sustainability

MDPI

Article

Are Commercial Financial Databases Reliable? New Evidence from Korea

Hyunjung Nam [1], Won Gyun No [2,*] and Youngsu Lee [3]

[1] Graduate School of International Studies, Dong-A University, Busan 49236, Korea; hjnam@dau.ac.kr
[2] Rutgers Business School, Rutgers, The Staute University of New Jersey, Newark, NJ 07102-3122, USA
[3] College of Business, California State University, Chico, CA 95929, USA; ylee54@csuchico.edu
* Correspondence: wn55@business.rutgers.edu; Tel.: +1-973-353-5985

Received: 15 July 2017; Accepted: 4 August 2017; Published: 9 August 2017

Abstract: The quality of financial information is crucial for the effective decision-making of practitioners and academics. A number of studies have shown the existence of errors in proprietary databases provided by financial data aggregators (e.g., Compustat and Value Line) in advanced markets like the U.S. However, no study has examined the quality of the financial data offered by aggregators in emerging markets. Research on such markets is needed as financial investment frequently occurs in emerging markets due to the globalization of capital. The purpose of this study is to fill this gap by investigating whether financial data provided by aggregators is the same as the data reported in firms' financial statements in emerging markets. Another purpose of this study is to examine the impact on academic research. Comparing the 18 most widely-used financial items found in the original filings of firms with the corresponding data provided by all three data aggregators currently available in South Korea (i.e., DataGuide, KisValue, and TS2000), we found a considerable number of differences; many of the differences are substantially greater than conventional materiality. We also found that the differences between data sources lead to different prediction results in bankruptcy prediction model.

Keywords: data quality; absolute difference; material difference; financial information aggregator; Ohlson's bankruptcy prediction model

1. Introduction

> *"Almost all businesses, government organization, hospital, educational institutions, and individuals have been hurt by data quality problems."* [1]

Credible financial information is the essential ingredient for the evaluation of business performance and provides a reliable basis for making business decisions [2–5]. While accurate financial information results in appropriate decision-making, inaccurate financial information distorts the financial performance of companies and eventually leads to incorrect decisions [6]. For instance, information quality problems cost U.S. businesses more than $600 billion annually [1]. Koziol [7] demonstrated that investment advice made by a bank based on inaccurate information led to investors' financial losses.

At present, the various users of financial information—including investors, financial institutions, creditors, policy makers, and researchers—generally obtain financial data from two popular sources: public data repositories and financial data aggregators (henceforth, "aggregators"). Around the world, many regulators and government agencies provide users with publicly accessible electronic data repositories to enhance the efficiency of capital markets by making financial information more accessible, timelier, and less costly. For instance, the U.S. Securities and Exchange Commission (SEC) has implemented and maintained the Electronic Data Gathering Analysis and Retrieval (EDGAR) system from which the public can obtain all documents filed with the SEC, such as quarterly reports

and annual reports. Another popular source of financial data is proprietary databases provided by financial data aggregators, such as Standard and Poors, Bloomberg, Thomson Reuters, and MSN Money in the U.S. These aggregators directly or indirectly extract data from companies' financial statements and provide it to users for a fee or free of charge to entice them to use their services. Since aggregators offer easily accessible data in a standardized format as well as value-added information that might not be found in the original data, many users, including researchers, prefer aggregators' databases to public data repositories as their main source of financial data.

The accuracy of aggregators' data, however, has been questioned by several researchers. A number of previous studies have shown the existence of errors in aggregators' data, despite aggregators being known for offering reliable and value-added financial data [8–12]. Most of these studies have focused on developed markets, particularly the U.S. market, in which data offered by aggregators is readily used, and in which a strict validation process is believed to be present. However, no study has examined the quality of financial data offered by aggregators in emerging markets. Research on such markets is needed as financial investment frequently occurs in emerging markets due to the globalization of capital. The purpose of this study is to fill this gap by investigating whether financial data provided by aggregators is the same as the data reported in firms' financial statements in emerging markets, especially in South Korea. We chose South Korea because it is one of the fastest-growing markets in the world. Globally renowned financial institutions include the Korean stock index as one of the fastest-growing market indices, and the number of international investors is increasing [13]. Furthermore, as one of the world's most digitally advanced countries, financial data services (i.e., aggregators) for business are also growing in South Korea [14].

In this study, we examined 1290 filings of 645 publicly held firms in South Korea for the two fiscal years of 2011 and 2012. In particular, we selected 18 commonly-used financial items and investigated the similarities and differences between the data (i.e., values) reported in firms' filings (i.e., financial statements) and the corresponding data found in all three data aggregators' databases currently available in South Korea: DataGuide, KisValue, and TS2000. The results indicate 3.5% to 25.4% differences between the data reported in firms' financial statements and the corresponding data provided by the three aggregators. We also examined Ohlson's [15] bankruptcy prediction model to address the potential effects of such differences on financial analysis and academic research. The results reveal that, depending on the sources of financial data, the bankruptcy prediction model presents different predictions of bankruptcy.

This study provides several contributions not only for aggregators but also for users who frequently rely on financial data offered by aggregators. First, it is often argued that the less accurate financial data of emerging markets incurs a so-called "discount" in the stock market [16–18]. In their analysis of the Korean stock market, Jeong, Kwak, and Hwang [19] estimated that the stock market value would increase by 35 billion dollars as of 2008, and interest costs for companies would decrease by 13 billion dollars, if financial reporting and financial data were more accurate and reliable. By assisting aggregators to identify where their data acquisition processes lead to differences between their data and the data reported by the firms themselves, the results of this study contribute to enhancing the data quality of aggregators and, eventually, to improving market valuation. In addition, users of the data provided by aggregators can benefit from the findings of this study to better understand the nature and extent of differences between firms' reported data and the data provided by aggregators.

Second, Asian stock markets, such as in China, South Korea, and Malaysia, are gaining attention as significant markets, and the number of academic research studies on emerging markets is increasing [20,21]. Although most of this research is based on data provided by aggregators, no attempt has been made to investigate differences between firms' reported data and aggregators' data in emerging markets. This study extends existing research by examining emerging markets.

Finally, quick and convenient data retrieval is one of the major drivers of data offered by aggregators; however, if differences exist between the data reported in firms' financial statements and the corresponding aggregators' data, then such differences might lead to different results depending on

the source of the data used and, thus, eventually might distort decision-making. Therefore, researchers who rely on aggregators as their data sources can benefit from this research by becoming better informed about differences in data that might affect their analysis.

The remainder of this paper is organized as follows. The next section reviews the literature on data quality and introduces research questions. This is followed by the details of the research methodology, including a description of the sample, data collection procedures, and research instruments. Next, the results of the study are presented. Finally, this paper concludes with a summary of the findings, implications, and limitations of the study.

2. Literature Review and Research Questions

It seems obvious that the same data should represent the same fact, regardless of the source of the data; that is, for the same financial fact, no difference should exist between the data reported in financial statements and the corresponding aggregators' data or between data found in different aggregators' databases; however, several prior studies have identified data differences among data sources. This section provides a summary of such studies.

Early studies on aggregators' data focus on the existence of differences. For example, Rosenberg and Houglet [10] made an initial effort to investigate aggregators' data. They examined differences between Compustat and the Center for Research in Security Price (CRSP) by comparing 35,357 monthly price relatives of 844 industrial firms between 1963 and 1968, and 5939 monthly price relatives of 97 utility firms between 1962 and 1968. They found a total of 1202 differences (2.19%): 1060 differences (2.99%) for industrial firms and 142 differences (2.39%) for utility firms. Bennin [8] updated Rosenberg and Houglet [10] by investigating more firms and for a longer period: 170,084 monthly returns of 1295 industrial firms and 17,376 monthly returns of 99 utility firms from 1962 to 1978. They found 934 errors (0.54%) for the industrial firms and eight errors (0.04%) for the utility firms, suggesting that the number of differences had decreased since Rosenberg and Houglet [10].

Subsequent studies have identified several reasons for differences in financial data provided by aggregators. Based on our reviews of existing literature, we identified four major underlying causes: (1) different coding policies across aggregators, (2) different currency units, (3) insufficient information about filing practices of raw data, and (4) unexplainable coding errors. The first two causes are due to systemic errors, whereas the other two causes come from random errors such as insufficient explanation in the original data and other errors that cannot be identified. In the following subsections, we discuss these two types of data differences: systematic difference and random difference.

2.1. Difference in Coding Policies

The first systemic difference occurs when aggregators use different coding schemes that are variant from what firms use to report their financial items and treat omitted and non-existing items as missing values. Financial data aggregators often adjust or reclassify original financial items in many different ways, not only to standardize financial items across firms for comparison purposes but also to provide the value-added information they claim. However, their adjustment procedure sometimes may confuse users and lead to the controversial classification of financial items. For example, Kinney and Swanson [9] compared 19 tax-related items retrieved from Compustat with those found in firms' financial statements from 1986 to 1988. Out of the total 4978 observations from 100 companies, the difference rates ranged from 0.76% to 11.65%. Items from footnotes showed high difference rates while items from balance sheets showed low difference rates. They found that the high difference rates were associated with the complex procedure that Compustat used to transform tax-related items and, thus, caused the differences. When the tax items in the firms' financial statements were entered into Compustat, some data were likely to be miscoded. Courtenay and Keller [22] compared distributions (i.e., stock dividends and stock splits) between CRSP and Moody's Dividend Record (MDR) for 1989. They found 142 differences (20%) out of 718 distributions, and 64% of the differences (i.e., 91 differences) occurred mainly due to CRSP's unspecified coding policy. More recently, by

comparing the financial data of 1479 firms from Compustat and Value Line between 1976 and 1981, Yang et al. [12] also showed that the major cause of differences was the differences in undisclosed coding rules. They found a significant number of differences in seven financial facts between two data sources: assets, net sales, inventory, net income before extraordinary items, current liabilities, depreciation, depletion and amortization, and gross plant. Out of the total of 10,353 comparisons, they discovered 1284 differences (12.5%) in which the magnitude of differences was larger than 1%; 520 differences (5.02%) were caused by missing values.

2.2. Use of Different Currency Units

Multinational firms often produce multiple financial reports in various languages to meet international investors' needs. While doing so, each firm might have different reporting practices for converting the original currency to foreign currencies. Kern and Morris [23] examined *Sales* and *Total Assets* retrieved from Compustat and Value Line. By comparing data found in annual reports with those from Compustat and Value Line, they identified 123 material differences in *Total Assets* and 378 material differences in *Sales* between 1985 and 1990. Kern and Morris [23] defined material difference as the degree to which difference of an item between two data sources is five percent or greater. Among those material differences, 56 discrepancies in *Total Assets* and 57 discrepancies in *Sales* were due to the difference in the monetary units used. Value Line used the original foreign currencies, whereas Compustat converted the foreign currencies into U.S. dollars. In addition, Yang et al. [12] found the difference between the monetary units used by Compustat and Value Line. Similar to Kern and Morris [23], Compustat consistently used U.S. dollars, while Value Line used foreign currencies for foreign companies.

