Alternative Data in Investment Management: Usage, Challenges and Valuation

Gene Ekster[1] and Petter N. Kolm[2]

October 20, 2020

Abstract

Alternative data in finance is an umbrella term for diverse non-traditional datasets used by quantitative and fundamental institutional investors to enhance portfolio returns. While the use of alternative data is a recent phenomenon, it was not until the last five years that it gained widespread acceptance and the sector had evolved into a complex ecosystem of data originators, intermediaries and investors. The alternative data industry faces several obstacles, including difficulty estimating a dataset's potential value to investors and technical challenges for leveraging these datasets efficiently at large scale. In this article, we provide an up-to-date description of the alternative data space as it relates to the institutional investment industry. We elaborate on what alternative data is and how it is used in investment management today. We identify and discuss some of the key challenges that arise when working with alternative data. In particular, we address issues such as entity mapping and ticker-tagging, panel stabilization and debiasing with modern statistical and machine learning approaches. We advance several methodologies for the valuation of alt-datasets,

[1] Gene Ekster (email: gene@altdg.com) is at the Alternative Data Group.

[2] Petter N. Kolm (email: petter.kolm@nyu.edu) is at the Courant Institute of Mathematical Sciences, New York University.

including an event study methodology we refer to as the Golden Triangle, the application of report cards, and the relationship between the structure of its information content and potential to enhance investment returns. To illustrate the effectiveness of the methods, we apply them to a case study analysis of real-world healthcare data, delivering an improvement of revenue prediction accuracy from an 88% mean absolute error to a 2.6% mean absolute error.

Keywords: Alternative data; alt-data; fundamental investing, investment management; investment strategies; machine learning; quantitative investing; unstructured data.

1 Introduction

Since 2014 the institutional investment industry has undergone an arms race of procuring and employing alternative datasets (alt-datasets) intending to boost investment performance (Hope 2015). This arms race has driven the industry into the utilization of diverse non-traditional unstructured datasets at a larger scale than before. The adoption of these datasets has been rapid. In 2014 only a few funds employed non-traditional data and the term "alternative data" (alt-data) did not exist in the investment vernacular (Ekster 2014). However, only five years later alt-data has become a tour-de-force with conferences attracting thousands of attendees, three out of four financial institutions housing alt-data teams and 90% of the firms expanding their alt-data strategy (Denev and Amen 2020). While there is no question that the buzz of alt-data has spread quickly in the industry, there is still skepticism about its ability to enhance investment returns. Much of the hesitation stems from the significant technical challenges required to leverage alt-data and the costs associated with processing and analyzing it. In this article, we identify some of these challenges and discuss possible solutions for them.

Alt-data is less structured and more obscure than its traditional data counterparts like security prices, options statistics, regulatory filings, or bulge bracket sell-side reports. The integration of non-traditional information is significantly more challenging than the methods driving standard non-alt-data.

This article makes several contributions. First, we provide an up-to-date description of the alt-data space as it relates to the institutional investment industry. We elaborate on what alt-data is and how it is used in investment management today. Second, we identify and discuss the key challenges that arise when using and working with alt-data. In particular, we address issues such as entity mapping and ticker tagging, panel stabilization and debiasing with modern statistical and machine learning approaches. Third, we advance several methodologies for the valuation of alt-datasets, including an event study methodology we refer to as the Golden Triangle, the application of report cards, and the relationship between the structure of its data content and potential to enhance investment returns. The latter sheds light on some of the differences between raw datasets and aggregated data products. Fourth, to illustrate the effectiveness of the methods, we apply them to a case study analysis of real-world healthcare purchasing data, thereby detailing the data processing pipeline we use to perform entity tagging, handle outliers and impute missing data. Our model delivers an improvement of revenue prediction accuracy from an 88% mean absolute error (MAE) to a 2.6% MAE. The article closes with some thoughts about the future of the alt-data space.

2 Background

2.1 Definition of Alternative Data

We define alternative data as any dataset which does not originate from the securities exchanges, regulatory disclosures, or economic releases and indicators. To the best of our knowledge, alt-data originated in 2014 as an industry-defining term, yet the non-traditional data industry has existed long before that. Today, alt-data is arguably mainstream, with many investment management companies having an alt-data team within their organization (Monk, Prins, and Rook 2019). Alt-datasets are frequently large and unstructured. They are also typically not purpose-made to be used by the financial industry, instead alt-data is often “exhaust” data from every-day business operations. Examples of alt-datasets include consumer transaction data, satellite imagery, vehicle movements, bills of lading, cargo locations, cell phone geolocation, and social media data. Of note is that since the identification of individuals is not useful for funds, data is frequently anonymized directly at the source.

2.2 The Alternative Data Ecosystem

The alt-data ecosystem consists of three types of constituents: (a) originators, (b) intermediaries and (c) consumers. We discuss each type next.

2.2.1 Originators

Originators represent the entry point of alt-datasets into the investment industry. While the distinction between originators and intermediaries is sometimes unclear, understanding the difference is pivotal to the success of the industry's veteran funds. First-line ownership and full down-stream distribution control define the originator organization. Whether the data ownership is literal or merely de-facto, the downstream impact is the same. The originator becomes the top-level source of the raw data, obtaining the data in several different ways:

- They have unrelated core operations and sell the data to the alt-data industry as a secondary business. For example, while Mastercard's (MC) primary business is facilitating global credit card transactions, they also aggregate, anonymize and sell consumer spending data to asset managers (Jaquez 2020).
- Disparate or overlooked publicly available data such as websites, obscure municipal or other public records. While accessible to anyone, the practical collection is resource prohibitive and is often limited to ongoing data, while historical data is only available from third parties. For example, Buildfax Property Intelligence Solutions aggregates residential and commercial permit data (Helm 2020).
- Originators that are not the ultimate source of the data may have long-term exclusive licensing and usage rights from the actual source(s), thus making them the de-facto source

data controllers. For example, Yodlee aggregates and controls the distribution of transaction-level consumer purchasing data (Hope 2015).

- Primary research companies conducting their own surveys are also considered originators. For example, Gallup Inc. is known for their worldwide public opinion polls.

As of this writing, hundreds of data originators exist, including main players such as Yodlee, SimilarWeb, Yipit, Thinknum, Verizon, Orbital Labs, Open Signal, Inrix, Jump Shot, Gallop, Pew and Nielson (Kolanovic and Smith 2019).

2.2.2 Intermediaries

Due to its non-traditional nature, alt-data is frequently unstructured, biased and not assembled with institutional investors in mind. While the data originators typically sell the data in its raw form, there is a rich ecosystem of go-between intermediaries who bridge the gap across data owners and the funds that use the data. Intermediaries can act as data brokers, full-service research boutiques, or anything in between. The intermediary space has evolved into several types of providers:

- A “yellow pages” listing for alt-data vendors such as alternativedata.org.
- Brokers of unaltered datasets. An example includes Nudata.
- Data curators that create dashboards and Excel products from raw data. The majority of intermediaries are included in this group. Examples of such intermediaries include 1010data, 7park, Earnest Research and Consumer Edge.

Financial equity research companies use alt-data to perform analysis for their research reports. Their products are comparable to traditional sell-side equity research reports and full-service analyst support, including some GAAP forecasts. Examples of such providers include M Science and UBS Evidence Labs.

If a company both originates and markets a data product, it is vertically integrated, blurring the distinction between an originator and intermediary (Jagtiani and Lemieux 2019). Mastercard fits the vertical integration description since they originate, assemble, aggregate and market their data directly.

Misunderstanding the difference between an originator and an intermediary is perhaps the most common mistake leading to costly errors by alt-data novices (Deloitte 2020). Although intermediaries indeed add immense value to the alt-data supply chain, buyers of alt-data need to keep the following intermediary dynamics in mind:

- Misaligned incentives: Alt-data research is most often priced per the number of companies under coverage and the number of distinct datasets offered. However, if the data product does not accurately forecast key investable metrics, its true value is lower, yet the price may not reflect it. Thus, for an intermediary a logical business choice is to increase the number of datasets and their coverage rather than invest into accuracy. This creates a problematic misalignment of incentives where the intermediaries are not compensated in a way that maximizes value for their buy-side clients. This problem will persist for as long as alt-data fee structures are based on coverage metrics rather than forecasting accuracy.
- Data aggregation mistakes: As we will discuss later in this article, alt-data product creation is a capital-intensive R&D process. Cost-saving motives can compel firms to gloss over some of the technical aspects that can lead to errors downstream. Mistakes in critical steps, including entity resolution, ticker tagging, paneling, debiasing or revenue modeling can result in low product accuracy. Most intermediaries' techniques and methodologies are black-box systems, not available for audits by customers, thus exacerbating aggregation errors because of a lack of transparency.

- Aggregated data has less alpha potential than the raw source because aggregation is often a destructive process that can lower the ability to find unique investment opportunities. Related to this is the concept of alpha decay which occurs when similar alt-data products are shared amongst competing funds, resulting in the insights from the data being priced into security prices at a faster pace (see section 4.3).

2.2.3 Investment Professionals

It is easy to mistake alt-data-driven investing as automated and systematic quantitative strategies. Unlike the structured data powering traditional quantitative systems, unstructured alt-datasets often require more effort to be monetized (Kolanovic and Smith 2019). Compared to deployment of traditional data, alt-data is more difficult to automate and systematize into the investment process. In practice, only experienced specialists have the skillset needed to analyze, process and monetize them.

Fundamental Funds

Given the challenges with alt-data described above, most alt-data operations amplify the benefit of having fundamentally driven teams, rather than provide substitutes for them.

While predicting revenues is likely the most apparent use of alt-data today, it is not necessarily the most lucrative. An informal survey of top fund managers suggests that investing using alt-data to inform fundamental thesis-based bets is the favored approach of seasoned alt-data informed teams (Eagle Alpha 2019). Nevertheless, revenue modeling has its place in an investing process, but more often than not, revenue modeling is better suited for validating a datasets predictive accuracy rather than using it to drive investment strategies.

It turns out that the relationship between revenue surprises and related equity return performance is small. Research measuring the relationship of revenue surprises to U.S. stock prices around earnings announcements finds that the correlation coefficient for the last 15 years is a paltry 0.18. For many traditional investors, this correlation is too low to be actionable after taking trading costs into account. In other words, if an investor acted on perfect revenue announcement information for every firm in the S&P 500, they would only profit 52% of the time (randomly guessing would end up at 50%). The statistical correlation can be improved by focusing efforts on only the fast-growing companies, but results in only a 0.1 point improvement from 0.18 to 0.28 (Ekster 2015b). Consequently, funds gravitate away from utilizing alt-data for revenue-surprise focused strategies towards using it more holistically.

