230.

Jaraitė, J., Jong, T., Kažukauskas, A., Zaklan, A., & Zeitlberger, A. (2013). *Matching EU ETS Accounts to historical parent companies: A technical note*. Social Science Research Network. https://ssrn.com/abstract=2384537

Jones, H. (2020, April 21). London Stock Exchange committed to Refinitiv deal in pandemic-hit markets. *Reuters*. https://www.reuters.com/article/us-lse-results/idUSKBN2230OO

Jorion, P., & Schwarz, C. (2017). *The fix is in: Properly backing out backfill bias. Social Science Research Network*. https://ssrn.com/abstract=3010469

Kahle, K. M., & Stulz, R. M. (2013). Access to capital, investment, and the financial crisis. *Journal of Financial Economics*, *110*(2), 280-299.

Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *Journal of Financial and Quantitative Analysis*, *31*(3), 309-335.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2019). *How to construct nationally representative firm level data from the Orbis Global Database: New facts and aggregate implications (No. w21558)*. National Bureau of Economic Research. https://www.nber.org/papers/w21558.pdf

Kaplan, S. N., & Lerner, J. (2016). Venture capital data: Opportunities and challenges (No. w22500). National Bureau of Economic Research. https://www.nber.org/papers/w22500.pdf

Katz, J. (2019). Place-based manufacturing subsidies and the spatial distribution of production. *Atlantic Economic Journal, 47*(4), 521-523.

Keasler, T. R., & Denning, K. C. (2009). A re-examination of corporate strategic alliances: new market responses. *Quarterly Journal of Finance and Accounting*, *48*(1) 21-47.

Keil, J. (2017). The trouble with approximating industry concentration from Compustat. *Journal of Corporate Finance*, *45*, 467-479.

Kern, B. B., & Morris, M. H. (1994). Differences in the Compustat and expanded Value Line databases and the potential impact on empirical research. *Accounting Review*, 274-284.

Kinney, M. R., & Swanson, E. P. (1993). The accuracy and adequacy of tax data in Compustat. *The Journal of the American Taxation Association*, *15*(1), 121.

Landis, C., & Skouras, S. (2018). *A granular approach to international equity data from Thomson Datastream*. Social Science Research Network. https://ssrn.com/abstract=3225371.

Lara, J. M. G., Osma, B. G., & Noguer, B. G. D. A. (2006). Effects of database choice on international accounting research. *Abacus*, *42*(3-4), 426-454.

Lee, J. (2017). *How do firms choose their debt types?*. http://www.fmaconferences.org/Boston/P1_201608.pdf

Library of Congress. (n.d.). *LC name authority file: Thomson Financial Securities Data (Firm)*. http://id.loc.gov/authorities/names/nr00006135.html

Lins, K. V., & Servaes, H. (2002). Is corporate diversification beneficial in emerging markets?. *Financial Management*, *31*(2), 5-31.

Ljungqvist, A., Malloy, C., & Marston, F. (2009). Rewriting history. *The Journal of Finance, 64*(4), 1935-1960.

Loughran, T., & Ritter, J. R. (1995). The new issues puzzle. *The Journal of Finance*, *50*(1), 23-51.

Lu, G., & Shang, G. (2017). Impact of supply base structural complexity on financial performance: Roles of visible and not-so-visible characteristics. *Journal of Operations Management*, 53–56, 23–44. https://doi.org/10.1016/j.jom.2017.10.001

Lu, X., Stambaugh, R. F., & Yuan, Y. (2017). *Anomalies abroad: Beyond data mining* (No. w23809). National Bureau of Economic Research. https://www.nber.org/papers/w23809

Ma, J., Pagan, J. A., & Chu, Y. (2009). Abnormal returns to mergers and acquisitions in ten Asian stock markets. *International Journal of business*, *14*(3). 235-250.

Ma, Z., Pant, G., & Sheng, O. R. (2011). Mining competitor relationships from online news: A network-based approach. *Electronic Commerce Research and Applications*, *10*(4), 418-427.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160-1177.

Mathers, A., & Giacomini, E. (2016). A note on Capital IQ's credit line data. *Financial Review, 51*(3), 435-461.

McElreath Jr, R. B., & Wiggins, C. D. (1984). Using the COMPUSTAT tapes in financial research: problems and solutions. *Financial Analysts Journal*, *40*(1), 71-76.

McGuire, J. B., James, B. E., & Papadopoulos, A. (2016). Do your findings depend on your data (base)? A comparative analysis and replication study using the three most widely used databases in international business research. *Journal of International Management*, *22*(2), 186-206.

Mills, L. F., Newberry, K. J., & Novack, G. F. (2003). How well do Compustat NOL data identify firms with US tax return loss carryovers?. *Journal of the American Taxation Association, 25*(2), 1-17.

Monasterolo, I., Battiston, S., Janetos, A. C., & Zheng, Z. (2017). Vulnerable yet relevant: the two dimensions of climate-related financial disclosure. *Climatic Change*, *145*(3-4), 495-507.