2.3. Insufficient Information about Filing Practice for Raw Data

Despite detailed guidelines by accounting regulation bodies, filing companies do not provide sufficient information about financial items. Noticing this, San Miguel [24] made an initial attempt to investigate the causes of differences observed in aggregators' data. By comparing the R&D expenses reported in the 10-Ks of a sample of 256 firms in 1972 with those retrieved from Compustat, they found 78 differences (30%) and argued that insufficient disclosure of 10-K filing practices resulted in 26 of the 78 differences. In particular, several R&D expenses were reported in different sections of 10-Ks, but Compustat did not present the exact R&D expenses by failing to reclassify them.

2.4. Unexplainable Coding Errors

Finally, some studies have demonstrated errors that are not identifiable. Tallapally et al. [11] compared the Cost of Goods Sold (COGS) of 26 firms retrieved from Compustat and EDGAR Online in 2009. They found that, among 26 firms, the COGS amounts of 23 companies were reported lower in Compustat than in EDGAR Online, whereas two companies showed the opposite values; however, they failed to identify the fundamental causes of the differences. Similarly, Kinney and Swanson [9] found differences in tax items between Compustat and firms' financial statements, but failed to identify the causes of the differences. Yang et al. [11] also found unknown errors in several items (e.g., net sales, inventory, and gross plant) and argued that these errors were unexplained coding errors.

2.5. Research Questions

As noted previously, no study has examined the differences between firms' reported data and the data offered by aggregators in emerging markets, while a substantial number of studies have examined developed markets, particularly in the U.S. As long as the original financial data are processed, even to a small degree, differences in values between the original data and aggregators' data are unavoidable. Therefore, given the growing importance of emerging markets in the global economy, it is essential to examine whether differences exist between the data reported in firms' financial statements and the corresponding data provided by aggregators and to compare the results with findings in developed

markets. In particular, this study investigates the differences between firms' reported data and the data provided by aggregators in South Korea.

Similar to the U.S., publicly-held companies in South Korea are required to report their financial statements to a government agency, the Financial Supervisory Service (FSS), by uploading their statements onto a public data repository—the Data Analysis, Retrieval, and Transfer (DART) system. DART is freely available to the public website provided by FSS [25] and, therefore, the public can access all financial statements reported by firms. Since publicly held firms are required to upload their audited financial statements directly onto DART, it is also believed that DART provides the most accurate data which reflect the firms' intended communications in their official financial reports. On the other hand, DART is inconvenient for accessing multiple financial statements for multiple firms at the same time because it does not provide financial data in a standardized format. This makes data retrieval from DART inconvenient. Consequently, other sources for financial data exist to facilitate data retrieval; that is, several aggregators provide financial data in various standardized formats to be downloaded into desktop applications, such as Microsoft Excel, and deliver value-added information (e.g., accounting ratios) that does not appear in the original financial statements. Many users, including academic researchers and investors, therefore prefer to use aggregator-provided data to DART data.

One would expect that no differences exist between the financial data in DART and aggregator-provided data, since aggregators in South Korea usually retrieve firms' financial data from DART and transfer the data into their repositories. Nevertheless, the potential for differences exists due to the different data-gathering, processing, and storing practices adopted by aggregators. For instance, some aggregators are known to manually transfer financial data from DART to their databases whereas others use subcontractors that employ either manual or automatic methods for data retrieval from DART and transfer financial data to the aggregators. Furthermore, some aggregators merge or reclassify financial items to incorporate firms' data into their data format. Considering manual or semi-automated data entry practices may cause errors in data (e.g., data entry errors), there might be differences between DART data and aggregators' data, and such differences, if they exist, might lead to unintended consequences in financial analysis and academic research. Therefore, this study introduces the following research questions:

RQ1: Do differences exist between financial data available in DART and data provided by aggregators in South Korea?
RQ2: Are any differences found material?
RQ3: Can the differences have an impact on academic research?

3. Research Method

3.1. Sample

We used the 764 publicly listed firms in DART as our initial sample. Out of the 764, we excluded 94 financial firms—such as banks, security, and insurance firms—because of their substantially different reporting practices. For example, interests from bank deposits for non-financial companies fall into non-operating income, while interests from deposits for financial companies fall into revenue due to the nature of their business. In addition, we dropped 19 firms whose fiscal year did not end on 31 December. We also excluded two firms that used different currencies (e.g., Hong Kong dollar) and two firms that had serious flaws in the use of units in their financial statements. Finally, we excluded two firms that did not provide their financial statements in DART. This procedure led to 645 firms as our final sample. In particular, we examined the official filings of 645 firms for two years (i.e., 2011 and 2012), in which international financial reporting standards (IFRS) were mandated by the Korean government. As a result, a total of 1290 filings of 645 firms were investigated in this study.

For financial items, we selected 18 financial items that are extensively used by investors and academic researchers to evaluate firms' financial statuses [26–29]. Among the 18 financial items, 12 items were from balance sheet: *Current Assets, Cash and Cash Equivalents, Inventories, Non-current*

Assets, *Tangible Assets*, *Intangible Assets*, *Total Assets*, *Current Liabilities*, *Non-current Liabilities*, *Total Liabilities*, *Retained Earnings*, and *Total Equity*. The remaining six items were from income statements: *Sales*, *Operating Profit*, *Earnings Before Interest and Taxes (EBIT)*, *Income Tax Expense*, *Net Income*, and *Earnings Per Stock (EPS)*. Finally, we collected the corresponding 18 financial items from three prominent data aggregators. We chose DataGuide, KisValue, and TS2000, since these aggregators are the only available commercial databases used by institutional and individual investors as well as academic researchers in South Korea. FnGuide owns DataGuide [30], NICE Information Service provides KisValue [31] and, finally, the Korean Listed Companies Association offers TS2000 [32]. This procedure led to a total of 23,220 financial facts (645 firms × 18 financial items × 2 years); however, some firms did not provide all 18 financial items in DART. As a consequence, the final sample was reduced to a total of 22,717 financial facts for each data source. Furthermore, we used the difference in a financial item for three comparisons (DART–DataGuide, DART–KisValue, and DART–TS2000) as the unit of analysis, since the purpose of this study is to examine whether there is a difference between the original data from DART and the data from three aggregators. This leads to a total number of data points of 68,151 (22,717 financial facts × 3 comparisons). Panel A of Table 1 summarizes our final sample.

Table 1. Sample.

Panel A. Summary of sample		
Filing	Number of Firms	645
	Sample Period (2 years)	2011–2012
	Total	1290
Financial Item	Balance Sheet Item	12
	Income Statement Item	6
	Total	18
Total Number of Financial Facts		23,220
Exclude: Financial facts that are not available		(503)
Total Number of Financial Facts Examined †		22,717
Three Comparisons ‡		×3
Total Number of Data Points		68,151
Panel B. Distribution of sample by industry		
Industry		***N*** [a] **(%)**
Fishery		5 (0.8%)
Raw Material Mining		1 (0.2%)
Manufacturing		430 (66.7%)
Utilities		11 (1.7%)
Construction		33 (5.1%)
Retailing and Wholesaling		51 (7.9%)
Transportation (Ground) and Logistics		10 (1.6%)
Transportation (Air and Marine)		11 (1.7%)
Publishing		5 (0.8%)
Broadcasting, Telecommunication, and Information Systems		11 (1.7%)
Real Estate and Car Rental		2 (0.3%)
Services		75 (11.6%)

† The total of 22,717 financial facts was examined for each data source (i.e., Data Analysis, Retrieval and Transfer (DART), DataGuide, KisValue, and TS2000); ‡ DART vs. DataGuide, DART vs. KisValue, and DART vs. TS2000; [a] Number of firms.

A summary of the sample distribution based upon industry classification by the Korean Industry Classification code shows that our sample firms cover 12 industry groups (see Panel B of Table 1) (Since DART adopts the industry classification scheme of the Commission of Statistics of South Korea, the industry classification in this study might not exactly match that of the Korean Stock Exchange). Among them, *Manufacturing* (66.7%) represented the highest number of firms, followed by *Services* (11.6%), *Retailing and Wholesaling* (7.9%), *Construction* (5.1%), *Utilities* (1.7%), *Transportation—Air and*

Marine (1.7%), *Broadcasting, Telecommunication and Information Systems* (1.7%), *Transportation (Ground) and Logistics* (1.6%), *Fishery* (0.8%), *Publishing* (0.8%), *Real Estate and Car Rental* (0.3%), and *Raw Material Mining* (0.2%).

3.2. Data Collection Procedure and Measures

To investigate our three research questions, we compared the 18 financial items collected from DART with those retrieved from DataGuide, KisValue, and TS2000. The comparison was conducted in three steps. In the first step, we gathered the 18 financial items from the financial statements of the 645 firms. Although DART allows anyone to access the financial data of publicly held firms online, it does not provide a feature that allows users to download financial statements of multiple firms at the same time; hence, we developed a software agent (i.e., a computer program) to retrieve the 18 financial items of the 645 sample firms from DART automatically. To validate the software agent, we then manually compared the facts of all firms gathered by the software agent to check whether the software agent performed accurately. All of the collected facts exactly matched with the data manually retrieved from DART. Next, we extracted the corresponding 18 financial items of the 645 firms from DataGuide, KisValue, and TS2000. Unlike DART, the three aggregators allow for downloading multiple financial items from multiple firms.

In the second step, we merged the data gathered from both DART and the three aggregators using a company identifier and fiscal year. We used the designated serial number of each company employed by the Korean Stock Exchange as the company identifier.

In the third step, to make units comparable and minimize any effect from different rounding practices across all databases, units were adjusted by using DART data as our bases. More specifically, DART uses several different base units: one (62,868 cases), one thousand (1803 cases), and one million (3480 cases). On the other hand, DataGuide uses a base unit as one whereas KisValue and TS2000 use a base unit as one thousand. We therefore adjusted all units used by the four aggregators in our comparisons. In particular, to address the effect of rounding to different decimals, we rounded the original amounts in DART to the nearest ten thousand, except for EPS, and compared them with the amounts in other aggregators. For the amounts in millions in DART, we rounded the amounts in three aggregators to the nearest ten million and compared them with the amounts in DART. Furthermore, we did not round the EPS amounts for our comparisons for DART and all three aggregators.

In the fourth step, to examine whether data from the three data aggregators matched those from DART, we calculated the absolute value difference between DART and each aggregator for the 18 financial items for each firm. A variable called "*ComparisonResults*" was created to code the comparison results. If the absolute value difference was equal to zero, then the variable was coded as 0, indicating no difference of data between DART and an aggregator. In contrast, if the absolute value difference was not equal to zero, then the variable was coded as 1, signifying a difference between DART and an aggregator. If a financial item was only available in DART but not in an aggregator, then the variable was coded as 2, denoting missing data.