Quantitative Funds

A common misconception about alt-data is that quantitative hedge funds are one of its biggest users. However, issues including shortage of historical data, a small coverage universe and various irregularities specific to each alt-dataset limit their applicability for algorithmic-only funds.

Despite these challenges, a few quantitative funds have been using alt-data longer than other types of asset managers (Denev and Amen 2020). Even if quantitative funds are not forthcoming with their methodology, what is known is that some are using alt-data as important inputs, usually combined with more traditional datasets. For instance, some systematic alt-data practitioners use approaches similar to latent factor analysis for predicting earnings quality (Du et al. 2020), including principal components (PCs) or hidden Markov models (HMMs), to predict revenue surprises and take appropriate directional bets.

As alt-data vendors mature in sophistication and consistency of their offerings, the number of quantitative firms buying at least some type of alt-data increases. The number of quantitative funds using alt-data has grown from less than 10% three years ago to over 80% today (Eagle Alpha 2019).

3 Challenges with Alternative Data

As alt-data is frequently unstructured, unprocessed and not purpose-built for the institutional investment customer, it comes with its challenges in processing, interpretation and application in investment management. Currently, there are few commercial tools available for data providers and practitioners to address these challenges. Nevertheless, some “best practices” are starting to emerge on alt-data related tasks including entity mapping and ticker tagging, panel stabilization and debiasing.

3.1 Entity Mapping and Ticker Tagging

What transactional, inventory, location, written language, web crawled and almost all other alt-datasets have in common is that in contrast to traditional financial datasets, they do not contain standardized company names or tickers. A data row from a transactional dataset might read “Doritos Locos Taco” along with the price of the purchase. How can we map this purchase to the appropriate company name and ticker symbol?

A simple manual solution is to use keyword lookups or regular expressions to map unstructured text to their corresponding entities. However, lookup tables are labor-intensive to create and maintain because they are manual and static. For instance, keyword tagging “Doritos Locos Taco” to Doritos’ parent company PepsiCo Frito-Lay seems like a fine mapping. However, it would be a mistake because “Doritos Locos Taco,” is actually a menu item at Taco Bell owned by YUM brands, not by PepsiCo.

Static (off-line) entity mappings are impractical in managing the many concurrently changing relationships between brands, products and parent companies. Practical entity mapping solutions need to be fully automated and scalable. Existing mappings require ongoing and systematic revisions to stay up to date. Automatically adjusting the mapping data is needed to account for the thousands of new products and brands which are added, modified and deleted daily. Efficient and scalable mapping algorithms are related to search engine technology in that they need to deliver structured, constantly changing results given unstructured input.

Due to the complexity of the problem, there are no public domain software solutions available and many companies opt to build in-house ticker tagging systems. Recently, commercial ticker tagging systems such as AltDG which use online machine learning approaches are showing great promise in delivering fast and scalable dynamic entity mapping technologies (Ekster 2018). AltDG’s online mapping engine correctly tags “Doritos Locos Taco” to Taco Bell (YUM) from the example above.

3.2 Panel Stabilization

A panel refers to a multi-dimensional dataset involving observations of each panel member over time. Typical panel members from alt-datasets include the collection of users, stores or business entities. In practical applications, panel constituents may change over time as some are added to the panel, whereas others are dropped. For instance, airline booking data is collected at the user level from travel sites such as kayak.com (Kolanovic and Smith 2019). As Kayak’s popularity has increased, so has its number of users. Therefore, the history of observations for each panel member will start and end on different dates. Regularly, at least one panel member will be unobserved at every period, resulting in an unbalanced panel.

Many users of alt-data prefer to work with a balanced panel, where at each point in time, there is an observation for every panel member. The process of turning a panel that is imbalanced to one that is balanced is called panel stabilization. Panel stabilization is related to the missing data problem.

In the case of airline booking data from Kayak, a naive but often used technique to stabilize the panel is to fill the missing entries with the average number of bookings per user at each time point. This approach disregards cross-sectional variation. A more appropriate method is to first reframe the panel stabilization problem as a missing data problem. A panel member's incomplete records are a form of data missing not at random (MNAR) where the missingness is related to both observed and unobserved values. Consequently, MNAR can then be addressed with various imputation algorithms, such as multiple imputation (MI) (Li, Stuart, and Allison 2015).

In the alt-data community, there is a debate about whether imputation is an acceptable method to stabilize a dataset. Imputing the "missing" data at the data's finest level of detail will make downstream modeling more manageable since it will appear as if there is no data missingness at all. However, the accuracy of imputation procedures relies on its modeling assumptions, which, if not valid, can lead to erroneous downstream conclusions. If possible, alt-data models should be built to handle imbalanced panels directly, thereby avoiding biases and errors introduced from imputation. If one chooses to impute any data, it is essential to indicate which data were original and which were imputed in the final dataset.

3.3 Debiasing

In statistics, a dataset is a sample drawn from the population of items or events of interest for a scientific question or experiment. Unless the sample is chosen randomly and entirely by chance, the resulting dataset will exhibit sampling bias. For any data-driven analysis to be meaningful, biases in the dataset need to be quantified and addressed. Without addressing biases, any statistical inference will be distorted or completely wrong.

Alt-datasets are seldom bias-free. An analyst must assess how representative a dataset is of the larger population of interest. Unfortunately, most biases are not describable through demographic, socio-economic and spatial characteristics such as age, gender or geographical location. For example, consider bank data sourced from consumers who prefer a particular bank because the branch manager's German accent is appealing. Demographics cannot do a good job of pinpointing this group. What type of person likes German accents anyway? An answer to this question is, "the type of person who's in your data." Of course, it would be convenient if demographics accurately isolated biases. Yet most often, the biases are not easy to identify because data is often incomplete and lacking the necessary attributes to make proper adjustments. In terms of sample size, alt-datasets are akin to polling data; they both sample a small minority of the total population, yet are used to infer the rest. Similarly, in both polling and alt-data, removing sample biases by weighting and resampling plays a large role in achieving reliable inferences and predictions.

4 On the Value of Alternative Datasets

The seemingly innocuous question of "how much is a dataset worth?" can quickly mushroom into a web of views as vast as the investment management field itself. In the discussion of the dataset valuation below, we will limit ourselves to traditional and quantitative non-high-frequency fundamental equity investing.

The purpose of alt-data valuation is: (1) to assess whether a dataset can help forecast security returns, and (2) to assess whether a dataset can help enhance other forecasting models of security returns. The purpose is not to build a full-fledged investment strategy, but rather to evaluate if the dataset should be purchased for further analysis. As such, an evaluation methodology needs to strike a balance between accuracy and cost.

While at the time of this writing, there is no known theoretical basis to alt-data valuation, practitioners are relying on two fundamental methods of evaluation:

1. Determine whether there is a significant correlation between the data or transformed versions thereof, and company operating metrics including revenues, same-store sales, etc.
2. Determine whether there is a significant correlation between the data or transformed versions thereof and security returns.

To isolate the impact of a new dataset on security returns, it is essential to first remove other well-known factors of covariation (Yi-Ou Li, Adali, and Calhoun 2007). Let us consider a new factor constructed from an alt-dataset. We can measure the impact of that factor by computing its partial correlation with security returns after controlling for other systematic pricing and risk factors. A straightforward way to operationalize this on a larger scale is residualizing security returns with respect to one or several of the many standard risk models available.

Perhaps somewhat surprisingly, the first method is the most commonly used approach in the investment management industry today. Although the second approach, which analyzes comovement or correlation of the data directly with security prices, appears to be more direct, it is not common practice for alt-data in fundamental investing.

Due to the low correlation with the residuals from standard security pricing and risk models, practitioners will most often use alt-data to model operating metrics of companies, not security prices. This is because modeling to operating metrics such as revenues has a higher signal-to-noise ratio and can be observed with a higher degree of confidence than modeling to security prices directly.

4.1 The Golden Triangle Event Study Methodology

Many alt-datasets have short series of historical data. It is not unusual that data analysts work with datasets with a history of only a few years (Denev and Amen 2020). For example, with three years of history, we only have twelve company reported quarterly data points for each panel member, making it challenging to build company-level models that forecast various quarterly operational metrics.

Therefore, instead of the correlation analysis discussed above, practitioners' resort to other methods to assess the value of a dataset. The Golden Triangle event study is a methodology which incorporates three steps:

1. Identify significant changes in the dataset. A change is referred to as a "data event" or "catalyst."
2. Identify associated real-world events from public information sources, including newswires, third-party data or company reports, that provide support for each data event.

3. Determine changes in the return of a tradable security around the time of each data event.

Exhibit 1 depicts the three steps of the Golden Triangle. We refer to steps two and three as the public information and market reaction tests, respectively. The public information test is qualitative in nature, using other data sources to find support for the events. The market reaction test measures the effect of the event on security, industry or market returns. If we can obtain positive confirmation for each data event from both tests, this indicates there may be some relationship between the data and returns of related financial securities. In addition, if we can establish temporal precedence of the three steps above, e.g.

- the data event from the dataset precedes a company news release, and
- the news release precedes a change in the stock return of the company;

then there is support for the dataset having predictive value to the investor.

If we can relate the dataset with security returns only (i.e. (1) and (3), but not (2)), then we cannot support the link between the dataset and real-world events, which significantly reduces our confidence in the value of the dataset. Similarly, a connection solely between the alt-data and company operating metrics (i.e. (1) and (2), but not (3)) suggests possible interdependence between the two, but also indicates that there is no market reaction and thus likely no investment potential.

The Golden Triangle is one specific tool for assessing the value of a dataset. However, finding multiple Golden Triangles for numerous unrelated events in the same dataset would provide further support for a dataset's value. A downside of the Golden Triangle methodology is that there are no automated tools for searching for the data - event - price movement trifecta, so each triangle can be time-consuming to establish. Experienced alt-data professionals are often needed to perform the analysis; thus, compiling golden triangles can be a laborious, high-cost endeavor, but can be rewarding if implemented correctly.

4.2 Report Cards

How valuable a particular dataset is to a buy-side manager depends on several factors, including:

- the scope, quality and representativeness of the data,
- how the data can contribute to the investment process, and
- whether the data is exclusive or more broadly known.

Since 2014, alt-data analysts have been using report cards, such as the one shown in Exhibit 2, to assess the viability of a dataset in an investment context. We provide a key for this report card in Appendix A. A report card may describe the main attributes of the dataset, characteristics of the data vendor, and its related ecosystem. For instance, dataset attributes may include company coverage, scope, representativeness and applicability. Information about the vendor may include items like data collection and distribution practices, as well as the quality of an entity's internal data-related operations. Attributes used to assess a dataset's ecosystem may include the following: (1) the awareness and pervasiveness of the data products' usage by institutional investors, (2) artificial caps on the number of funds with access to the data (scarcity), and (3) the duration of availability in the marketplace. Often an investment firm will develop its own report card with a unique system for scoring data attributes. Most report cards are qualitative, thus requiring less time-consuming analysis than would a quantitative deep-dive. Their advantage is in providing a big-picture summary of the data and its potential use cases, which can aid in quickly comparing and narrowing down a collection of datasets. However, due to the lack of quantitative data validation means report cards should not be the only tool used to assess the value of a dataset.