Mulherin, H., & Aziz Simsir, S. (2015). Measuring deal premiums in takeovers. *Financial Management, 44*(1), 1-14.

Nam, H., No, W. G., & Lee, Y. (2017). Are commercial financial databases reliable? New evidence from Korea. *Sustainability*, *9*(8), 1406.

Netter, J., Stegemoller, M., & Wintoki, M. B. (2011). Implications of data screens on merger and acquisition analysis: A large sample study of mergers and acquisitions from 1992 to 2009. *The Review of Financial Studies*, 24(7), 2316-2357.

New Research Center. (1965). *Banking*, *58*(3), 95.

Nobes, C., & Stadler, C. (2018). Investigating international differences in financial reporting: Data problems and some proposed solutions. *The British Accounting Review*, *50*(6), 602-614.

Olbrys, J., & Majewska, E. (2014). On some empirical problems in financial databases. *Pensee Journal, 76*(9), 2-9.

Payne, J. L., & Thomas, W. B. (2003). The implications of using stock-split adjusted I/B/E/S data in empirical research. *The Accounting Review, 78*(4), 1049-1067.

Phillips, C. H. (2012). S&P Capital IQ. *Journal of Business & Finance Librarianship*, *17*(3), 279-286.

Ramnath, S., Rock, S. & Shane, P. (2005). Value Line and I/B/E/S earnings forecasts. *International Journal of Forecasting, 21*(1), 185-198.

Rauh, J. D., & Sufi, A. (2012). Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance, 16*(1), 115–155.

Refinitiv. (2019). *Datastream*. Retrieved June 2, 2020, from https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/datastream

Refinitiv. (2020). *SDC Platinum*. Retrieved June 23, 2020, from https://www.refinitiv.com/en/products/sdc-platinum-financial-securities

Ritter, J. R. (2016). *Initial public offerings: Technology stock IPOs* [Unpublished Manuscript]. Warrington College of Business. The University of Florida. Retrieved June 3, 2020, from https://site.warrington.ufl.edu/ritter/files/IPOs2019Tech-Stock.pdf

Ritter, J. R. (2019). *SDC corrections from Jay R. Ritter of the University of Florida* [Unpublished Manuscript]. Warrington College of Business. The University of Florida. Retrieved June 3, 2020, from https://site.warrington.ufl.edu/ritter/files/2019/04/SDC-corrections.pdf

Ritter, J. R. (2020). *Initial public offerings: Updated statistics* [Unpublished Manuscript]. Warrington College of Business. The University of Florida. Retrieved June 3, 2020, from https://site.warrington.ufl.edu/ritter/files/IPOs2019Statistics.pdf

Roger, T. (2017). Reporting errors in the I/B/E/S earnings forecast database: J. Doe vs. J. Doe. *Finance Research Letters, 20*, 170-176.

Rogers, F. T. (2020). Patent text similarity and cross-cultural venture-backed innovation. *Journal of Behavioral and Experimental Finance, 26*, 100319.

Röhm, P., Merz, M., & Kuckertz, A. (2019). Identifying corporate venture capital investors - a data-cleaning procedure. *Finance Research Letters, 32*, 101092.

Rosenberg, B., & Houglet, M. (1974). Error rates in CRSP and Compustat data bases and their implications. *The Journal of Finance*, *29*(4), 1303-1310.

Rossi, F. (2011). *U.K. cross-sectional equity data: do not trust the dataset! The case for robust investability filters.* MPRA Paper. https://mpra.ub.uni-muenchen.de/38303/1/

S&P Global Market Intelligence. (2017). *Compustat® Data*. Retrieved June 23, 2020, from https://www.spglobal.com/marketintelligence/en/documents/compustat-brochure_digital.pdf

San Miguel, J. G. (1977). The reliability of R&D data in Compustat and 10-K reports. *Accounting Review, 52* (3), 638-641.

Sarig, O. & Warga, A. (1989). Bond price data and bond market liquidity. *Journal of Financial and Quantitative Analysis. 24*(3), 367-378.

Schwarz, C. G., & Potter, M. E. (2016). Revisiting mutual fund portfolio disclosure. *The Review of Financial Studies*, *29*(12), 3519-3544.

Shi, L., & Zhang, H. (2011). On alternative measures of accruals. *Accounting Horizons*, *25*(4), 811-836.

Shumway, T. (1997). The delisting bias in CRSP data. *The Journal of Finance*, *52*(1), 327-340.

Shumway, T., & Warther, V. A. (1999). The delisting bias in CRSP's Nasdaq data and its implications for the size effect. *The Journal of Finance*, *54*(6), 2361-2379.