In addition, another variable called "*MaterialResults*" was created to explore whether the differences were material (i.e., considerably large or not). Based on Leslie [33] and Eilifsen and Messier [34], we define "material differences" as the degree to which absolute value differences are larger than 0.5% of *Total Assets* for balance sheet items and 5% of *Earnings Before Interest and Taxes (EBIT)* for income statement items except EPS; therefore, if the absolute value difference was larger than 0.5% of *Total Assets* or 5% of *EBIT*, then the variable was coded as 1; otherwise, it was coded as 0. For *Earnings Per Share (EPS)*, we applied a materiality of 50 Korean Won (KRW); that is, if the material difference for EPS was greater than 50 KRW, then the variable was coded as 1; otherwise, it was coded as 0. Finally, additional variables called "*DifferenceType*" was created to code the types of differences between DART data and aggregators' data. As we discussed earlier, if a difference is due to different coding policies across aggregators and different currency units, we counted these as systemic difference and coded them as 1. On the other hand, if a difference is owing to random errors such as

insufficient explanation in the original data and other errors that cannot be identified, we deemed these as a systemic difference and coded them as 0.

4. Results

4.1. Descriptive Statistics

Table 2 shows the descriptive statistics of the 18 financial items of the 645 firms for 2011 and 2012. Currency units used for all balance sheet and income statement items are in hundred million KRW and the unit for EPS is KRW. It is noteworthy that there are differences among the financial items across the four data sources; for instance, the mean value of *Total Assets* in DART is 17,414 in 2011, while the values in DataGuide, KisValue, and TS2000 for the same year are 17,415, 16,761, and 20,932, respectively. Likewise, the mean value of *Sales* in DART is 26,552 in 2011, whereas the values in DataGuide, KisValue, and TS2000 are 26,553, 26,016, and 32,574, respectively. The results from the descriptive statistics suggest that differences exist between the data reported in DART and the corresponding data provided by the three aggregators.

Table 2. Descriptive statistics of 18 financial items across data sources.

Panel (A), DART.										
Financial Item [a]	**2011**					**2012**				
	N [b]	**M** [c]	**SE** [d]	**Min** [e]	**Max** [f]	**N**	**M**	**SE**	**Min**	**Max**
Current Asset	642	28,948	113,557	101	1,558,003	643	30,296	123,456	85	1,810,716
Cash and Cash Equivalents	642	11,551	44,973	25	715,021	642	11,724	49,310	42	872,690
Inventories	644	11,045	56,698	1	1,123,849	644	11,743	61,756	0	1,223,761
Non−Current Asset	645	1824	8180	0	146,918	645	1764	9018	0	187,915
Tangible Assets	635	2749	10,897	0	157,167	636	2723	10,915	0	177,474
Intangible Assets	556	1162	5467	0	86,106	557	1224	5559	0	85,229
Total Assets	641	17,414	74,370	3	1,227,003	641	18,550	80,852	2	1,322,193
Current Liabilities	645	16,410	62,047	24	826,639	645	16,903	66,155	6	950,886
Non−current Liabilities	629	9910	34,015	11	443,190	629	9638	33,404	4	469,331
Total Liabilities	641	6625	34,301	0	649,227	641	7415	39,163	0	762,715
Retained Earnings	645	12,469	55,334	−657	1,013,136	645	13,295	62,575	−6760	1,214,802
Total Equity	638	8442	47,833	−41,823	976,229	638	8776	57,454	−290,293	1,199,857
Sales	618	26,552	106,724	42	1,650,018	618	28,652	119,907	59	2,011,036
Operating Profit	639	1600	8491	−7635	156,443	639	1307	14,094	−163,934	290,493
EBIT	627	351	2224	−10,681	34,329	629	326	2864	−9854	60,697
Income Tax Expense	641	1427	9773	−45,601	171,919	641	1033	16,637	−248,710	299,150
Net Profit	630	1160	7850	−45,601	137,590	630	769	14,614	−248,710	238,453
EPS	598	1791	31,785	−386,966	350,909	598	2694	14,996	−164,485	161,280
Panel (B), DataGuide.										
Financial Item [a]	**2011**					**2012**				
	N [b]	**M** [c]	**SE** [d]	**Min** [e]	**Max** [f]	**N**	**M**	**SE**	**Min**	**Max**
Current Asset	642	28,949	113,557	101	1,558,003	643	30,296	123,456	85	1,810,716
Cash and Cash Equivalents	642	11,568	44,981	25	715,021	642	11,764	49,321	42	872,690
Inventories	644	11,045	56,699	1	1,123,849	644	11,743	61,756	0	1,223,761
Non−Current Asset	645	1825	8181	0	146,918	645	1765	9019	0	187,915
Tangible Assets	633	2764	10,918	0	157,167	633	2715	10,932	0	177,474
Intangible Assets	554	1241	5791	0	86,106	555	1328	6032	0	85,229
Total Assets	641	17,415	74,373	3	1,227,003	641	18,550	80,853	2	1,322,193
Current Liabilities	645	16,410	62,047	24	826,639	645	16,903	66,155	6	950,886
Non−current Liabilities	629	9920	34,029	22	443,190	629	9660	33,420	4	469,331
Total Liabilities	641	6629	34,302	0	649,227	640	7429	39,193	0	762,715
Retained Earnings	645	12,476	55,333	−657	1,013,136	645	13,301	62,573	−6760	1,214,802
Total Equity	638	8594	47,803	−10,230	976,229	638	9327	56,229	−13,459	1,199,857
Sales	618	26,553	106,724	42	1,650,018	618	28,652	119,907	59	2,011,036
Operating Profit	639	1608	8492	−7635	156,443	639	1561	12,480	−6987	290,493
EBIT	615	450	2229	−2931	34,329	618	390	2881	−9854	60,697
Income Tax Expense	641	1504	9595	−24,731	171,919	641	1421	13,386	−40,633	299,150
Net Profit	630	1233	7625	−11,548	137,590	630	1163	10,698	−15,882	238,453
EPS	261	1201	23,904	−334,933	134,856	259	2002	16,833	−164,485	161,280

Table 2. *Cont.*

Panel (C) KisValue.										
Financial Item [a]	**2011**					**2012**				
	N [b]	**M** [c]	**SE** [d]	**Min** [e]	**Max** [f]	**N**	**M**	**SE**	**Min**	**Max**
Current Asset	642	28,447	113,671	0	1,558,003	643	29,699	123,587	0	1,810,716
Cash and Cash Equivalents	642	11,324	45,048	0	715,021	642	11,478	49,368	0	872,690
Inventories	644	10,903	56,858	0	1,124,308	644	11,563	61,946	0	1,224,132
Non−Current Asset	645	1791	8186	0	146,918	645	1729	9023	0	187,915
Tangible Assets	635	2694	10,914	0	157,167	636	2640	10,919	0	177,474
Intangible Assets	556	1213	5758	0	86,106	557	1303	5999	0	85,229
Total Assets	641	16,761	71,978	0	1,223,419	641	17,784	78,204	0	1,322,193
Current Liabilities	645	16,176	62,102	0	826,639	645	16,632	66,216	0	950,886
Non−current Liabilities	629	9734	34,061	0	443,190	629	9466	33,452	0	469,331
Total Liabilities	641	6573	34,310	0	649,227	641	7278	39,129	0	762,715
Retained Earnings	645	12,209	55,384	−657	1,013,136	645	12,977	62,631	−6760	1,214,802
Total Equity	638	8431	47,827	−10,230	976,229	638	9050	56,216	−13,459	1,199,857
Sales	618	26,016	106,838	0	1,650,018	618	28,045	120,034	0	2,011,036
Operating Profit	639	1574	8493	−7635	156,443	639	1534	12,479	−6987	290,493
EBIT	627	433	2209	−2931	34,329	629	376	2857	−9854	60,697
Income Tax Expense	641	1477	9594	−24,731	171,919	641	1402	13,386	−40,633	299,150
Net Profit	630	1113	7156	−11,562	133,826	630	1039	10,268	−12,859	231,854
EPS	483	2748	30,874	−386,966	350,909	473	2838	16,499	−164,485	161,280
Panel (D) TS2000.										
Financial Item [a]	**2011**					**2012**				
	N [b]	**M** [c]	**SE** [d]	**Min** [e]	**Max** [f]	**N**	**M**	**SE**	**Min**	**Max**
Current Asset	515	35,401	125,926	179	1,556,313	512	37,298	137,496	125	1,810,716
Cash and Cash Equivalents	516	13,634	46,890	61	715,021	512	13,976	51,609	42	872,690
Inventories	518	13,487	62,981	1	1,123,849	514	14,435	68,874	4	1,223,761
Non−Current Asset	518	2228	9084	2	146,918	514	2171	10,064	0	187,915
Tangible Assets	511	3343	12,079	0	157,167	507	3312	12,141	0	177,474
Intangible Assets	452	948	3813	0	45,919	450	1015	3929	0	44,221
Total Assets	515	20,932	80,269	11	1,223,419	511	22,458	87,763	60	1,322,193
Current Liabilities	518	20,094	68,633	29	826,639	514	20,871	73,590	10	950,886
Non−current Liabilities	507	12,014	37,569	22	443,190	503	11,802	37,026	4	469,331
Total Liabilities	516	8138	38,045	0	649,227	512	9171	43,645	0	762,715
Retained Earnings	518	15,191	61,616	−657	1,018,453	514	16,285	69,788	−6760	1,214,802
Total Equity	505	10,552	53,546	−7521	976,229	383	8077	31,352	−13,533	403,465
Sales	495	32,574	118,684	60	1,650,018	492	35,227	133,611	61	2,011,036
Operating Profit	513	2	9	−13	155	509	2	14	−8	290
EBIT	506	541	2444	−2931	34,249	505	468	3183	−9854	60,697
Income Tax Expense	514	1858	10,666	−24,731	171,590	510	1762	14,989	−40,634	299,150
Net Profit	503	1527	8498	−11,548	137,341	499	1388	11,922	−15,882	238,453
EPS	476	773	7931	−33,291	134,856	473	527	11,978	−164,485	161,280

[a] Unit: All balance sheet and income statement items (in hundred million of Korean Won) and EPS (in Korean Won); [b] Number of financial facts; [c] Mean; [d] Standard deviation; [e] Minimum; [f] Maximum.

4.2. Comparisons between DART and the Three Aggregators

This section summarizes the comparison results between DART and the three aggregators. Table 3 provides a summary of the comparison results of the 18 financial items.