4.3 Relationship Between a Dataset's Structure and Investment Performance

Alt-data are available in a wide range of formats, from raw and unprocessed to aggregated and structured. On one end of the spectrum, examples of unprocessed data include free-form text from web discussion boards, raw images of municipal construction permits and images of cars in parking lots. On the other end of the spectrum, there are aggregated and structured datasets such as total daily website visitors, the number of monthly construction permits per construction company and daily vehicle counts at specific points of interest.

A raw and unstructured dataset can be different than a corresponding aggregated and structured dataset in a number of ways. For example, (1) the unstructured data's information content is at least as comprehensive, and often more comprehensive, as that of data aggregated from the same source; (2) frequently, it can be transformed and aggregated in various ways that are complementary to the already aggregated data. Consequently, a raw dataset may have many more potential investment applications. In addition, it is likely that competing investment teams will discover different ways to process and interpret the data, even if starting from the same raw dataset.

Naturally, there is a trade-off between the cost to analyze and explore alt-data and the uniqueness of any investment insights resulting from those efforts. Depending on the dataset in question, the costs associated with the development of an investment strategy and/or trading signal can be high. Funds find themselves having to decide between two fundamentally different approaches:

1. Obtain processed, aggregated and/or structured alt-data products from intermediaries, or

2. Obtain raw and/or unstructured datasets and develop in-house capabilities to process, analyze and perform research with the data.

Using alt-data without committing to large analysis expenses is often cited as a benefit of purchasing ready-made products. In reality, few funds profit from alt-data without investing into their own in-house capabilities (Ekster 2015a).

5 Case Study: Healthcare Purchasing Data

In this section, we apply the concepts discussed in this article to a large real-world alt-dataset that comprises medical device purchasing information from national medical facilities. Such datasets are provided and processed by vendors including MedMine, 7park, AltDG, HealthVerity, IHS Markit and GuidePoint. For investors, this data provides a unique insight into the medical sector without having to grapple with privacy issues as device purchases are presented at the facility level, not at the level of the individual patient. In particular, the data's value for investment professionals is the visibility it offers into the revenues of major medical device manufacturers. For instance, it provides an analytical deep-dive into daily product sales, such as the popularity of a hotly anticipated new MRI machine model and its post-release sales volumes at medical institutions.

5.1 Dataset Overview

We obtained samples of daily U.S. medical-purchasing activity by healthcare facilities including hospitals, ambulatory surgery centers and physician offices. These facilities' inventory management systems (IMS) contain useful information about purchasing activity, such as the medical device manufacturers' identities, their volumes and pricing. The alt-data product we use is updated and delivered on a daily basis. Our sample covers purchasing activity from 2015 through 2017. The data is sourced from 778 medical facilities and, as is often the case with alt-data, the sample representativeness is determined mostly by the mechanisms underlying the collection process. In this case, hospitals are in the sample if they happen to use an IMS for record-keeping. As we will see below, the dataset presents a number of modeling challenges such as IMS user-input errors, outliers, unstructured item descriptions and missing information. Exhibit 3 provides some summary metrics of our data.

5.2 Report Card

We summarize key opportunities and costs of the dataset using the report card from Section 4.2 in Exhibit 4 below. In this case, we did not perform a Golden Triangle case study. Using the report card analysis, we found that the dataset covers the medical device manufacturing sector thoroughly, with the largest 20 publicly-traded medical device companies represented including Abiomed (ABMD), Coloplast (COLO), ConvaTec Group (CTEC), Edwards Lifesciences (EW), Globus Medical (GMED), Intuitive Surgical (ISRG), K2M Group Holdings (KTWO), Medtronic (MDT), Penumbra (PEN), Wright Medical Group (WMGI) and Zimmer Biomet (ZBH). Furthermore, the dataset's granularity at the level of individual products, facilities on a daily basis creates a great opportunity for various forms of high-dimensional analysis. Thus, it has the potential to provide a number of diverse investment themes and profit opportunities.

However, we note that while the U.S. hosts over 30,000 medical facilities, the dataset contains only 778, which represents less than 2.6% of the total. Qualitatively, given that the facilities appear in the data due to their selection of an internal IMS system, this 2.6% sample is not a random but a biased sample (Weinick, Bristol, and DesRoches 2009). This relatively low sampling rate, single-sector coverage, and known bias limits the dataset's potential usage for systematic investors who value big sample sizes and large sector-coverage rates. We discuss below how some of these limitations can be addressed by properly pre-processing the data.

5.3 Processing the Data

The unprocessed healthcare medical data is recorded at the device / device part level for each facility and updated on a daily basis, see Exhibit 5 for an example of the data. Key fields in the data include:

- UID: a unique numerical ID for each medical facility,
- Date: when the device was sold or used,
- Item Description: the description of the medical device used, and
- Price: the purchasing price of the device.

Inspecting the data at the most granular level, we found that it contains small errors and gaps where data is missing, as is common in many alt-datasets. This prevents meaningful direct aggregation and analysis without first preprocessing the data.

There are several other issues that needs to be addressed during the preprocessing of the data. Below we discuss the following preprocessing steps:

- Entity tagging,
- Outlier detection, and
- Missing data imputation (panel stabilization).

5.3.1 Limits of Direct Aggregation

We found that simple summation and scaling of total spending was not predictive of company-reported operating metrics such as revenues without first addressing outliers and missing data. The difference (error) between simply summing the spending reported in the dataset and the company-reported revenues was as high as 1200%, with an MAE of 88% as per Exhibit 7. Such large errors in unprocessed data are often seen by alt-data practitioners. For comparison, traditional equity research analysts, not using alt-data, are able to forecast revenues for medical device companies such as COLO, EW and ISRG with a MAE of 7.5% over the 2015 - 2017 time period.

5.3.2 Entity Tagging

Like many other alt-datasets, this dataset does not contain standardized company names or other structured entity identifiers. Hence, to identify companies that manufactured a given product, we rely on a field describing the medical component, which is expressed as unstructured text. Entity identification is challenging for this dataset as many invoices are created manually at the medical facilities, often with staff members using cryptic short-hand or abbreviations, leading to a lack of consistency across facilities. For example, consider the text string “cath dlvr enveo coreviv.” After searching through product catalogues of various medical manufacturers we conclude that this is the EnVeo R Catheter CoreValve system manufactured by Medtronic (MDT). Manually performing searches of this kind is too time-consuming to be practical in large scale applications. Creating a static map for all products and all possible abbreviations is not feasible either because of continuous product additions to the data. Thus, even if we created such a mapping table, it would require continuous maintenance. Only an automated and dynamic system is a plausible solution for the large-scale datasets such as the one in this case study. Therefore, we applied an automatic merchant mapper to tag each transaction with its associated manufacturer, revenue segment and product category (Ekster 2019). Exhibit 8 shows an example of raw data with tagged transactions.

5.3.3 Outlier Detection and Resolution

Not surprisingly, as depicted in Exhibit 9, our dataset contains transaction-level outliers including price errors and quantity miscounts.

To address the outliers, we first clustered transactions based on manufacturer, product, product category, purchasing facility and time. Then, as prices in each cluster appeared multimodal, we chose to fit a multimodal distribution to each. Then we used this data model to identify and winsorize outliers. By addressing outliers at a device level rather than at the facility level allowed us to preserve the complex pricing dynamics which are present in actual transactions between medical facilities and medical device manufacturers.

We define a facility-month as the scalar product of all transaction volumes and their respective prices per facility per month. Any outliers at the facility-month level indicate some system error and reduce the accuracy of the data. By way of example, consider Exhibit 10 where we observe that some facilities contain months which are far below or above the expected value for that facility-month. We decided to remove such data points using a local outlier factor (LOF) approach and impute them using the methods described below in section 5.3.4 below (Sugidamayatno and Lelono 2019).

5.3.4 Panel Stabilization and Imputation

None of the medical facilities of this dataset provide complete reporting at all times. Facility-level data has time gaps during which no data is reported. These gaps occur because each facility does not continuously report its data throughout the panel's entire history. For instance, the panel constituents may exit the panel early, or they may enter the panel somewhere in the middle. This causes an incomplete panel that may have to be addressed with imputation (Honaker, King, and Blackwell 2011).

Based on the needs of downstream de-biasing and revenue prediction models, we used a multivariate time-series imputation technique described and implemented in the Amelia2 program (Honaker, King, and Blackwell 2011). We applied the imputation on a facility-product-month level. The configuration setup parameters for Amelia2 included a square root distribution transform, prediction bounds and a choice of ridge prior. In addition, we ran a second imputation at the monthly-facility-total-sales level in order to validate the previously computed, facility-product-month level imputation.

5.3.5 Imputation Error Estimation

To estimate the error of imputation, we used a customized leave-one-out (LOO) cross validation. The customization modified the LOO algorithm's removal of a uniform random datapoint. Instead, we removed existing data points in such a way that we emulate the missing data's proximity to existing data points. Because many data points are clustered together (see Exhibit 11), if we delete a random datapoint for cross validation, we find that the average proximity of the deleted datapoint is about 1.6 months from the nearest existing datapoint. However, the missing data has an average temporal proximity of 3.8 months to existing (non-outlier) datapoints. Thus, we modified the LOO random selection criteria such that we match the average proximity of the real missing data. The customized LOO cross-validation procedure resulted in an absolute error estimate of 6.1% for imputed facility-months (see Exhibit 12).

After imputation, hospital-level data shows that individual characteristics of each hospital's device-sales mix is preserved (see Exhibit 13).

5.4 Modeling and Investment Insights

After tagging and imputing the data, we can query the dataset for metrics relevant to investment performance such as year-over-year revenue growth rates. We were able to forecast segment revenues with an MAE of 2.6%, a significant improvement over the 88.0% mean error from the unprocessed data (see Exhibit 14 below for examples of segment revenue prediction for COLO, EW and ISRG). Since we modeled data at a device-facility-month level, we are able to analyze the revenue totals by device-sales mix. This form of analysis is valuable in a fundamental research context. Our model provides insights into the sales performance of new products, a metric which is closely followed by financial market participants and impacts related stock returns. At an aggregate level, revenue surprises for medical device manufacturers impact market performance and are relevant for making investment decisions (Jegadeesh and Livnat 2006, 147-171).

5.5 Discussion

In this case study, we (1) review the process of transforming unstructured transaction descriptions into product names, categories, and manufacturers, (2) identify and resolve transaction and facility-level outliers, (3) impute missing data, and (4) develop a model for revenues and sales performance at a granular level.