Silva, A. E. P. D. (2017). *Testing dynamic agency predictions to corporate finance*[Doctoral dissertation]. FGV Digital Repository. https://bibliotecadigital.fgv.br/dspace/handle/10438/18243

Stasch, M. (2014, September). *Vendors' methodologies for assigning industry codes* [PowerPoint Slides]. https://files.stlouisfed.org/files/htdocs/conferences/btn2014/docs/papers/stasch.pdf

Succurro, M., & Costanzo, G. D. (2019). Ownership structure and firm patenting activity in Italy. *Eurasian Economic Review, 9*(2), 239–266. https://doi.org/10.1007/s40822-018-0109-1

Suret, J. M., Morrill, C., & Morrill, J. (1998). Availability and accuracy of accounting and financial data in an emerging market: The case of Malaysia. *Asia-Pacific Journal of Accounting, 5*(1), 95-126.

Tallapally, P. (2009). *The association between data intermediaries and bond rating classification model prediction accuracy*. [Doctoral dissertation]. Louisiana Tech Digital Commons. https://digitalcommons.latech.edu/dissertations/474

Tallapally, P., Luehlfing, M. S., & Motha, M. (2011). The partnership of EDGAR online and XBRL-should Compustat care?. *Review of Business Information Systems (Rbis), 15*(4), 39-46.

Tallapally, P., M. S. Luehlfing, and M. Motha. (2012). *Data differences - XBRL versus Compustat*. [Unpublished Manuscript]. Slippery Rock University. Retrieved June 24, 2020, from http://www.aabri.com/manuscripts/11798.pdf

Thomson Reuters. (2006). *2006 factbook*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/fact-book

Thomson Reuters. (2008). *2008 factbook*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/fact-book

Thomson Reuters. (2011). *2010 Thomson Reuters Annual Report*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/annual-reports

Thomson Reuters. (2012). *2012 factbook*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/fact-book

Thomson Reuters. (2013). *Worldscope Database: Data definitions guide*. Retrieved June 2, 2020, from https://blogs.cul.columbia.edu/business/files/2014/02/Worldscope-Data-Definition-Guide-Issue-14.2.pdf

Thomson Reuters. (2016). *Worldscope Fundamentals*. Retrieved June 2, 2020, from https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/fundamentals-worldscope-fact-sheet.pdf

Thomson Reuters. (2018). *2017 Thomson Reuters Annual Report*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/annual-reports

Thomson Reuters. (2019). *2019 factbook*. Retrieved June 2, 2020, from https://ir.thomsonreuters.com/financial-information/fact-book

Tobek, O., & Hronec, M. (2018). *Does the source of fundamental data matter?*. IES Working Paper. Retrieved from https://www.econstor.eu/bitstream/10419/203194/1/1027882285.pdf

Ulbricht, N., & Weiner, C. (2005). *Worldscope meets Compustat: A comparison of financial databases*. Social Science Research Network. https://ssrn.com/abstract=871169.

Utke, S. (2018). Miscodings in Compustat's auditor variable: issues, identification, and correction. *Advances in Accounting, 43*, 56-59.

Waszczuk, A. (2014). Assembling international equity datasets - Review of studies on the cross-section of returns. *Procedia Economics and Finance, 15,* 1603-1612.

Weiß, G. N., & Mühlnickel, J. (2014). Why do some insurers become systemically relevant?. *Journal of Financial Stability, 13*, 95-117.

Williams, K. L. (2015). *The prediction of future earnings using financial statement information: Are XBRL company filings up to the task*. [Doctoral dissertation]. The University of Mississippi eGrove. https://egrove.olemiss.edu/etd/353/

Wisen, C. H. (2002). *The bias associated with new mutual fund returns*. Social Science Research Network. https://ssrn.com/abstract=90463

World Heritage Encyclopedia. (1991). *Bureau Van Dijk*. http://self.gutenberg.org/articles/eng/Bureau_van_Dijk

Yan, Y. (2007). *Research impacts and correction of the upward biased CRSP daily equal-weighted index*. Social Science Research Network. https://ssrn.com/abstract=971073

Yan, Y., Dong, J. Q., & Faems, D. (2020). Not every coopetitor is the same: The impact of technological, market and geographical overlap with coopetitors on firms' breakthrough inventions. *Long Range Planning*, 53(1), 101873.

Yang, D.C., Vasarhelyi, M.A. and Liu, C. (2003), A note on the using of accounting databases, *Industrial Management & Data Systems, 103*(3), 204-210. https://doi.org/10.1108/02635570310465689

Zhang, X., Zhang, Q., Chen, D., & Gu, J. (2019). Financial integration, investor protection and imbalanced optimistically biased information timeliness in emerging markets. *International Review of Financial Analysis*, *64*, 38-56.

Zhang, Y., & Alexander, B. (2016). Half a century of research on Value Line: a comprehensive review. *Managerial Finance*, 42(8), 799 – 816.

Zhu, Q. (2020). The missing new funds. *Management Science*, *66*(3), 1193-1204. https://doi.org/10.1287/mnsc.2019.3454

Zuckerman, G. (2004, September 9). S&P is set to acquire information provider in $200 million deal. *Wall Street Journal - Eastern Edition*, *244*(49), C5.