The overall results in the last row of Table 3 indicate that DataGuide (93.3%) had the highest number of matches, followed by KisValue, (73.6%) and TS2000 (68.4%). However, the results also show a considerable number of mismatches: DataGuide (3.5%), KisValue (25.4%) and TS2000 (11.1%). Specifically, the comparison results in the "DART vs. DataGuide" column show that, out of the total of 22,717 financial facts, 804 facts (3.5%) in DataGuide did not match with the corresponding facts in DART. The number of mismatches in the 18 financial items ranged from 12 (1%) to 96 (7.4%) facts. *Current Assets* (1202), *Cash and Cash Equivalents* (1194), and *Retained Earnings* (1196) were the top three mismatched financial items. In addition, 709 facts (3.1%) across five financial items were not available in DataGuide, while corresponding facts were available in DART. Among them, *EPS* (676) had the highest number of missing values followed by *Income Tax Expense* (23), *Inventories* (5), *Intangible Assets* (4), and *Non-Current Liabilities* (1).

Table 3. Comparison of 18 financial facts between DART and three aggregators.

Financial Item	N [a]	DART vs. DataGuide						DART vs. KisValue						DART vs. TS2000					
		Match [b]		Mismatch [c]		Missing [d]		Match		Mismatch		Missing		Match		Mismatch		Missing	
Current Asset	1284	1202	(93.6%)	82	(6.4%)	0	(0.0%)	961	(74.8%)	323	(25.2%)	0	(0.0%)	929	(72.4%)	99	(7.7%)	256	(19.9%)
Cash and Cash Equivalents	1290	1194	(92.6%)	96	(7.4%)	0	(0.0%)	997	(77.3%)	294	(22.8%)	0	(0.0%)	935	(72.5%)	97	(7.5%)	258	(20.0%)
Inventories	1271	1237	(97.3%)	29	(2.3%)	5	(0.4%)	1000	(78.7%)	271	(21.3%)	0	(0.0%)	976	(76.8%)	42	(3.3%)	253	(19.9%)
Non-Current Asset	1282	1233	(96.2%)	49	(3.8%)	0	(0.0%)	972	(75.8%)	310	(24.2%)	0	(0.0%)	946	(73.8%)	80	(6.2%)	256	(20.0%)
Tangible Assets	1288	1256	(97.5%)	32	(2.5%)	0	(0.0%)	931	(72.3%)	357	(27.7%)	0	(0.0%)	988	(76.7%)	44	(3.4%)	256	(19.9%)
Intangible Assets	1113	1058	(95.1%)	51	(4.6%)	4	(0.4%)	813	(73.0%)	300	(27.0%)	0	(0.0%)	433	(38.9%)	469	(42.1%)	211	(19.0%)
Total Assets	1285	1256	(97.7%)	29	(2.3%)	0	(0.0%)	1004	(78.1%)	281	(21.9%)	0	(0.0%)	981	(76.3%)	46	(3.6%)	258	(20.1%)
Current Liabilities	1258	1222	(97.1%)	36	(2.9%)	0	(0.0%)	980	(77.9%)	278	(22.1%)	0	(0.0%)	955	(75.9%)	55	(4.4%)	248	(19.7%)
Non-current Liabilities	1282	1251	(97.6%)	30	(2.3%)	1	(0.1%)	996	(77.7%)	286	(22.3%)	0	(0.0%)	969	(75.6%)	59	(4.6%)	254	(19.8%)
Total Liabilities	1290	1261	(97.8%)	29	(2.2%)	0	(0.0%)	1009	(78.2%)	281	(21.8%)	0	(0.0%)	982	(76.1%)	50	(3.9%)	258	(20.0%)
Retained Earnings	1276	1196	(93.7%)	80	(6.3%)	0	(0.0%)	959	(75.2%)	317	(24.8%)	0	(0.0%)	723	(56.7%)	165	(12.9%)	388	(30.4%)
Total Equity	1290	1257	(97.4%)	33	(2.6%)	0	(0.0%)	1010	(78.3%)	280	(21.7%)	0	(0.0%)	981	(76.0%)	51	(4.0%)	258	(20.0%)
Sales	1236	1202	(97.2%)	34	(2.8%)	0	(0.0%)	968	(78.3%)	268	(21.7%)	0	(0.0%)	899	(72.7%)	88	(7.1%)	249	(20.1%)
Operating Profit	1278	1210	(94.7%)	68	(5.3%)	0	(0.0%)	979	(76.6%)	299	(23.4%)	0	(0.0%)	751	(58.8%)	271	(21.2%)	256	(20.0%)
EBIT	1282	1233	(96.2%)	49	(3.8%)	0	(0.0%)	1000	(78.0%)	282	(22.0%)	0	(0.0%)	945	(73.7%)	79	(6.2%)	258	(20.1%)
Income Tax Expense	1256	1200	(95.5%)	33	(2.6%)	23	(1.8%)	999	(79.5%)	257	(20.5%)	0	(0.0%)	943	(75.1%)	68	(5.4%)	245	(19.5%)
Net Profit	1260	1228	(97.5%)	32	(2.5%)	0	(0.0%)	244	(19.4%)	1016	(80.6%)	0	(0.0%)	953	(75.6%)	49	(3.9%)	258	(20.5%)
EPS	1196	508	(42.5%)	12	(1.0%)	676	(56.5%)	890	(74.4%)	66	(5.5%)	240	(20.1%)	240	(20.1%)	709	(59.3%)	247	(20.7%)
Overall	22,717	21,204	(93.3%)	804	(3.5%)	709	(3.1%)	16,712	(73.6%)	5766	(25.4%)	240	(1.1%)	15,529	(68.4%)	2521	(11.1%)	4667	(20.5%)

[a] Number of financial facts; [b] Number of matches (match is coded as 0 if the absolute value difference between DART and each aggregator is equal to zero); [c] Number of mismatches (mismatch is coded as 1 if the absolute value difference between DART and each aggregator is not equal to zero); [d] Number of missing values (missing value is coded as 2 if the financial item is not available from aggregator). Differences includes mismatches and missing values in the current study.

Concerning KisValue, the overall comparison results indicate that 5766 facts (25.4%) in KisValue did not match with the corresponding facts in DART, with a range of 66 (5.5%) to 1016 (80.6%) mismatches across the 18 financial items. The top three mismatched financial items were *Current Assets* (323), *Tangible Assets* (357), and *Net Profit* (1016). Furthermore, 240 facts (1.1%) for one financial item (i.e., EPS) were not available in KisValue. It should be noted that KisValue is known to enter "0" for missing data, which makes it very difficult to determine whether a financial fact with a "0" is missing or the actual value of the fact is, indeed, zero. Given that financial items examined in this study rarely have zero values, we speculate that the majority of the mismatches in KisValue are due to the use of "0" for missing data. Therefore we calculate mismatch due to "0" and mismatch except mismatch caused by "0" coding in Table 4.

Table 4. Mismatch caused by "0" coding of Kisvalue.

Financial Item	N	DART vs. KisValue					
		Mismatch		Mismatch Caused by "0" Coding		Mismatch Except Mismatch Caused by "0" Coding	
Current Asset	1284	323	25.16%	248	19.31%	75	5.84%
Cash and Cash Equivalents	1290	294	22.79%	217	16.82%	77	5.97%
Inventories	1271	271	21.32%	229	18.02%	42	3.30%
Non-Current Asset	1282	310	24.18%	248	19.34%	62	4.84%
Tangible Assets	1288	357	27.72%	244	18.94%	113	8.77%
Intangible Assets	1113	300	26.95%	129	11.59%	171	15.36%
Total Assets	1285	281	21.87%	250	19.46%	31	2.41%
Current Liabilities	1258	278	22.10%	240	19.08%	38	3.02%
Non-current Liabilities	1282	286	22.31%	240	18.72%	46	3.59%
Total Liabilities	1290	281	21.78%	250	19.38%	31	2.40%
Retained Earnings	1276	317	24.84%	244	19.12%	73	5.72%
Total Equity	1290	280	21.71%	250	19.38%	30	2.33%
Sales	1236	268	21.68%	241	19.50%	27	2.18%
Operating Profit	1278	299	23.40%	248	19.41%	51	3.99%
EBIT	1282	282	22.00%	250	19.50%	32	2.50%
Income Tax Expense	1256	257	20.46%	219	17.44%	38	3.03%
Net Profit	1260	1016	80.63%	250	19.84%	766	60.79%
EPS	1196	66	5.52%	0	0.00%	66	5.52%
Overall	22,717	5766	25.38%	3997	17.59%	1769	7.79%

In addition, the overall results in the "DART vs. TS2000" column reveal that 2521 (11.1%) of the total of 22,717 financial facts in TS2000 did not match with the corresponding facts in DART. The number of mismatches across the 18 financial items ranged from 42 (3.3%) to 709 (59.3%). *Intangible Assets* (469), *Operating Profit* (271), and *EPS* (709) were the top three mismatched financial items. In addition, 20.5% of the financial facts (4667) available in DART were not available in TS2000. These omissions were observed across all 18 financial items, with a range of 211 (19.0%) to 388 (30.4%) omitted facts.

We also examined whether no positive earnings were coded as "NA" or "−999." The largest number of missing values was EPS. Therefore, we delineated EPS depending on whether the original data has positive values or negative values in Table 5. As shown Table 5, missing values occurred both in positive EPS as well as negative EPS. As can be seen from these results, it is considered that the error of the financial information company is not related to the attempt to adjust the enterprise profit.

Table 5. Missing on EPS.

Financial Item	N	DART vs. DataGuide		DART vs. KisValue		DART vs. TS2000	
EPS	1196	676 (56.5%)		240 (20.1%)		247 (20.7%)	
		Positive	Negative	Positive	Negative	Positive	Negative
		513 (0.76%)	163 (0.24%)	177 (0.74)	63 (0.26)	182 (0.74)	65 (0.26)

We also examined the types of differences and whether the identified differences were considerably large. In particular, we measured the types of differences (i.e., *DifferenceType* variable) and material differences between DART and each aggregator (i.e., *MaterialResults* variable) and analyzed them. The types of differences are categorized into systematic difference and random difference. The systematic difference results from different coding policies across data aggregators and differences in currency units used. For example, TS2000 includes *Goodwill* in *Intangible Assets* whereas DART, DataGuide, and KisValue do not. Such variance in coding policies causes differences in *Intangible Assets* between DART data and TS2000 data. We refer these types of differences coming from coding policies to the systematic difference. Another example of the systematic difference is the use of "0" for missing data in KisValue. In the case of a missing value, a data aggregator should treat the missing value as null or missing; however, KisValue has a coding policy that replaces a missing value with "0 (zero)" instead. This coding policy leads to differences between DART data and KisValue data. In addition, the use of different currency units also leads to a systematic difference. In particular, DART allows firms to use a different unit of measure (e.g., to denote the numbers in thousands or millions of Korean Won). However, the unit of measure used by all three data aggregators is fixed in thousands of Korean Won. Therefore, differences were often observed between DART data and aggregators' data when the aggregators mistakenly covert numbers denoted in a different unit of measure into their unit of measure. On the other hand, a random difference comes from unexplainable coding errors due to insufficient information about coding practices of raw data. In this study, we classified an unexplainable difference between DART data and aggregators' data as random error, since we could not identify specific causes of the differences due to insufficient information about coding practices in aggregators. Table 6 summarizes the types of differences as well as material differences between DART data and the three aggregators' data.