This case study is an illustration of how alt-data is used in the industry today to generate new investment insights. While each dataset is different, many of the issues we discussed throughout the workflow are common.

An investment team looking to integrate alt-datasets into their investment decision process have to choose whether to work with raw datasets or aggregated products. In our experience, the key to succeeding with alt-data is (1) leveraging the greater information content present in high-dimensional raw datasets, and (2) carefully addressing underlying data issues such as unstructured transactions, biases, entity tagging and entity churn. A properly processed and modeled dataset has the potential to provide more valuable insights and opportunities to investors than aggregated products.

6 Trends in the Alternative Data Space

6.1 Cost-Benefit Analysis of Intermediaries vs. Originators

The discovery of a dataset's potential is predicated on transforming raw data into structured data. This poses a dilemma for funds interested in acquiring a large and costly dataset because its value is unknown a priori. In addition, its valuation frequently requires expensive and labor-intensive data processing and analysis. This costly information acquisition has been standing in the way of alt-data drastically transforming the investment management landscape.

In an asymmetrically-informed client-base environment, even low-quality data providers prosper and high-quality ones are punished as is in the classical lemons problem (von Weizsäcker 2016, 91-96). However, new emerging automatic tools and analytics will bring about a wave of changes within the industry. Specifically, these developments will (1) likely separate out and possibly eliminate providers supplying questionable products, and (2) elevate firms with high-quality data products. This may create a winner-take-all market, similar to that of the online software sector, and result in concentrated consolidations. Needless to say, automatic processing and valuation of the data is not automatic alpha generation. Human analysts will still need to carefully analyze the datasets in order to monetize them for investment purposes.

7 Conclusions

In this article we have addressed a number of issues pertinent to the application of alt-data in investment management, including (1) what alt-data is and how it is used, (2) the commercial ecosystem for alt-data, (3) key challenges that arise when using and working with alt-data, (4) statistical and machine learning techniques for processing alt-data, (5) valuation methodologies for alt-datasets, and (6) the relationship between the underlying data structure and its value in producing investment insights and predictions. In addition, we presented a case study using a large healthcare alt-dataset for forecasting revenues of publicly traded medical device companies. As part of the case study, we discussed the considerations that went into designing the data processing pipeline and the downstream analytics. Our model delivered an improvement of revenue prediction accuracy from an 88% MAE to a 2.6% MAE.

Finally, we discussed some of the trends in the continuously evolving alt-data sector. We emphasized the changes the new emerging automatic tools and analytics will bring to the alt-data space. In particular, they will provide great growth opportunities for the industry, but will likely result in a dramatic industry consolidation due to the high costs of implementation.

References

Deloitte. "Maximising Data Value." . https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-maximising-data-value-a-vendors-perspective.pdf.

Denev, Alexander and Saeed Amen. *The Book of Alternative Data*. Newark: John Wiley & Sons, 2020.

Du, Kai, Steven Huddart, Lingzhou Xue, and Yifan Zhang. "Using a Hidden Markov Model to Measure Earnings Quality." *Journal of Accounting & Economics* 69, no. 2-3 (Apr, 2020): Article 101281.

Eagle Alpha. *Alternative Data Use Cases Edition 5*, 2019.

Ekster, Gene. "Alternative Data Group Launches a New API Platform for Wrangling Alternative Data." *PR Newswire,* Sep 27, 2018.

Ekster, Gene. "Driving Investment Performance with Alternative Data," 2015. http://www.integrity-research.com/driving-investment-performance-with-alternative-data.

Ekster, Gene. *Finding and Using Unique Datasets by Hedge Funds*, 2014. https://www.hedgeweek.com/2014/11/03/212370/finding-and-using-unique-datasets-hedge-funds.

Ekster, Gene. *Revenue Surprise and Equity Performance Analysis*, AltDG, 2015.

Ekster, Gene. *Solving the Ticker / Merchant Mapping Problem in Alternative Data*, 2019. https://www.Linkedin.Com/Pulse/Solving-Ticker-Merchant-Mapping-Problem-Alternative-Data-Gene-Ekster/?articleId=6534115284498817024.

Helm, Burt. "Credit Card Companies are Tracking Shoppers Like Never Before: Inside the Next Phase of Surveillance Capitalism." *Fast Company*, 2020.

Honaker, James, Gary King, and Matthew Blackwell. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45, no. 7 (2011): pp. 1-47. doi:10.18637/jss.v045.i07.

Hope, Bradley. "Provider of Personal Finance Tools Tracks Bank Cards, Sells Data to Investors; Yodlee's Side Business shows Escalation in Race among Investors Trying to Turn Data into

Profits." *The Wall Street Journal. Eastern Edition,* Aug 7, 2015. https://search.proquest.com/docview/1701997147.

Jagtiani, Julapa and Catharine Lemieux. "The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform." *Financial Management* 48, no. 4 (2019): pp. 1009-1029. doi:10.1111/fima.12295. https://onlinelibrary.wiley.com/doi/abs/10.1111/fima.12295.

Jaquez, Casey. "Mastercard SpendingPulse: Estimated $53 Billion in Additional U.S. E-Commerce Sales as Pandemic Drives Consumers Online in April and May." *Contify Banking News* (Jun 10, 2020).

Jegadeesh, Narasimhan and Joshua Livnat. "Revenue Surprises and Stock Returns." *Journal of Accounting and Economics* 41, no. 1 (2006): 147-171. doi:10.1016/j.jacceco.2005.10.003. http://econpapers.repec.org/article/eeejaecon/v_3a41_3ay_3a2006_3ai_3a1-2_3ap_3a147-171.htm.

Kolanovic, Marko and Robert Smith. *Big Data and AI Strategies* J.P. Morgan, 2019.

Li, Peng, Elizabeth A. Stuart, and David B. Allison. "Multiple Imputation: A Flexible Tool for Handling Missing Data." *JAMA : The Journal of the American Medical Association* 314, no. 18 (Nov 10, 2015).

Monk, Ashbey, Marcel Prins, and Dane Rook. "Rethinking Alternative Data in Institutional Investment." *The Journal of Financial Data Science* (Feb 1, 2019). doi:10.3905/jfds.2019.1.1.014.

Sugidamayatno, Silvano and Danang Lelono. "Outlier Detection Credit Card Transactions Using Local Outlier Factor Algorithm (LOF)." *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 13, no. 4 (Oct 31, 2019): pp. 409-420. doi:10.22146/ijccs.46561.

von Weizsäcker, Carl Christian. "Akerlof, George A. and Shiller, Robert J.: Phishing for Phools: The Economics of Manipulation and Deception." *Journal of Economics (Vienna, Austria)* 118, no. 1 (May, 2016): pp. 91-96. doi:10.1007/s00712-016-0471-y. https://search.proquest.com/docview/1780742841.

Weinick, Robin M., Steffanie J. Bristol, and Catherine M. DesRoches. "Urgent Care Centers in the U.S.: Findings from a National Survey." *BMC Health Services Research* 9, no. 1 (May 15, 2009): pp. 79. doi:10.1186/1472-6963-9-79.

Yi-Ou Li, T. Adali, and V. D. Calhoun. "A Multivariate Model for Comparison of Two Datasets and its Application to FMRI Analysis." In 2007 *IEEE Workshop on Machine Learning for Signal Processing*, pp. 217-222. IEEE, 2007.

Appendix A – Report Card Key

Coverage

- Asset Classes: List the asset classes covered (e.g. equities, sovereign bonds, commodities).
- Sectors: List the number of sectors under coverage, such as retail, energy, etc.
- Liquidity: A 1-5 estimate of the overall liquidity of the underlying security, where 5 is the most liquid.

Scope

- History: Amount of history available for the dataset.
- Temporal granularity delivery frequency and lag. Time-frequency can be reported as quarterly, monthly weekly or daily, etc.

Representativeness

How representative is the dataset of the population it is aiming to measure? A 1-5 estimate or exact metric where applicable, where 5 is the highest.

Applicability

- Investment Ideas: The number of varieties of investment ideas that the dataset can support.
- KPIs/GAAP: A list or number of GAAP- level predictions that are possible to obtain from the dataset.
- Surprise Sensitivity: If the dataset is able to predict revenues, rate how sensitive are the covered companies to revenue surprises.

Data Collection / Process Quality

Expertise of the Data Originator in creating, packaging, delivering, and supporting the product. A 1-5 estimate, where 5 is the most experienced.

Structure

Is the data mapped to tickers, attributes or other structured information? Yes/no.

Consistency

How stable and reliable is the data product over time. A 1-5 estimate, where 5 is the most consistent.

Scarcity

- Market Awareness: Measures the uniqueness of the dataset. List how many other investment organizations are aware of the product.
- User Base: List how many investment organizations are using the product.
- Time in Market: List how long the dataset has been commercially available for.

Direct Costs

The monetary price of acquiring and/or subscribing to an ongoing feed of the data. List the costs.

Costs Risk

The estimated cost of the risks inherent in the data feed including compliance, counterparty and PR/headline.

R&D

Estimated costs of the labor and computing resources needed to research, validate and develop an investment product based on the dataset. List the costs / resources needed.