As shown in Table 6, the numbers of systematic differences are 7 (0.0003%) for DataGuide, 4004 (17.63%) for KisValue, and 365 (1.61%) for TS2000. As discussed above, there are three primary reasons for such differences: (1) KisValue replaces missing value of the original data with "0 (zero)", (2) TS2000 has a policy to report *Intangible Assets* and *Goodwill* together for *Intangible Assets*, and (3) DataGuide, KisValue, and TS2000 use a fixed unit of measure which is often different from the unit of measure in the original DART data. On the other hand, the number of random differences is 797 (3.51%) for DataGuide, 1762 (7.76%) for KisValue, and by 2156 (9.49%) of TS2000, which mostly result from unexplainable coding errors.

In addition, the overall comparison results of material differences in Table 6 indicate that, out of the total of 22,717 financial facts, 206 facts (0.9%) in DataGuide, 4827 facts (21.3%) in KisValue, and 1476 facts (6.5%) in TS2000 were considerably larger or smaller than corresponding facts in DART. The results also show substantial changes in frequency between differences and material differences. The number of discrepancies was reduced from 804 absolute differences (3.5%) to 206 material differences (0.9%) in DataGuide, from 5766 absolute differences (25.4%) to 4827 material differences (21.3%) in KisValue, and from 2521 absolute differences (11.1%) to 1476 material differences (6.5%) in TS2000. Among the 18 financial items, *Current Assets*, *Tangible Assets*, and *Retained Earnings* were three financial items with the highest degree of changes in DataGuide. The three financial items with most changes in KisValue were *Current Assets*, *Retained Earnings*, and *Net Profit* whereas *Intangible Assets*, *Operating Profit*, and *EPS* in TS2000.

In addition to the differences discussed above, we also conducted a sensitivity analysis and checked differences by industry. First, as a sensitivity analysis, we also examined the differences by doubling the materiality level (i.e., 1% of total assets, 10% of income before extraordinary items, 10% of net increase/decrease in cash and cash equivalents, and 10 cents for EPS) to determine how many material errors remain. Similar results were found. Second, for industry effects, we found no systematic differences across the industries.

Table 6. Comparison of types of differences and material differences between DART and three aggregators.

Financial Item	*N* [a]	DART vs. DataGuide						DART vs. KisValue						DART vs. TS2000					
		Types of Difference [b]				Material Difference [b]		Types of Difference				Material Difference		Types of Difference				Material Difference	
		Systemic Difference		Random Difference				Systemic Difference		Random Difference				Systemic Difference		Random Difference			
Current Asset	1284	0	(0.00%)	82	(6.39%)	30	(2.34%)	248	(19.31%)	75	(5.84%)	279	(21.73%)	0	(0.00%)	99	(7.71%)	45	(3.50%)
Cash and Cash Equivalents	1290	0	(0.00%)	96	(7.44%)	8	(0.62%)	217	(16.82%)	77	(5.97%)	224	(17.36%)	0	(0.00%)	97	(7.52%)	10	(0.78%)
Inventories	1271	0	(0.00%)	29	(2.28%)	1	(0.08%)	229	(18.02%)	42	(3.30%)	232	(18.25%)	0	(0.00%)	42	(3.30%)	7	(0.55%)
Non-Current Asset	1282	0	(0.00%)	49	(3.82%)	15	(1.17%)	248	(19.34%)	62	(4.84%)	271	(21.14%)	0	(0.00%)	80	(6.24%)	33	(2.57%)
Tangible Assets	1288	0	(0.00%)	32	(2.48%)	25	(1.94%)	244	(18.94%)	113	(8.77%)	277	(21.51%)	0	(0.00%)	44	(3.42%)	10	(0.78%)
Intangible Assets	1113	0	(0.00%)	51	(4.58%)	18	(1.62%)	129	(11.59%)	171	(15.36%)	158	(14.20%)	268	(24.08%)	201	(18.06%)	246	(22.10%)
Total Assets	1285	0	(0.00%)	29	(2.26%)	1	(0.08%)	250	(19.46%)	31	(2.41%)	253	(19.69%)	0	(0.00%)	46	(3.58%)	11	(0.86%)
Current Liabilities	1258	0	(0.00%)	36	(2.86%)	5	(0.40%)	240	(19.08%)	38	(3.02%)	245	(19.48%)	0	(0.00%)	55	(4.37%)	15	(1.19%)
Non-current Liabilities	1282	0	(0.00%)	30	(2.34%)	5	(0.39%)	240	(18.72%)	46	(3.59%)	245	(19.11%)	0	(0.00%)	59	(4.60%)	13	(1.01%)
Total Liabilities	1290	0	(0.00%)	29	(2.25%)	0	(0.00%)	250	(19.38%)	31	(2.40%)	251	(19.46%)	0	(0.00%)	50	(3.88%)	9	(0.70%)
Retained Earnings	1276	2	(0.00%)	78	(6.11%)	48	(3.76%)	246	(19.28%)	71	(5.56%)	285	(22.34%)	3	(0.24%)	162	(12.70%)	76	(5.96%)
Total Equity	1290	0	(0.00%)	33	(2.56%)	5	(0.39%)	250	(19.38%)	30	(2.33%)	256	(19.84%)	0	(0.00%)	51	(3.95%)	14	(1.09%)
Sales	1236	0	(0.00%)	34	(2.75%)	6	(0.49%)	241	(19.50%)	27	(2.18%)	250	(20.23%)	0	(0.00%)	88	(7.12%)	60	(4.85%)
Operating Profit	1278	1	(0.00%)	67	(5.24%)	15	(1.17%)	249	(19.48%)	50	(3.91%)	270	(21.13%)	92	(7.20%)	179	(14.01%)	163	(12.75%)
EBIT	1282	2	(0.00%)	47	(3.67%)	9	(0.70%)	248	(19.34%)	30	(2.34%)	254	(20.12%)	2	(0.16%)	77	(6.01%)	42	(3.28%)
Income Tax Expense	1256	0	(0.00%)	33	(2.63%)	3	(0.24%)	219	(17.44%)	38	(3.03%)	224	(17.83%)	0	(0.00%)	68	(5.41%)	29	(2.31%)
Net Profit	1260	2	(0.00%)	30	(2.38%)	7	(0.56%)	248	(19.68%)	764	(60.63%)	827	(65.63%)	0	(0.00%)	49	(3.89%)	20	(1.59%)
EPS	1196	0	(0.00%)	12	(1.00%)	5	(0.42%)	0	(0.00%)	66	(5.52%)	22	(1.84%)	0	(0.00%)	709	(59.28%)	673	(56.27%)
Overall	22,717	7	(0.00%)	797	(3.51%)	206	(0.91%)	4004	(17.63%)	1762	(7.76%)	4827	(21.25%)	365	(1.61%)	2156	(9.49%)	1476	(6.50%)

[a] Number of financial facts; [b] Number of types of difference (if a difference is due to different coding policies across aggregators and different currency units, counted as systemic difference whereas if a difference is owing to random errors such as insufficient explanation in the original data and other errors that cannot be identified, counted as a systemic difference.); [c] Number of material differences (material differences are defined as the degree to which absolute value differences are larger than 0.5% of *Total Assets* for balance sheet items, 5% of EBIT for income statement items, and 50 Korean Won for EPS).

4.3. Significance of the Data Differences

To test whether differences were statistically significant, we measured the absolute value difference between DART and each aggregator for the 18 financial items and performed one-tailed, one-sample t-tests. One-tailed, one-sample *t*-tests were performed at a significance level of $p = 0.10$, since the purpose of comparison was to find out the significance of the differences regardless of signs of the differences. Table 7 summarizes the test results.

The results in Table 7 reveal significant differences between DART and DataGuide across eight financial items: *Current Assets* ($p = 0.01$), *Cash and Cash Equivalent* ($p = 0.005$), *Non-current Assets* ($p = 0.003$), *Intangible Assets* ($p = 0.042$), *Current Liabilities* ($p = 0.068$), *Non-current Liabilities* ($p = 0.054$), *Total Equity* ($p = 0.065$), and EPS ($p = 0.001$). With respect to KisValue, all financial items except for Operating Profit and EBIT were significantly different from DART: *Current Assets* ($p = 0.001$), *Cash and Cash Equivalents* ($p = 0.001$), *Inventories* ($p = 0.001$), *Non-current Assets* ($p = 0.01$), *Tangible Assets* ($p = 0.001$), *Intangible Assets* ($p = 0.014$), *Total Assets* ($p = 0.001$), *Current Liabilities* ($p = 0.001$), *Non-current Liabilities* ($p = 0.001$), *Total Liabilities* ($p = 0.001$), *Retained Earnings* ($p = 0.011$), *Total Equity* ($p = 0.001$), *Sales* ($p = 0.001$), *Income Tax Expense* ($p = 0.001$), *Net Profit* ($p = 0.041$), and *EPS* ($p = 0.033$). As addressed before, these results should, however, be interpreted with caution, since KisValue treats missing values as zero. Finally, 13 financial items in TS2000 were significantly different from DART: *Cash and Cash Equivalents* ($p = 0.003$), *Non-current Assets* ($p = 0.098$), *Tangible Assets* ($p = 0.008$), *Intangible Assets* ($p = 0.001$), *Total Assets* ($p = 0.009$), *Current Liabilities* ($p = 0.041$), *Non-current Liabilities* ($p = 0.066$), *Total Liabilities* ($p = 0.088$), *Retained Earnings* ($p = 0.001$), *Total Equity* ($p = 0.024$), *Sales* ($p = 0.027$), *Income Tax Expense* ($p = 0.016$), and EPS ($p = 0.001$).

Table 7. One-sample t-test results of absolute differences between DART and three aggregators.