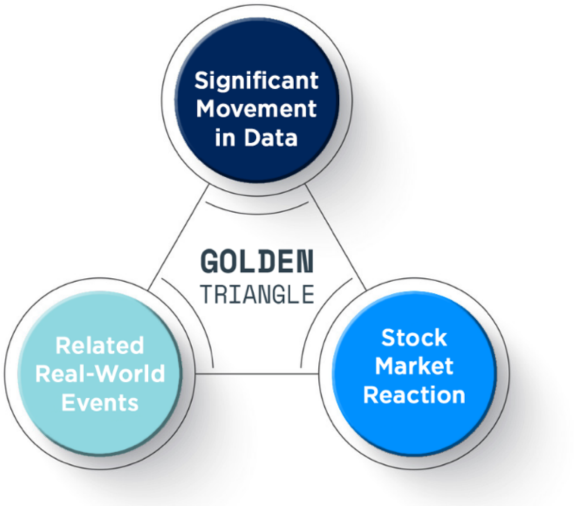
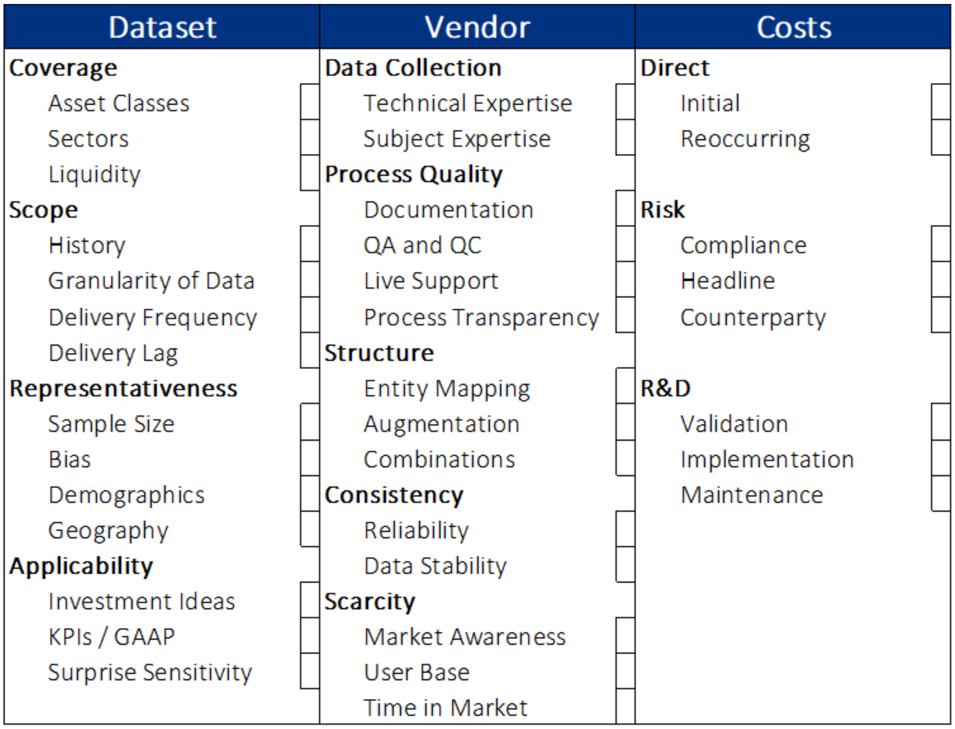


Exhibit 1. The Golden Triangle event study methodology. A dataset evaluation technique that compares the data with real-world news events and stock price reactions.

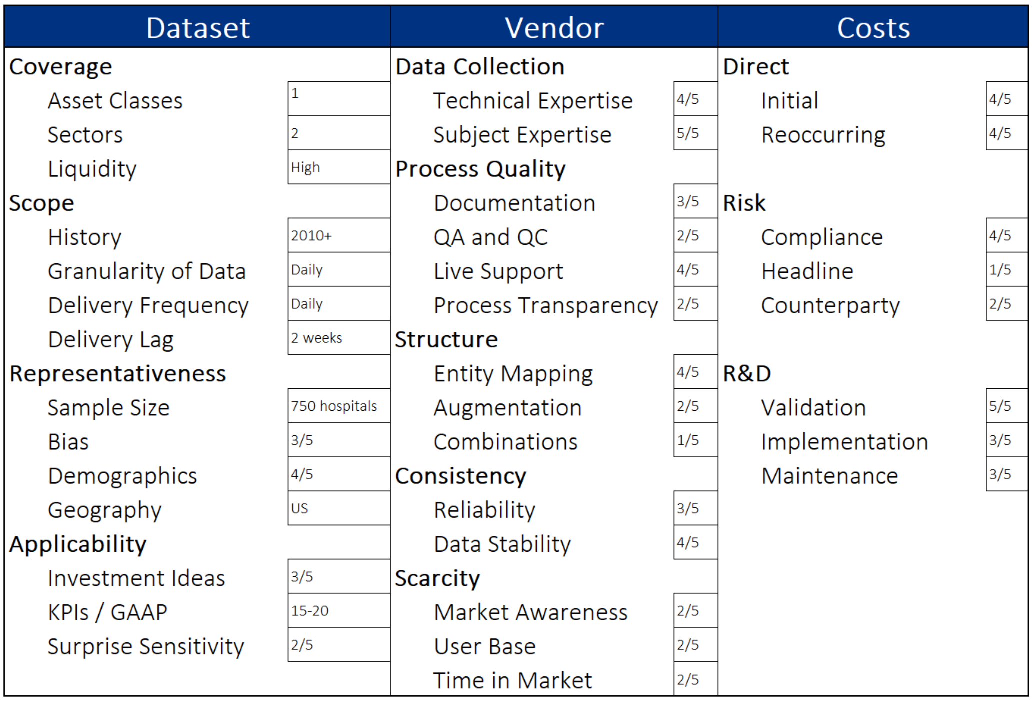


Dataset		Vendor		Costs	
Coverage		**Data Collection**		**Direct**	
Asset Classes	□	Technical Expertise	□	Initial	□
Sectors	□	Subject Expertise	□	Reoccurring	□
Liquidity	□	**Process Quality**			
Scope		Documentation	□	**Risk**	
History	□	QA and QC	□	Compliance	□
Granularity of Data	□	Live Support	□	Headline	□
Delivery Frequency	□	Process Transparency	□	Counterparty	□
Delivery Lag	□	**Structure**			
Representativeness		Entity Mapping	□	**R&D**	
Sample Size	□	Augmentation	□	Validation	□
Bias	□	Combinations	□	Implementation	□
Demographics	□	**Consistency**		Maintenance	□
Geography	□	Reliability	□		
Applicability		Data Stability	□		
Investment Ideas	□	**Scarcity**			
KPIs / GAAP	□	Market Awareness	□		
Surprise Sensitivity	□	User Base	□		
		Time in Market	□		

Exhibit 2. An alt-data report card which evaluates data with a qualitative questionnaire. Appendix A provides a key for this report card.

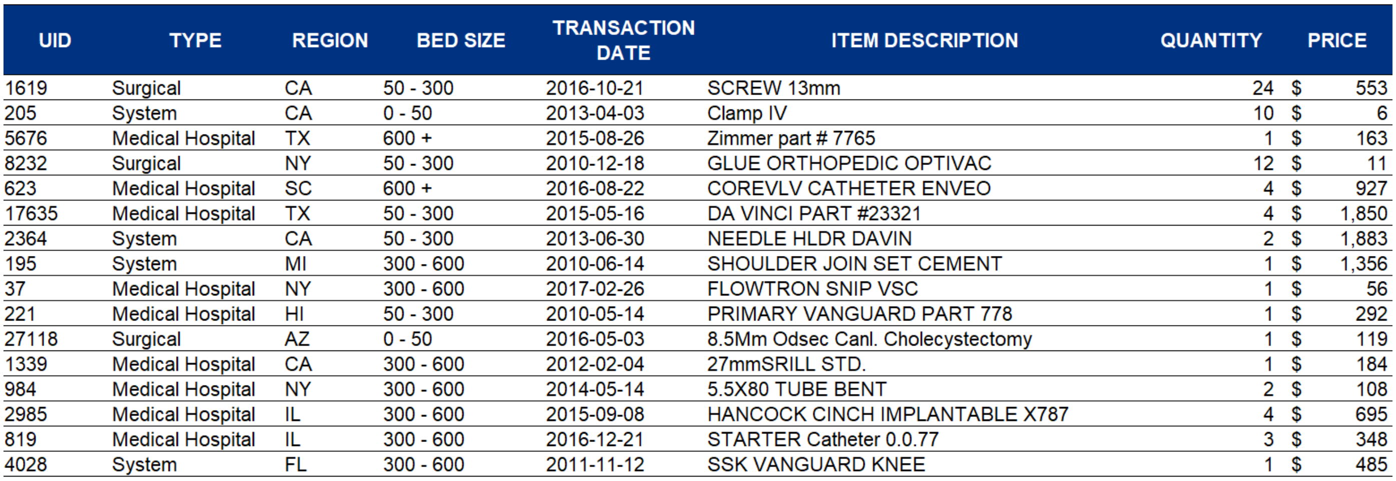
	Summary Metrics
Rows	213,390,788
Unique SKUs	67,563
Total Spending	$331M
Time Granularity	Daily
Geographical Coverage	U.S. State
Number of Facilities	778
History Start	January 2015

Exhibit 3. Healthcare case study dataset summary metrics.



Dataset		Vendor		Costs	
Coverage		Data Collection		Direct	
Asset Classes	1	Technical Expertise	4/5	Initial	4/5
Sectors	2	Subject Expertise	5/5	Reoccurring	4/5
Liquidity	High	Process Quality			
Scope		Documentation	3/5	Risk	
History	2010+	QA and QC	2/5	Compliance	4/5
Granularity of Data	Daily	Live Support	4/5	Headline	1/5
Delivery Frequency	Daily	Process Transparency	2/5	Counterparty	2/5
Delivery Lag	2 weeks	Structure			
Representativeness		Entity Mapping	4/5	R&D	
Sample Size	750 hospitals	Augmentation	2/5	Validation	5/5
Bias	3/5	Combinations	1/5	Implementation	3/5
Demographics	4/5	Consistency		Maintenance	3/5
Geography	US	Reliability	3/5		
Applicability		Data Stability	4/5		
Investment Ideas	3/5	Scarcity			
KPIs / GAAP	15-20	Market Awareness	2/5		
Surprise Sensitivity	2/5	User Base	2/5		
		Time in Market	2/5		

Exhibit 4. Report card for the healthcare dataset. Source: AltDG, https://developer.altdg.com.



UID	TYPE	REGION	BED SIZE	TRANSACTION DATE	ITEM DESCRIPTION	QUANTITY	PRICE
1619	Surgical	CA	50 - 300	2016-10-21	SCREW 13mm	24	$ 553
205	System	CA	0 - 50	2013-04-03	Clamp IV	10	$ 6
5676	Medical Hospital	TX	600 +	2015-08-26	Zimmer part # 7765	1	$ 163
8232	Surgical	NY	50 - 300	2010-12-18	GLUE ORTHOPEDIC OPTIVAC	12	$ 11
623	Medical Hospital	SC	600 +	2016-08-22	COREVLV CATHETER ENVEO	4	$ 927
17635	Medical Hospital	TX	50 - 300	2015-05-16	DA VINCI PART #23321	4	$ 1,850
2364	System	CA	50 - 300	2013-06-30	NEEDLE HLDR DAVIN	2	$ 1,883
195	System	MI	300 - 600	2010-06-14	SHOULDER JOIN SET CEMENT	1	$ 1,356
37	Medical Hospital	NY	300 - 600	2017-02-26	FLOWTRON SNIP VSC	1	$ 56
221	Medical Hospital	HI	50 - 300	2010-05-14	PRIMARY VANGUARD PART 778	1	$ 292
27118	Surgical	AZ	0 - 50	2016-05-03	8.5Mm Odsec Canl. Cholecystectomy	1	$ 119
1339	Medical Hospital	CA	300 - 600	2012-02-04	27mmSRILL STD.	1	$ 184
984	Medical Hospital	NY	300 - 600	2014-05-14	5.5X80 TUBE BENT	2	$ 108
2985	Medical Hospital	IL	300 - 600	2015-09-08	HANCOCK CINCH IMPLANTABLE X787	4	$ 695
819	Medical Hospital	IL	300 - 600	2016-12-21	STARTER Catheter 0.0.77	3	$ 348
4028	System	FL	300 - 600	2011-11-12	SSK VANGUARD KNEE	1	$ 485

Exhibit 5. Example of unprocessed device-level data.

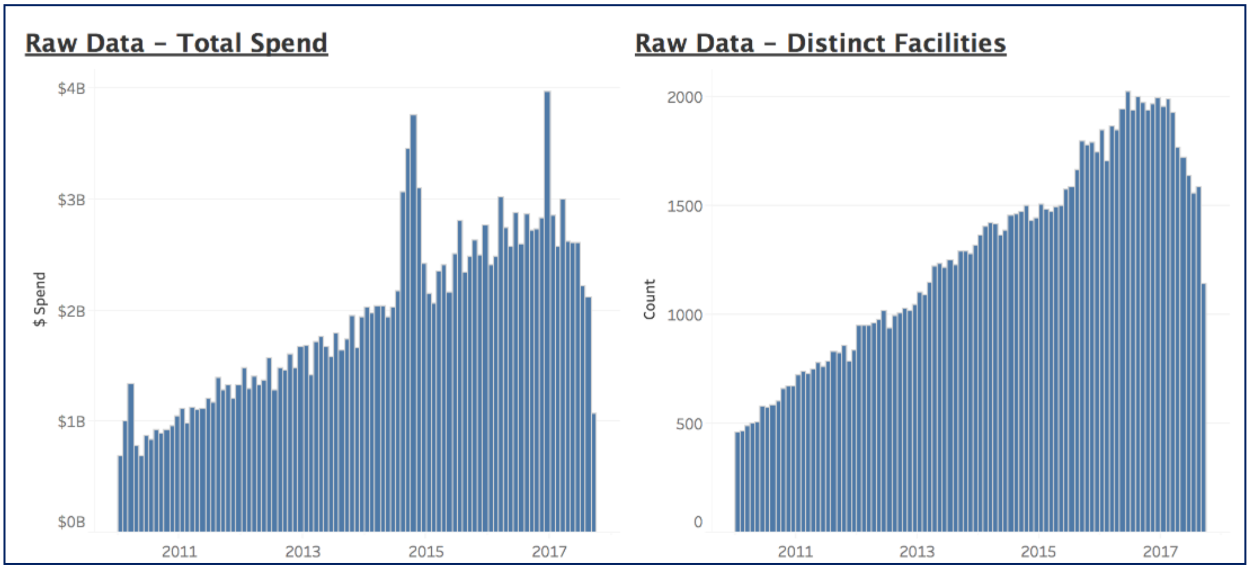


Exhibit 6. Transactions across all manufacturers and facilities in the raw data. Spending systematically increases over time as more facilities are included in the dataset. The drop-off in spending and facilities towards the end of the dataset is due to delays in reporting.

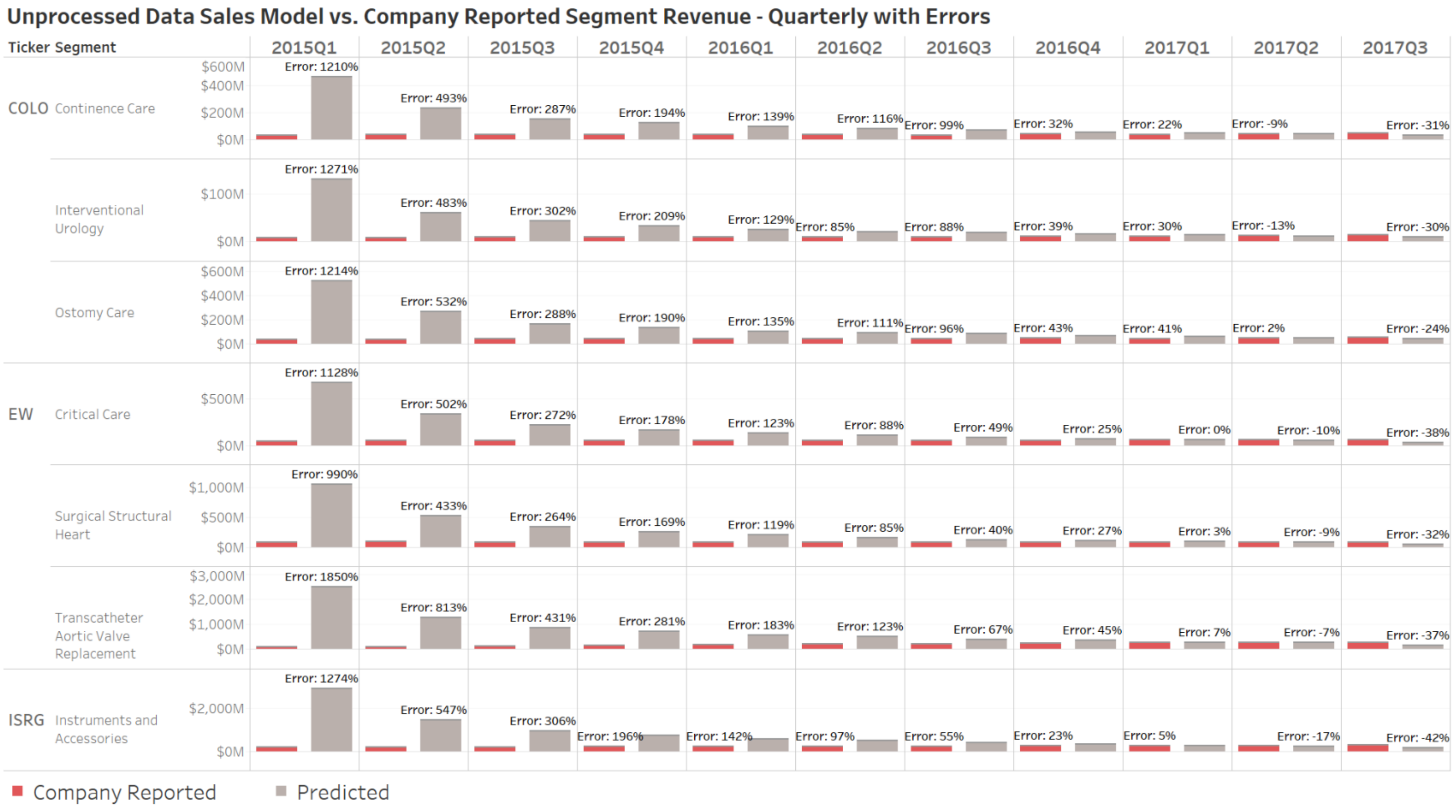
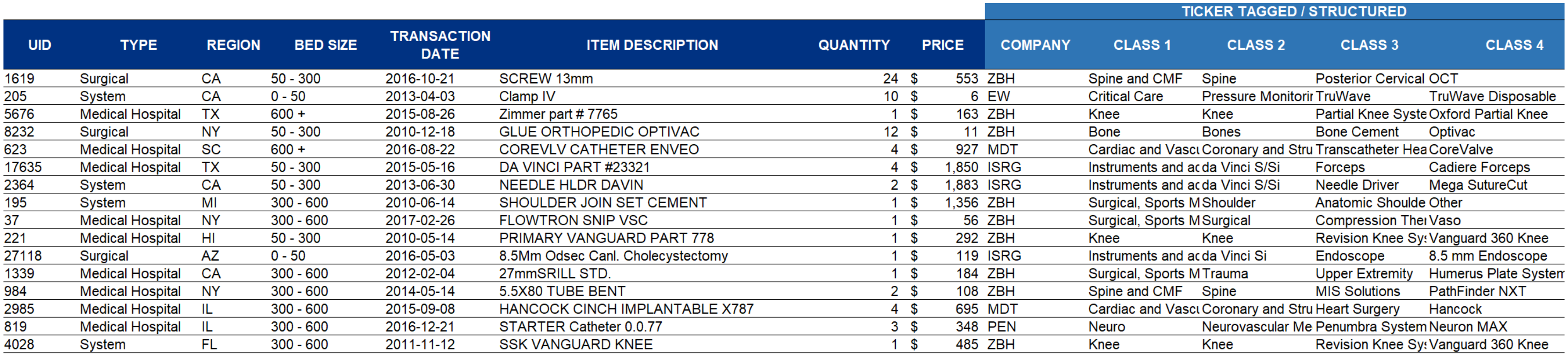
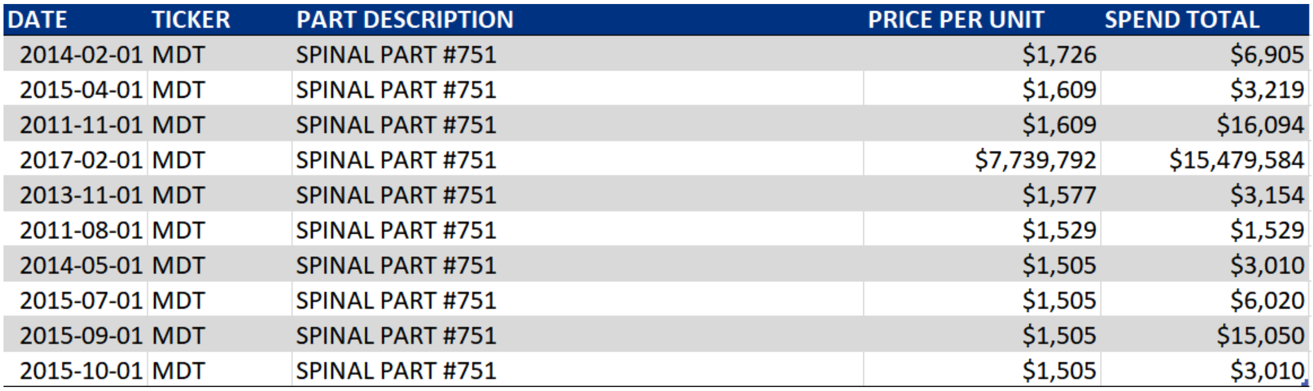


Exhibit 7. Company-reported segment revenues vs. forecasts using non-processed raw data. The substantial difference between the two create large errors.