Financial Item	Absolute Differences (DART vs. DataGuide)					Absolute Differences (DART vs. KisValue)					Absolute Differences (DART vs. TS2000)				
	M	**SD**	**T**	**df**	**p**	**M**	**SD**	**t**	**df**	**P**	**M**	**SD**	**t**	**df**	**p**
Current Asset	2.85	39.72	2.57	1283	0.010	28.09	88.18	11.41	1283	0.001	36.50	802.45	1.63	1283	0.103
Cash and Cash Equivalents	0.11	1.37	2.84	1289	0.005	3.43	13.89	8.87	1289	0.001	0.10	1.28	2.94	1289	0.003
Inventories	0.31	10.95	1.01	1270	0.313	7.56	47.81	5.64	1270	0.001	0.44	11.24	1.39	1270	0.166
Non-Current Asset	0.57	6.78	3.01	1281	0.003	71.04	980.83	2.59	1281	0.010	33.19	717.74	1.66	1281	0.098
Tangible Assets	0.04	0.89	1.56	1287	0.120	24.19	99.54	8.72	1287	0.001	0.41	5.50	2.66	1287	0.008
Intangible Assets	8.67	142.33	2.03	1112	0.042	10.55	142.85	2.46	1112	0.014	39.18	254.03	5.15	1112	0.001
Total Assets	0.05	1.77	1.08	1284	0.282	55.04	170.82	11.55	1284	0.001	0.75	10.21	2.62	1284	0.009
Current Liabilities	1.62	31.50	1.83	1257	0.068	18.70	61.43	10.80	1257	0.001	4.06	70.48	2.04	1257	0.041
Non-current Liabilities	0.33	6.12	1.93	1281	0.054	9.75	103.32	3.38	1281	0.001	1.07	20.79	1.84	1281	0.066
Total Liabilities	0.01	0.19	1.09	1289	0.277	25.22	90.21	10.04	1289	0.001	1.00	21.07	1.70	1289	0.088
Retained Earnings	35.12	829.39	1.51	1275	0.131	60.69	849.04	2.55	1275	0.011	78.88	709.75	3.97	1275	0.001
Total Equity	0.66	12.85	1.84	1289	0.065	30.49	101.11	10.83	1289	0.001	1.25	19.83	2.26	1289	0.024
Sales	0.05	1.30	1.47	1235	0.141	57.37	190.41	10.59	1235	0.001	7.17	113.71	2.22	1235	0.027
Operating Profit	13.51	458.40	1.05	1277	0.292	17.65	458.53	1.38	1277	0.169	15.58	458.81	1.21	1277	0.225
EBIT	23.33	705.51	1.18	1281	0.237	28.23	705.63	1.43	1281	0.152	23.83	705.52	1.21	1281	0.227
Income tax expense	0.10	3.30	1.08	1255	0.282	1.09	5.57	6.92	1255	0.001	0.42	6.23	2.41	1255	0.016
Net profit	23.38	711.53	1.17	1259	0.244	41.88	725.85	2.05	1259	0.041	23.47	711.53	1.17	1259	0.242
EPS	567	496	39.57	1195	0.001	434	7043	2.13	1195	0.033	4633	22,503	7.12	1195	0.001

4.4. Data Differences among Aggregators and Financial Items

We performed an additional analysis to examine whether there were statistically significant differences in data differences among the three aggregators and 18 financial items. In particular, an analysis of covariance (ANCOVA) was conducted to determine the effects of SOURCE (i.e., the three data aggregators) and ITEM (i.e., the 18 financial items) on the *ComparisonResults* variable (i.e., comparison results measure) after controlling for firm size and industry. ANCOVA was used as it enables the examination of the effect of endogenous variables (i.e., data aggregators and financial items) while parsing out the effects of exogenous variables or control variables (i.e., firm size and industry). When exogenous variables and dependent variables are correlated, ANCOVA is usually considered superior at finding treatment effects over ANOVA (analysis of variance). For firm size, we used *Total Assets*. For industry type, we classified industries based on industry classification using the first two digits of the Korean Industry Classification code (see Panel B of Table 1). The ANCOVA results are shown in Table 8.

Table 8. Analysis of covariance results.

Panel (A) ANCOVA results					
Item	**SS**	**Df**	**MS**	**F**	***P***
TA [a]	1.52	1	1.52	15.31	0.001
INDUSTRY [b]	5.11	1	5.12	51.46	0.001
ITEM [c]	248.23	17	14.60	146.82	0.001
SOURCE [d]	516.30	2	258.15	2595.67	0.001
*ITEM * SOURCE*	752.06	34	22.12	222.41	0.001
Error	6189.70	62,237	0.10		

Panel (B) Bonferroni-corrected mean comparison results			
Financial Item	**DataGuide vs. KisValue**	**DataGuide vs. TS2000**	**KisValue vs. TS2000**
Current Asset	(−) ***	(−) *	(+) ***
Cash and Cash Equivalents	(−) ***		(+) ***
Inventories	(−) ***		(+) ***
Non−current Asset	(−) ***	(−) ***	(+) ***
Tangible Assets	(−) ***		(+) ***
Intangible Assets	(−) ***	(−) ***	(−) ***
Total Assets	(−) ***		(+) ***
Current Liabilities	(−) ***	(−) *	(+) ***
Non−current Liabilities	(−) ***	(−) **	(+) ***
Total Liabilities	(−) ***	(−) *	(+) ***
Retained Earnings	(−) ***	(−) ***	(+) ***
Total Equity	(−) ***		(+) ***
Sales	(−) ***	(−) ***	(+) ***
Operating Profit	(−) ***	(−) ***	
EBIT	(−) ***	(−) ***	(+) ***
Income Tax Expense	(−) ***	(−) ***	(+) ***
Net Profit	(−) ***		(+) ***
EPS		(−) ***	

Note: (**A**). [a] TA: Total assets; [b] INDUSTRY: Industry classification based on the first two-digit of Korean Industry Classification code; [c] ITEM: 18 financial items; [d] SOURCE: 3 aggregators, ITEM * SOURCE: Multiplication of item and source (**B**). If DataGuide > KisValue, (+); otherwise (−); If DataGuide > TS2000, (+); otherwise (−); If KisValue > TS2000, (+); otherwise (−); * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

The ANCOVA results in Panel A of Table 8 indicate significant main effects for ITEM ($F_{(17, 62237)} = 146.82$, $p = 0.001$) and SOURCE ($F_{(2, 62237)} = 2595.67$, $p = 0.001$), indicating that data inconsistences among financial items and aggregators were significantly different. The results also show a significant interaction effect between SOURCE and ITEM ($F_{(34, 62237)} = 222.41$, $p = 0.001$), suggesting that the data differences for financial items differed in DataGuide, KisValue, and TS2000.

When a significant interaction occurs, it is generally preferable to consider effects on the individual levels of the other factors instead of interpreting the main effects themselves; therefore, we examined the effect of SOURCE at each level of ITEM. We used Bonferroni-corrected mean comparisons to compare aggregators within each financial item while controlling for firm size and industry. Comparison results (not provided in tabular form) are summarized in Panel B of Table 8. For instance, the results indicate significant differences between DataGuide results and KisValue results ($p < 0.01$) for all of the financial items except for EPS, suggesting that the data differences in DataGuide were significantly lower than those in KisValue (see comparison results in Table 3).

4.5. Estimation of Ohlson's Bankruptcy Prediction Model

Inaccurate financial data distort companies' financial performances and eventually leads to ineffective decisions. Financial institutions are in need of effective prediction models in order to make appropriate lending decisions. Many recent techniques have been employed develop bankruptcy prediction models [35,36]. Liang et al. [37], in particular, extend the outcome of their research to non-financial information. To address the potential effects of differences between the data found in DART and the corresponding data provided by aggregators on financial analysis and academic research, we employ traditional and popular prediction models for general estimation. Hillegeist et al. [38] suggest Ohlson [15] as being one of the most popular prediction models, and therefore we, as an illustration, compared the estimated parameters of Ohlson's bankruptcy prediction model using four data sources. When the significance of coefficients for predictive variables varies depending on the data sources used in the analysis, one should take special care when using the aggregator's data. We chose Ohlson's model for two primary reasons. First, the original Ohlson's model is known to be robust for predicting failure (i.e., bankruptcy) and non-failure firms, and examining factors without considering the comparable size of firms. Hence, despite the small number of bankrupt firms, one can predict bankruptcy within the same model. Second, the model is simple to apply, and, therefore, practical both for practitioners and academicians. Ohlson [15] proposed nine critical financial items from financial statements to predict bankruptcy for a firm. We used the nine items as independent variables and a bankruptcy dummy (0 for a non-bankrupt firm and 1 for a bankrupt firm) as a dependent variable. The bankruptcy prediction model is estimated using the following equation:

$$Y_{jkt} = a_0 + a_1 \times SIZE_{jkt-1} + a_2 \times TLTA_{jkt-1} + a_3 \times WCTA_{jkt-1} + a_4 \times CLCA_{jkt-1} + a_5 \times OENEG_{jkt-1} + a_6 \times NITA_{jkt-1} + a_7 \times FUTL_{jkt-1} + a_8 \times INTWO_{jkt-1} + a_9 \times CHIN_{jkt-1} + e_k$$

where Y = 1 if bankruptcy, or 0 if non-bankruptcy; $SIZE$ = Log (*Total Assets*/*GNP price-level index*); $TLTA$ = *Total Liabilities*/*Total Assets*; $WCTA$ = *Working Capital*/*Total Assets*; $CLCA$ = *Current Liabilities*/*Current Assets*; $OENEG$ = 1 if *Total Liabilities* exceed *Total Assets*, 0 otherwise; $NITA$ = *Net Income*/*Total Assets*; $FUTL$ = Funds provided by operations/*Total Liabilities*; $INTWO$ = 1 if *Net Income* was negative for the last two years, 0 otherwise; $CHIN = (NI_t - NI_{t-1})/(|NI_t| + |NI_{t-1}|)$, where NI is *Net Income* for each firm, j, and each kth data aggregators (i.e., DART, KisValue, DataGuide, and TS2000) for the tth time period (i.e., t = 2012 and $t - 1$ = 2011).

Following Ohlson's approach, logit regression analyses were conducted due to the dummy dependent variable (i.e., non-bankruptcy or bankruptcy). The purpose of the analysis was to examine whether the predictions of the model using the data provided by each aggregator were virtually the same compared to the predictions of the model using the original data found in DART. Since missing values appeared randomly across the four data sources (i.e., DART, KisValue, DataGuide, and TS2000), the model comparisons after removing the common missing cases from all of the data sources might not have warrantied appropriate prediction results. We therefore removed any cases both from the model with DART data and from the model with an aggregator's data if the cases were not found in either DART or the aggregator. We then performed a separate logit regression for each model

comparison (i.e., three pairs of logit regressions). Table 9 summarizes the logit regression results. We predicted that the different data sources would result in different predicting variables.

Table 9. Estimation results of the Ohlson bankruptcy prediction model.