								TICKER TAGGED / STRUCTURED				
UID	TYPE	REGION	BED SIZE	TRANSACTION DATE	ITEM DESCRIPTION	QUANTITY	PRICE	COMPANY	CLASS 1	CLASS 2	CLASS 3	CLASS 4
1619	Surgical	CA	50 - 300	2016-10-21	SCREW 13mm	24	$ 553	ZBH	Spine and CMF	Spine	Posterior Cervical	OCT
205	System	CA	0 - 50	2013-04-03	Clamp IV	10	$ 6	EW	Critical Care	Pressure Monitorir	TruWave	TruWave Disposable
5676	Medical Hospital	TX	600 +	2015-08-26	Zimmer part # 7765	1	$ 163	ZBH	Knee	Knee	Partial Knee Syste	Oxford Partial Knee
8232	Surgical	NY	50 - 300	2010-12-18	GLUE ORTHOPEDIC OPTIVAC	12	$ 11	ZBH	Bone	Bones	Bone Cement	Optivac
623	Medical Hospital	SC	600 +	2016-08-22	COREVLV CATHETER ENVEO	4	$ 927	MDT	Cardiac and Vascı	Coronary and Stru	Transcatheter Hea	CoreValve
17635	Medical Hospital	TX	50 - 300	2015-05-16	DA VINCI PART #23321	4	$ 1,850	ISRG	Instruments and ac	da Vinci S/Si	Forceps	Cadiere Forceps
2364	System	CA	50 - 300	2013-06-30	NEEDLE HLDR DAVIN	2	$ 1,883	ISRG	Instruments and ac	da Vinci S/Si	Needle Driver	Mega SutureCut
195	System	MI	300 - 600	2010-06-14	SHOULDER JOIN SET CEMENT	1	$ 1,356	ZBH	Surgical, Sports M	Shoulder	Anatomic Shoulde	Other
37	Medical Hospital	NY	300 - 600	2017-02-26	FLOWTRON SNIP VSC	1	$ 56	ZBH	Surgical, Sports M	Surgical	Compression Thei	Vaso
221	Medical Hospital	HI	50 - 300	2010-05-14	PRIMARY VANGUARD PART 778	1	$ 292	ZBH	Knee	Knee	Revision Knee Sy:	Vanguard 360 Knee
27118	Surgical	AZ	0 - 50	2016-05-03	8.5Mm Odsec Canl. Cholecystectomy	1	$ 119	ISRG	Instruments and ac	da Vinci Si	Endoscope	8.5 mm Endoscope
1339	Medical Hospital	CA	300 - 600	2012-02-04	27mmSRILL STD.	1	$ 184	ZBH	Surgical, Sports M	Trauma	Upper Extremity	Humerus Plate Systen
984	Medical Hospital	NY	300 - 600	2014-05-14	5.5X80 TUBE BENT	2	$ 108	ZBH	Spine and CMF	Spine	MIS Solutions	PathFinder NXT
2985	Medical Hospital	IL	300 - 600	2015-09-08	HANCOCK CINCH IMPLANTABLE X787	4	$ 695	MDT	Cardiac and Vascı	Coronary and Stru	Heart Surgery	Hancock
819	Medical Hospital	IL	300 - 600	2016-12-21	STARTER Catheter 0.0.77	3	$ 348	PEN	Neuro	Neurovascular Me	Penumbra System	Neuron MAX
4028	System	FL	300 - 600	2011-11-12	SSK VANGUARD KNEE	1	$ 485	ZBH	Knee	Knee	Revision Knee Sy:	Vanguard 360 Knee

Exhibit 8. Example of raw data (left side, in dark blue) and tagged transactions (right side, in light blue) using an automatic merchant mapper.



DATE	TICKER	PART DESCRIPTION	PRICE PER UNIT	SPEND TOTAL
2014-02-01	MDT	SPINAL PART #751	$1,726	$6,905
2015-04-01	MDT	SPINAL PART #751	$1,609	$3,219
2011-11-01	MDT	SPINAL PART #751	$1,609	$16,094
2017-02-01	MDT	SPINAL PART #751	$7,739,792	$15,479,584
2013-11-01	MDT	SPINAL PART #751	$1,577	$3,154
2011-08-01	MDT	SPINAL PART #751	$1,529	$1,529
2014-05-01	MDT	SPINAL PART #751	$1,505	$3,010
2015-07-01	MDT	SPINAL PART #751	$1,505	$6,020
2015-09-01	MDT	SPINAL PART #751	$1,505	$15,050
2015-10-01	MDT	SPINAL PART #751	$1,505	$3,010

Exhibit 9. Example of device-level price outliers.

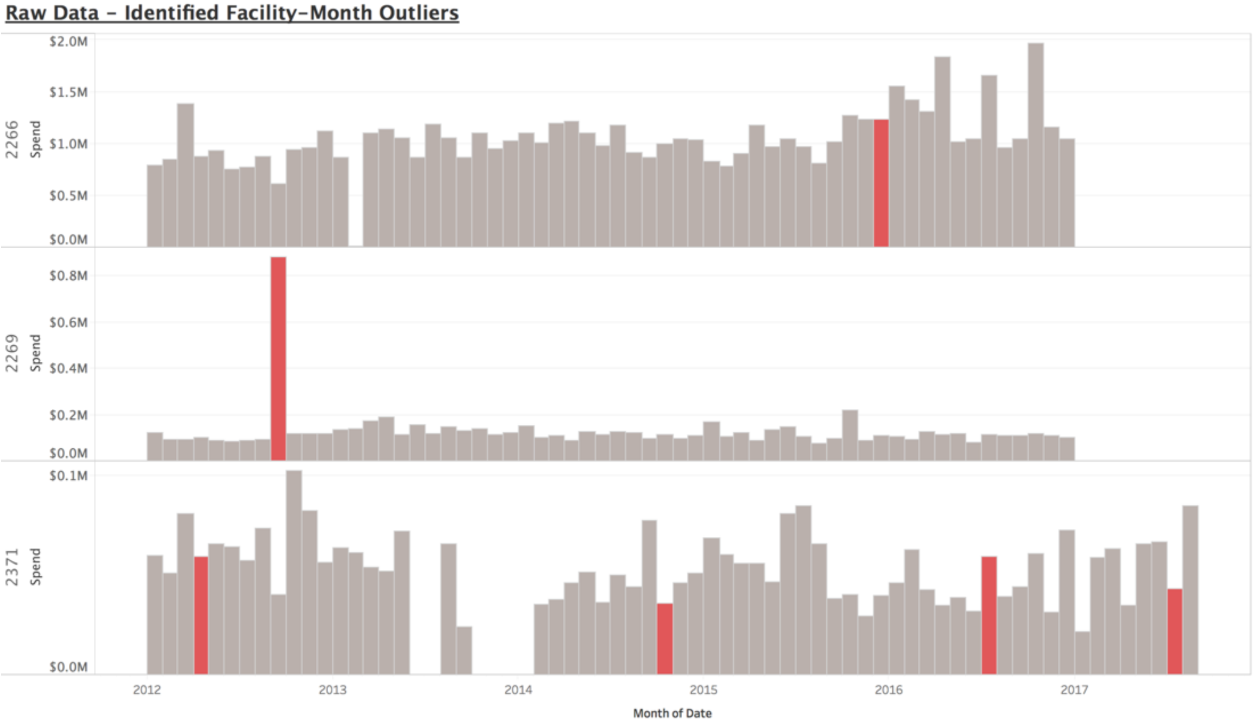


Exhibit 10. Example of identified facility-month outliers (in red).

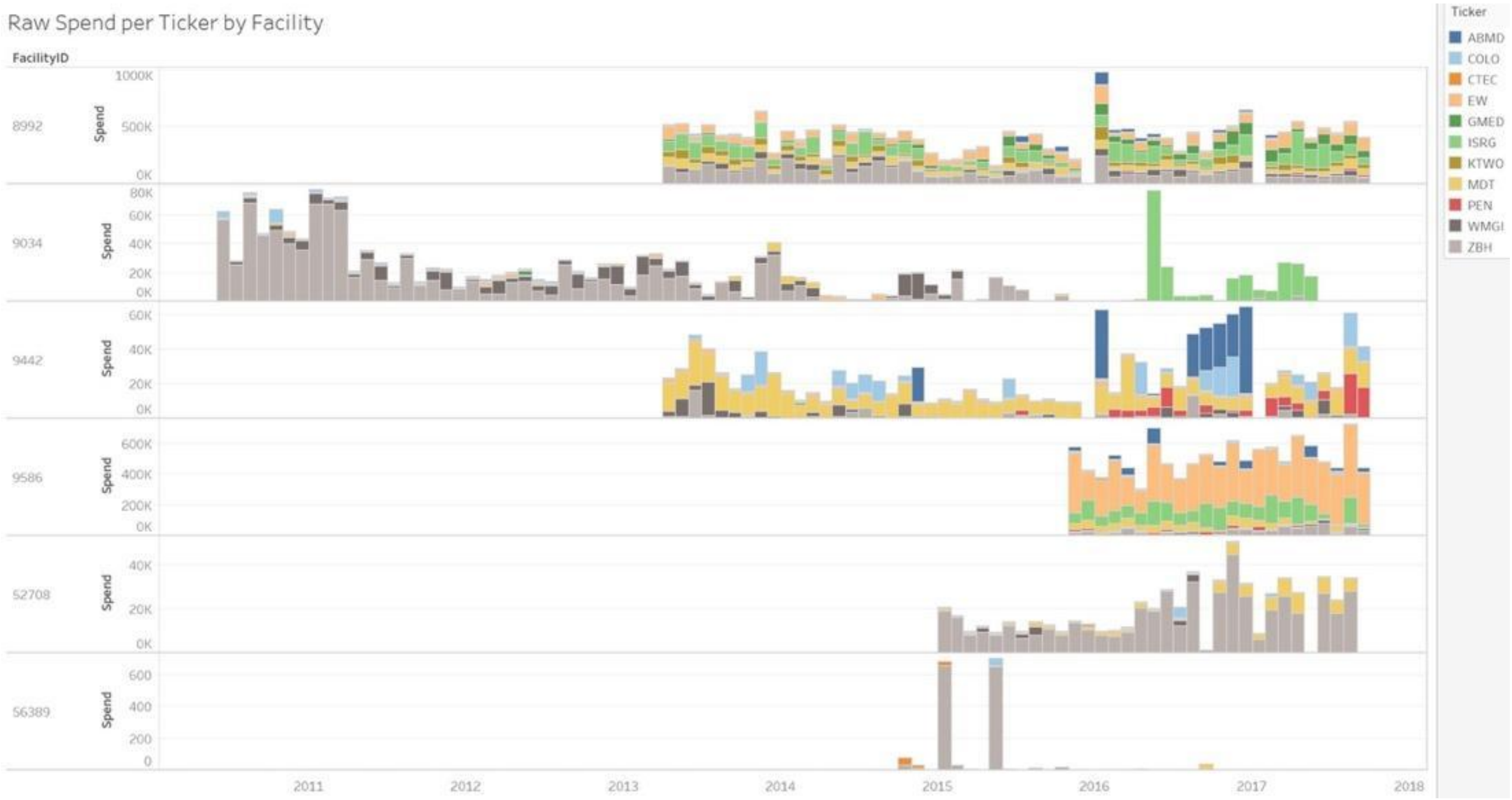


Exhibit 11. Monthly transactions by company at the facility level from the raw data. Gaps and outliers in the data can be clearly seen.

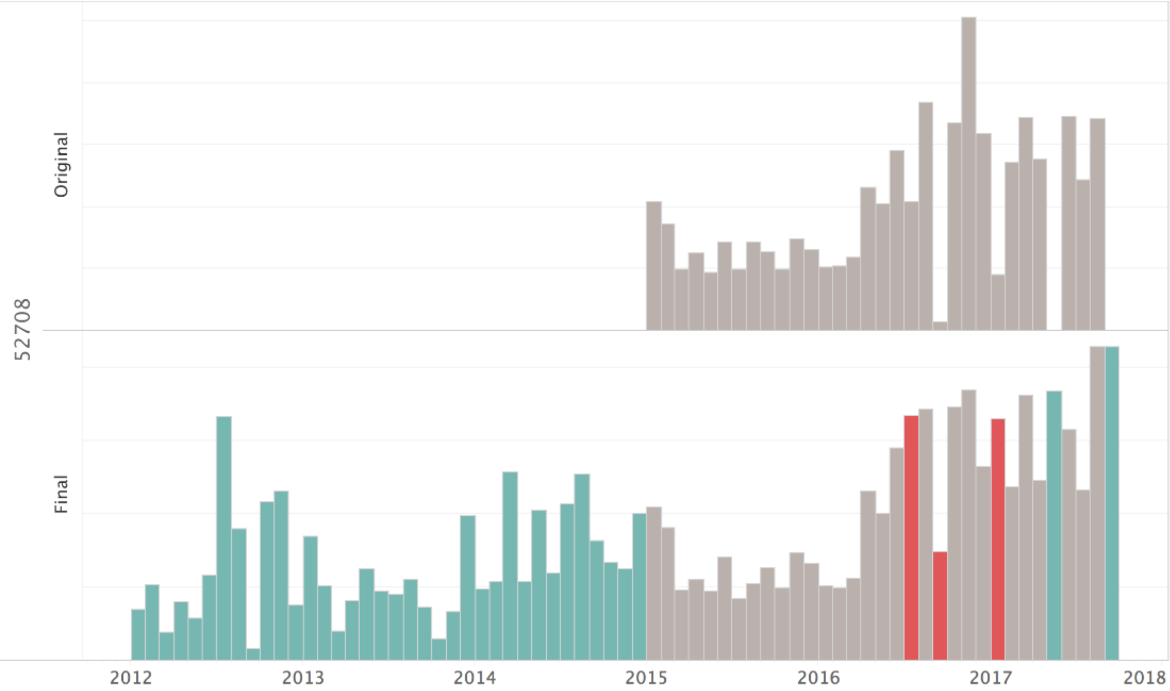


Exhibit 12. Examples of facility-level imputation of missing data (teal) and outliers (red).

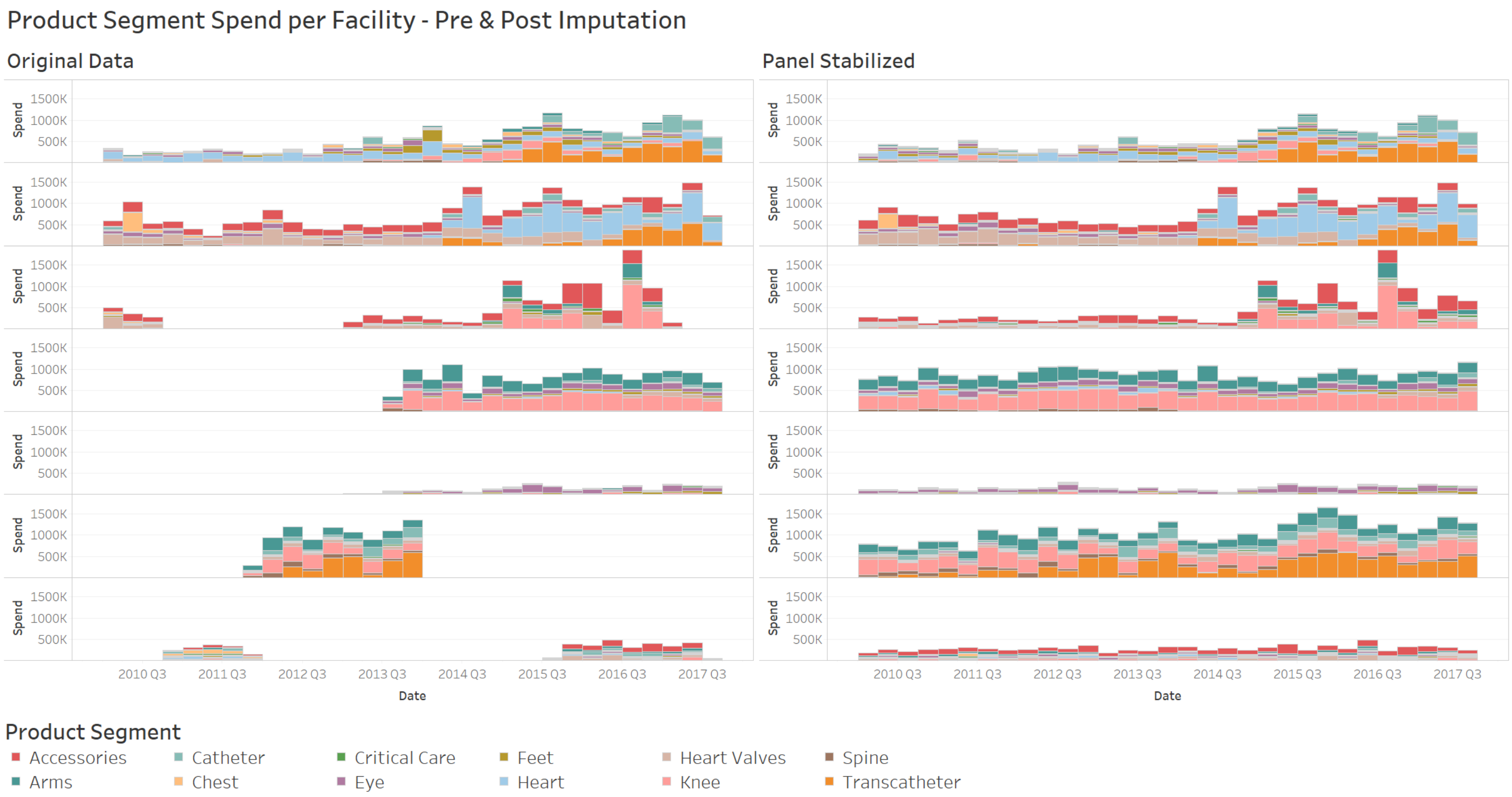


Exhibit 13. Comparison of the original raw data before performing panel-stabilizing imputation (left) and after (right). Note that each facility delivers a specific mix of product categories (in different colors) and their mix of characteristics are preserved after imputation at the facility level.

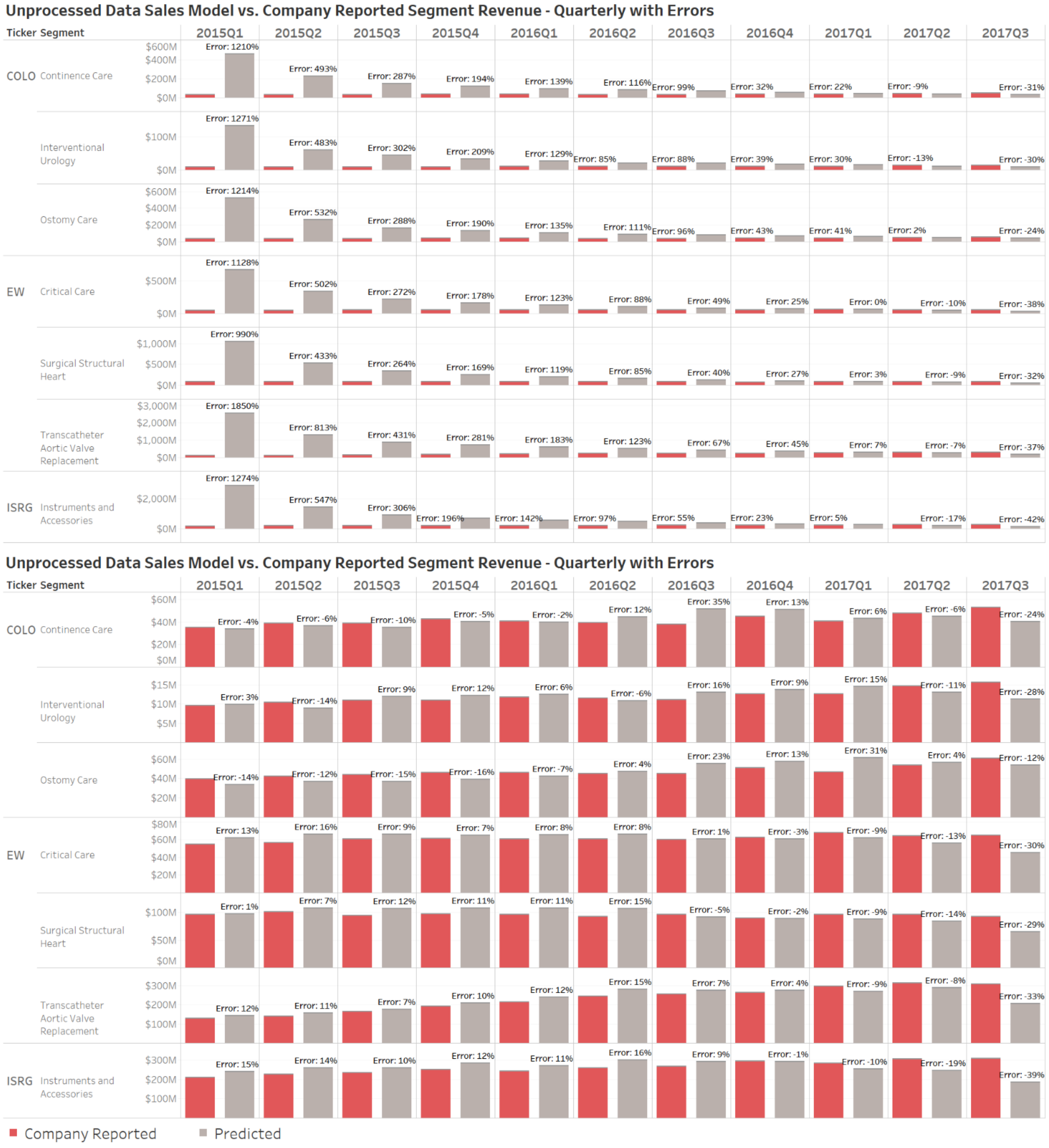


Exhibit 14. Reported segment revenues vs. forecasts for both unprocessed and processed data.