Financial Item	DART vs. DataGuide		DART vs. KisValue		DART vs. TS2000	
	DART	DataGuide	DART	KisValue	DART	TS2000
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
SIZE *	0.09 * 1.10 * (0.47) *	0.05 1.05 (0.26)	0.18 1.20 (0.68)	0.25 1.29 (0.88)	0.37 1.45 (1.30)	0.46 1.59 (1.56)
TLTA *	5.95 ** 387.55 (2.30)	6.72 *** 834.37 (2.66)	6.54 ** 697.53 (1.99)	6.58 ** 722.69 (2.08)	4.37 79.39 (1.35)	3.53 34.40 (0.97)
WCTA *	−0.08 0.92 (−0.03)	0.74 2.10 (0.30)	−0.35 0.70 (−0.10)	−0.03 0.97 (−0.16)	−1.30 0.27 (−0.37)	−3.39 0.03 (−0.93)
CLCA *	0.70 2.02 (1.41)	0.65 1.93 (1.37)	0.89 2.45 (1.07)	1.09 * 2.98 (1.78)	0.88 2.43 (1.50)	0.65 1.92 (1.24)
OENEG *	−1.95 0.14 (−1.27)	−1.94 0.14 (−1.29)	−7.60 *** 0.00 (−2.78)	−7.03 *** 0.00 (−2.72)	−5.84 ** 0.00 (−2.28)	−3.42 0.03 (−1.45)
NITA *	−5.65 *** 0.00 (−2.81)	−5.29 *** 0.01 (−2.65)	−9.42 *** 0.00 (−2.92)	−10.72 *** 0.00 (−3.14)	−14.48 *** 0.00 (−3.82)	−8.43 ** 0.00 (−2.36)
FUTL *	−2.19 0.11 (−1.01)	−1.97 0.14 (−0.88)	−3.83 0.02 (−1.01)	−3.60 0.03 (−0.91)	−6.13 0.00 (−0.02)	−2067.23 0.00 (−0.70)
INTWO *	−0.60 0.55 (−0.72)	−0.56 0.57 (−0.68)	0.24 1.28 (0.21)	0.33 1.40 (0.27)	−0.82 0.44 (−0.74)	−0.83 0.43 (−0.67)
CHIN *	−1.17 * 0.31 (−1.93)	−1.16 ** 0.31 (−1.97)	−3.75 ** 0.02 (−2.53)	−4.26 0.01 (0.013)	6.72 ** 0.03 (2.10)	−3.21 0.04 (−0.97)
Constant	−10.97 ** (−2.36)	−10.33 ** (−2.22)	−15.73 ** (−2.45)	−18.18 *** (−2.60)	−16.91 *** (−2.58)	−18.08 *** (−2.68)
Observations	608	608	481	481	473	473
R−square	0.41	0.40	0.58	0.60	0.53	0.50
Pseudo R−square	0.41	0.40	0.58	0.60	0.53	0.50

* *SIZE* = Log(Total Assets/GNP price-level index); * *TLTA* = Total Liabilities/Total Assets; * *WCTA* = Working Capital/Total Assets; * *CLCA* = Current Liabilities/Current Assets; * *OENEG* = One if Total Liabilities exceed Total Assets, zero otherwise; * *NITA* = Net Income/Total Assets; * *FUTL* = Funds provided by operations/Total Liabilities; * *INTWO* = One if Net Income was negative for last two years, zero otherwise; * *CHIN* = $(NI_t - NI_{t-1})/(|NI_t| + |NI_{t-1}|)$, where NI is Net Income; j: coefficient; k: odd ratio; l: *t* value; Note: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$; Parentheses indicate *t*-values.

The logit regression results in Table 9 indicate that important predictors of bankruptcy differed across data sources. It should be noted that the sample sizes for three matched data are different because of the different number of missing values for each database. Overall, although we did not intend to compare the size of *R-squares*, the analysis results of three data sets show large discrepancies in *R-squares*, which represent model fits; the *R-squares* are 0.40 for DataGuide, 0.60 for KisValue, and 0.50 for TS2000, respectively.

Furthermore, the significance of coefficients for predictive variables varied as well. For example, for the comparison between DART and DataGuide, two variables were statistically significant for both DART and DataGuide: *TLTA* (a_2: coefficient = 5.95, odd ratio = 387.55, $p < 0.05$ for DART and a_2: coefficient = 6.72, odd ratio = 834.37, $p < 0.01$ for DataGuide) and *NITA* (a_6: coefficient = −5.65, odd

ratio = 0.00, $p < 0.01$ for DART and a_6: coefficient = −5.29, odd ratio = 0.01, $p < 0.01$ for DataGuide). *CHIN* was significant for DataGuide (a_9: coefficient = −1.16, odd ratio = 0.31, $p < 0.05$) but marginally significant for DART (a_9: coefficient = −1.17, odd ratio = 0.31, $p < 0.10$). On the other hand, the model comparison results between DART and KisValue show that three variables were statistically significant for both DART and KisValue: *TLTA* (a_2: coefficient = 6.54, odd ratio = 697.53, $p < 0.05$ for DART and a_2: coefficient = 6.58, odd ratio = 722.69, $p < 0.05$ for KisValue), *OENEG* (a_5: coefficient = −7.60, odd ratio = 0.00, $p < 0.01$ for DART and a_5: coefficient = −7.03, odd ratio = 0.00, $p < 0.01$ for KisValue) and *NITA* (a_6: coefficient = −9.42, odd ratio = 0.00, $p < 0.01$ for DART and a_6: coefficient = −10.72, odd ratio = 0.00, $p < 0.01$ for KisValue). However, *CHIN* (a_9: coefficient = −3.75, odd ratio = 0.02, $p < 0.01$) was significant only for DART, and *CLCA* was marginally significant for KisValue (a_4: coefficient = 1.09, odd ratio = 0.01, $p < 0.10$). With respect to the comparison between DART and TS2000, only *NITA* (a_6: coefficient = −14.48, odd ratio = 0.00, $p < 0.01$ for DART and a_6: coefficient = −8.43, odd ratio = 0.00, $p < 0.05$ for TS2000) was statistically significant for both DART and TS2000. *OENEG* (a_5: coefficient = −5.84, odd ratio = 0.00, $p < 0.05$) and *CHIN* (a_9: coefficient = 6.72, odd ratio = 0.03, $p < 0.05$) were significant for DART, but not for TS2000. The results of the logit regression analysis seem to be consistent with the results of the differences versus material differences shown in Tables 3 and 4. In other words, the least inconsistent data in DataGuide provides fairly consistent results with DART.

Overall, our analysis shows that the different data sources used in the bankruptcy prediction model might lead to different predictions of bankruptcy for firms. In other words, the differences between the data reported in DART and the corresponding data provided by the aggregators might cause inappropriate analysis, which, in turn, might lead to unintended consequences in financial analysis and academic research.

5. Conclusions and Discussion

This study examined 18 financial items of 645 firms in South Korea for the two fiscal years of 2011 and 2012. In particular, we compared the data reported in firms' original filings (i.e., DART) with the corresponding data provided by three prominent aggregators in South Korea to investigate the similarities and differences between them. The results of the study show a significant number of differences between DART and the three aggregators; specifically, KisValue (25.4%) had the largest number of differences, followed by TS 2000 (11.1%), and DataGuide (3.5%). Although it seems obvious, the number of material differences was much smaller than the number of differences. KisValue (21.2%) had the largest number of material differences, followed by TS 2000 (6.5%), and DataGuide (0.9%). In addition, financial items with material differences varied across the aggregators. Compared to DART, DataGuide had statistically significant differences in eight financial items, while KisValue and TS2000 had 16 items and 13 items, respectively. Finally, the ANCOVA results suggest that there were statistically significant differences in data differences among the three aggregators and 18 financial items.

5.1. Reasons for Data Differences

Given that DART provides the most accurate data that reflect firms' intended communications in their official financial reports, the data differences between DART and the aggregators put the quality of data provided by the commercial data aggregators in question. This data quality issue might have a serious impact not only on investors' decision-making but also on academic research, as we demonstrated in the estimation of Ohlson's [15] bankruptcy prediction model.

There are two primary reasons for data differences between DART and aggregators. First, differences occur when aggregators re-organize firms' financial data to make them more comparable across companies. Although aggregators in South Korea usually rely on DART as a primary source of firms' financial data, they often merge or reclassify original data to fit into their proprietary data formats based on their own operationalized definition of financial items. In other words, they frequently maintain financial item names used in DART but use their criteria to make data more comparable (i.e.,

a unique standardization process). For example, we found that *Intangible Assets* in TS2000 represents the sum of *Intangible Assets* and *Goodwill* in DART; therefore, users of TS2000 would consider the sum of *Intangible Assets* and *Goodwill* as *Intangible Assets* while users of DART and other aggregators would treat them separately.

Second, differences also arise from differences in coding rules and errors during data entry processes. Yang et al. [12] found that many data errors result from unidentifiable or unexplainable coding errors. Positive numbers are often entered as negative numbers, and vice versa, depending on the definitions used to define the attributes of items. For instance, we found that KisValue had a total of 257 errors in *Income Tax Expenses*. Out of the 257 errors, 64 positive values in DART were simply coded as negative values in KisValue while 32 negative values were coded as positive values; furthermore, some differences occurred while coding values. Some aggregators keep the same base unit (e.g., thousands) shown in DART even though they use different base units (e.g., millions); that is, the accuracy of values varies across aggregators. If the data aggregators correct these errors resulting from coding practices and currency units, a large portion of data differences shown in Table 3 will be reduced.

5.2. Implications, Limitations, and Future Research

The overall implication of this study is that financial data offered by aggregators in South Korea have inconsistences due to differences in coding rules, coding errors, and different standardization processes. Our findings can help aggregators to improve their data quality by demonstrating that their data acquisition processes lead to differences between their data and DART data. If the differences are not due to errors but due to differences in coding rules and standardization processes designed to make data more comparable, then aggregators should clearly inform their users of such differences. In addition, users who mainly rely on data provided by aggregators can benefit from the findings of this study by understanding the nature and extent of the differences and how such differences lead to unintended consequences in financial analysis and academic research.

It is very critical for investors and academic researchers to make decisions, even bankruptcy predictions, based on reliable and accurate financial data. However, unlike common belief, our empirical research showed that commercial databases have very high error rates at 25.4%, 11.1%, and 3.5% for KisValue, TS2000, and Dataguide respectively. Therefore, our findings are of paramount importance. Bankruptcy prediction is a very important task for many related financial institutions [37], and, recently, many techniques have been employed to develop bankruptcy prediction models [36,39]. To that end, the reliability of financial data before prediction for bankruptcy should be considered and ensured. It is urgent to tightly control and redeem even minor errors and improve data quality.

As our study has shown that outstanding financial data may differ from the original corporate financial data in South Korea, we contribute to the studies of Rosenberg and Houglet [10], Bennin [8], Kinney and Swanson [9] and Yang et al. [12]. These studies have mostly examined western databases and contexts; data quality researches need to investigate other economic regions. A lot of foreign institutional investors are investing in Asian markets, but investing without checking data reliability and accuracy can be dangerous. For this reason, this field of research is needed in a wider variety of countries. As with any study, there are several limitations that must be considered when interpreting the findings. This study classified the types of differences into systematic difference and random difference. We found that several differences arise from differences in code systematically (e.g., the merging of *Intangible Assets* and *Goodwill*, and treatment of missing values). However, for some differences, we could not identify the causes of differences, and thus we classified these as random differences. Future studies are necessary to examine underlying causes of these random differences.

For comparison purposes, our sample excludes financial firms (i.e., firms in finance and insurance industries) due to different accounting rules and regulations. Non-financial firms report *Current Assets* and *Non-Current Assets* in balance sheets and *Sales* in income statements, while financial firms do not report these items due to the nature of their businesses. In addition, we examined 18 common financial

items for the two fiscal years of 2011 and 2012; thus, the findings of the study might not necessarily reflect the practices of firms in finance and insurance industries as well as other financial items and years. Future research should not only include firms in finance and insurance industries but should also consider other financial items and prior and/or post years.

Despite some limitations, the findings of this study are noteworthy. Commercial financial data aggregators often claim that they contribute to augmenting the value of original data by providing supplementary analyses and additional data that cannot be found in original financial statements; however, such claims become groundless if differences exist between original data and their data due to their data acquisition processes. Given the number of differences identified in this study, the use of data provided by aggregators should be carefully scrutinized. In particular, we suggest utilizing an existing technology called extensible business reporting language (XBRL). XBRL is a standardized method for preparing, publishing, and exchanging business information, especially financial information. It was developed to make financial information more accessible and easier to analyze for analysts, investors, regulators, and related parties. Regulators and government agencies around the world are increasingly implementing XBRL for mandatory filings. South Korea has also implemented XBRL. Since 2011, all publicly listed firms are required to prepare their financial statements in XBRL format and to file XBRL-based financial reports (e.g., annual and quarterly reports) with DART [40,41]. Since XBRL is a machine-readable format, users of XBRL reports can easily extract data and use them for analysis. There is early evidence that XBRL provides more reliable market information and improves decision quality [42,43]. Aggregators can also benefit from XBRL by enhancing their data acquisition processes (i.e., automatically acquiring original data). Before 2016, all XBRL reports were not publicly available in DART. The FSS allowed firms to decide whether they desire to make their XBRL reports publicly available in DART, and most firms decided not to do so. However, the FSS recently made all XBRL reports available in DART, meaning that both aggregators and users of financial data can obtain credible financial data and use them in their analysis and decision-making since June 2016.

Finally, future studies need to extend the current study by investigating the reliabilities of financial databases used in other emerging markets such as Malaysia, Hong Kong, Australia, as analysts and researchers alike rely on commercial databases more often. In addition, given that the data reported in firms' filings differ from the aggregators' corresponding data, it is worthwhile to examine the effects of errors in the databases on the time-series properties of aggregated data in the long run. Finally, although much attention is now paid to the analytical approaches of "big data" in the financial sector, research is needed to examine the reliability of raw data as fundamentals of those "big data" approaches.

Author Contributions: Hyunjung Nam and Won Gyun No conceived and designed the research; Hyunjung Nam undertook the primary research and Won Gyun No was in charge of analysis of the data. Youngsu Lee contributed to organizing all sections and critical revision. All of authors read and approved the final manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Fisher, C.; Lauria, E.; Chengalur-Smith, S. *Introduction to Information Quality*; AuthorHouse: Bloomington, Indiana, 2012.
2. Tushman, M.L.; Nadler, D.A. Information processing as in an integrating concept in organizational design. *Acad. Manag. Rev.* **1978**, *3*, 613–624.
3. Bryce, M.; Ali, M.J.; Mather, P.R. Accounting quality in the pre-/post-IFRS adoption periods and the impact on audit committee effectiveness—Evidence from Australia. *Pac.-Basin Financ. J.* **2015**, *35*, 163–181. [CrossRef]
4. Allen, A.; Cho, J.Y.; Jung, K. Earnings forecast errors: Comparative evidence from the Pacific-Basin capital markets. *Pac.-Basin Financ. J.* **1997**, *5*, 115–129. [CrossRef]
5. Gao, J.; Wang, J. Is Working Capital Information Useful for Financial Analysts?—Evidence from China. *Emerg. Mark. Financ. Trade* **2017**, *53*, 1135–1151. [CrossRef]
6. Baesens, B.; Mues, C.; Martens, D.; Vanthienen, J. 50 years of data mining and OR: Upcoming trends and challenges. *J. Oper. Res. Soc.* **2009**, *60*, S16–S23. [CrossRef]

7. Koziol, H. Incorrect advice to investors and the liability of banks. *J. Contemp. Roman-Dutch Law* **2011**, *74*, 1–11.
8. Bennin, R. Error rates in CRSP and COMPUSTAT: A second look. *J. Financ.* **1980**, *35*, 1267–1271. [CrossRef]
9. Kinney, M.; Swanson, E. The accuracy and adequacy of tax data in Compustat. *J. Am. Tax. Assoc.* **1992**, *15*, 121–135.
10. Rosenberg, B.; Houglet, M. Error rates in CRSP and Compustat data bases and their implications. *J. Financ.* **1974**, *29*, 1303–1310. [CrossRef]
11. Tallapally, P.; Luehlfing, M.; Motha, M. The partnership of EDGAR Online and XBRL-Should Compustat care? *Rev. Bus. Inf. Syst.* **2011**, *15*, 39–46. [CrossRef]
12. Yang, D.; Vasarhelyi, M.; Liu, C. A note on the using of accounting databases. *Ind. Manag. Data Syst.* **2003**, *103*, 204–210. [CrossRef]
13. Choi, S.; Kim, Y. Market mispricing on real earnings management and foreign investors. *Korean Account. Rev.* **2013**, *38*, 113–144.
14. Chung, S. Innovation, competitiveness, and growth: Korean experiences. *Annu. World Bank Conf. Develop. Econ.* **2010**, 333–357. Available online: http://www.rrojasdatabank.info/wbdevecon10-22.pdf (accessed on 5 July 2013).
15. Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
16. Financial News. From Korea Discount to Korea Premium via Transparent Accounting Practices. *Financial News*, 26 November 2013. Available online: http://www.fnnews.com/news/201311261702244009 (accessed on 27 November 2013). (In Korean)
17. Ko, K.C.; Lin, S.J.; Su, H.J.; Chang, H.H. Value investing and technical analysis in Taiwan stock market. *Pac.-Basin Financ. J.* **2014**, *26*, 14–36. [CrossRef]
18. Ehalaiye, D.; Tippett, M.; Zijl, T. The predictive value of bank fair values. *Pac.-Basin Financ. J.* **2016**, *41*, 111–127. [CrossRef]
19. Jeong, S.; Kwak, S.; Hwang, I. Accounting Transparency and National Competitive Advantage. In *Proceedings of the 2010 Annual Conference on Accounting*; Korean Acconting Association: Seoul, Korea, 2010; pp. 1–24. (In Korean)
20. Winkler, J.; Kuklinski, C.P.J.-W.; Moser, R. Decision making in emerging markets: The Delphi approach's contribution to coping with uncertainty and equivocality. *J. Bus. Res.* **2015**, *68*, 1118–1126. [CrossRef]
21. Le, T.H.; Kim, J.; Lee, M. Institutional Quality, Trade Openness, and Financial Sector Development in Asia: An Empirical Investigation. *Emerg. Mark. Financ. Trade* **2016**, *52*, 1047–1059. [CrossRef]
22. Courtenay, S.M.; Keller, S.B. Errors in databases revisited—An examination of the CRSP shares-outstanding data. *Account. Rev.* **1994**, *69*, 285–291.
23. Kern, B.B.; Morris, M.H. Differences in the Compustat and expanded value line databases and the potential impact on empirical research. *Account. Rev.* **1994**, *69*, 274–284.
24. San, M.J. The reliability of R&D data in Compustat and 10-K Reports. *Account. Rev.* **1977**, *52*, 638–641.
25. Financial Supervisory Service. Available online: http://englishdart.fss.or.kr (accessed on 5 July 2013).
26. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
27. Altman, E.I.; Haldeman, R.G.; Narayanan, P. ZETA analysis: A new model to identify bankruptcy risk of corporations. *J. Bank. Financ.* **1977**, *1*, 29–54. [CrossRef]
28. Becker, C.L.; Defond, M.L.; Jiambalvo, J.; Subramayam, K.R. The effect of audit quality on earnings management. *Contemp. Account. Res.* **1998**, *15*, 1–24. [CrossRef]
29. Lo, K.; Lys, T. The Ohlson model: Contribution to valuation theory, limitations, and empirical applications. *J. Account. Audit. Financ.* **2000**, *15*, 337–367. [CrossRef]
30. Dataguide.co.kr. [Homepage on the Internet]. FnGuide Inc.: Seoul, Korea. Available online: http://www.dataguide.co.kr (accessed on 30 June 2013).
31. Kisvalue.com. [Homepage on the Internet]. NICE Information Service: Seoul, Korea. Available online: www.kisvalue.com (accessed on 3 July 2013).
32. Kocoinfo.co.kr. [Homepage on the Internet]. Korean Listed Companies Association: Seoul, Korea. Available online: http://www.kocoinfo.com (accessed on 30 June 2013).

33. Leslie, D.A. *Materiality: The Concept and Its Application to Auditing*; The Canadian Institute of Chartered Accountants: Toronto, ON, Canada, 1985.
34. Eilifsen, A.; Messier, W.F., Jr. Materiality guidance of the major public accounting firms. *J. Pract. Theory* **2015**, *34*, 3–26. [CrossRef]
35. Balcaen, S.; Ooghe, H. 35 years of studies on business failure: An overview classic statistical methodologies and their related problems. *Br. Account. Rev.* **2006**, *38*, 63–93. [CrossRef]
36. Zainudin, E.F.; Hashim, H.A. Detecting fraudulent financial reporting using financial ratio. *J. Financ. Rep. Account.* **2016**, *14*, 266–278. [CrossRef]
37. Liang, D.; Lu, C.C.; Tsai, C.F.; Shin, G.A. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *Eur. J. Oper. Res.* **2016**, *252*, 561–572. [CrossRef]
38. Hillegeist, S.K.; Keating, E.K.; Cram, D.P.; Lundstedt, K.G. Assessing the probability of bankruptcy. *Rev. Account. Stud.* **2004**, *9*, 5–34. [CrossRef]
39. Kumar, P.R.; Ravi, V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques—A review. *Eur. J. Oper. Res.* **2007**, *180*, 1–28. [CrossRef]
40. Money Today. XBRL, A Revolution in Financial Reporting. *Money Today*. 10 November 2014. Available online: http://www.mt.co.kr/view/mtview.php?type=1&no=2014110915383631556&outlink=1 (accessed on 11 November 2014).
41. Financial Supervisor Service, IFRS-Based XBRL Public Disclosure System. 2010. Available online: http://dart.fss.or.kr/dsaa003/selectGuide.do (accessed on 30 June 2013). (In Korean)
42. Efendi, J.; Park, J.D.; Smith, L.M. Do XBRL filings enhance informational efficiency? Early evidence from post-earnings announcement drift. *J. Bus. Res.* **2014**, *67*, 1099–1105.
43. Yoon, H.; Zo, H.; Ciganek, A.P. Does XBRL adoption reduce information asymmetry? *J. Bus. Res.* **2011**, *64*, 157–163. [CrossRef]