Identifying Corporate Venture Capital Investors – A Data-Cleaning Procedure¹

Patrick Röhm - Researcher and Doctoral Student - University of Hohenheim
Wollgrasweg 49, 70599 Stuttgart, Germany
+49 711 459 24824 - pgroehm@uni-hohenheim.de

Markus Merz - Researcher and Doctoral Student - University of Tübingen
Nauklerstraße 47, 72074 Tübingen, Germany
+49 7071 29 76745 - markus.merz@wiwi.uni-tuebingen.de

Andreas Kuckertz - Chaired Professor - University of Hohenheim †
Wollgrasweg 49, 70599 Stuttgart, Germany
+49 711 459 24820 - andreas.kuckertz@uni-hohenheim.de

† Corresponding author

¹ This is a preprint version. For the final version refer to Röhm, P., Merz, M. & Kuckertz, A. (2020). Identifying Corporate Venture Capital Investors – A Data-Cleaning Procedure. Finance Research Letters. 32 (January): 101092.

Electronic copy available at: https://ssrn.com/abstract=3583310
Identifying Corporate Venture Capital Investors – A Data-Cleaning Procedure

Abstract

The majority of research on corporate venture capital (CVC) relies on data retrieved from secondary databases. The various databases however define CVC differently. Generally, researchers rely on the definition of the used database. As a result, empirical CVC research is not readily comparable, and replicability across databases is often impossible. This article examines the scope and consistency of the most popular databases in CVC research: Eikon from Thomson Reuters and Dow Jones’ VentureSource. The outcome is a replicable data-cleaning procedure based on an appropriate CVC definition. The article provides a necessary basis for the future discourse on CVC.

Highlights

- Research on corporate venture capital (CVC) primarily relies on two databases.
- These databases are Eikon from Thomson Reuters and Dow Jones’ VentureSource.
- Definitions of CVC firms vary in these databases.
- We propose a data-cleaning procedure to identify CVC firms and align both databases.
- Employing the procedure helps to enhance comparability of future research efforts.

Keywords

Corporate venture capital, CVC, databases

Electronic copy available at: https://ssrn.com/abstract=3583310
1. Introduction

Corporate venture capital (CVC) is increasingly becoming a means through which established firms gain an edge in today’s business. Investment funds, or in this case CVC units, are established within a parent company (Dushnitsky 2006). The funds target nascent firms with promising technologies that are usually strategically aligned with the mother firm (Ernst et al. 2005). CVC investments provide start-ups with capital and industry knowledge, and in turn, the parent companies acquire access to potentially disruptive technologies and emerging markets (e.g., Dushnitsky and Lenox 2005). The increased CVC activity has stimulated academic interest in the topic, resulting in a rapidly growing body of research (see Röhm 2018 for an overview). However, empirical research into its workings and impact has been hindered by data limitations and the absence of a common definition of CVC. This makes it particularly difficult to gauge the progression of CVC research.

There have been some attempts to propose a common theoretically-grounded CVC definition for future empirical work. Chemmanur et al. (2014), for example, define several dimensions that a firm should comply with in order to be considered a CVC. In their view, CVCs are stand-alone subsidiaries of nonfinancial corporations that strategically invest in new ventures on behalf of their corporate parents to enhance competitive advantage. CVCs typically pursue both strategic and financial goals and are characterized by a managerial compensation practice that is tied to the parent company’s performance.

In contrast, the majority of empirical studies base their definition of CVC on presets from the corresponding data providers, each of which has its own slightly different CVC definition. VentureSource classifies investors as a CVC if they invest in ventures through a dedicated fund to simultaneously achieve financial and strategic objectives (personal communication, September 2017; VentureSource 2018b). Eikon treats corporate subsidiaries as CVCs if they are actively involved in private equity (PE) related investments (personal communication, September to October 2017). According to Crunchbase, a CVC is an arm of a corporation that invests in innovative start-up companies, whereas Pitchbook considers all forms of equity investment to be CVCs. CB Insights defines CVCs as specialized divisions of larger companies, that directly invest.
in external private companies. Even for the same database, it is hard to replicate empirical results because the understanding of CVC activities varies among researchers (see e.g., Dushnitsky 2006 for an overview) and most studies give no detailed information on the applied search settings within the commercial databases. Additionally, researchers have reported inconsistencies among databases (e.g., Kaplan et al. 2002; Lerner 1994, 1995; Maats et al. 2011). In fact, we are unaware of any detailed comparison of CVC databases.

Based on the theoretical literature, we define CVC units as wholly-owned subsidiaries of nonfinancial corporations that invest in start-ups on behalf of their corporate parent (e.g., Souitaris et al. 2012; Chemmanur et al. 2014). Using this definition, we propose a replicable data-cleaning procedure for the two most popular CVC research databases: *Eikon* from Thomson Reuters and *Dow Jones’ VentureSource*. We thereby help to put future CVC research on a common footing, which would facilitate academic discussion and promote coherence across future research. Additionally, we contribute to the literature on the consistency and reliability of venture capital (VC) related databases by shedding light on the scope of CVC data in the two most extensively used databases.

### 2. Relevant databases for CVC research

To identify the most prominent databases for CVC research, we conducted an extensive literature review based on Elsevier’s Scopus database. We searched Scopus for occurrences of the search strings *venture capital* or *corporate venture capital* in either the title, abstract, or keywords. Additionally, we limit the results to academic papers published in journals before March 2018 and written in English; applying the initial criteria meant we downloaded 2,128 articles. To extract information about the underlying databases used by the articles, we drew on the text analysis program Linguistic Inquiry and Word Count (LIWC) 2015 from Pennebaker et al. (2015), and controlled for inconsistencies in spelling. With 551 appearances, *Eikon* (also known as Thomson One, VentureXpert, or Venture Economics and with a history of data collection going back to 1961) is used most extensively, followed by *VentureSource* (also known as VentureOne that has been collecting data since 1994) with 95 appearances. Other databases such as *Crunchbase* (26

\[\text{Prequin}\] does not provide any specific definition of CVC but considers it a subgroup of VC.

---

Electronic copy available at: https://ssrn.com/abstract=3583310
appearances), Prequin (31 appearances), Pitchbook (9 appearances) and CB Insights (9 appearances) play only a minor role.\(^3\) The results are similar to those of Da Rin et al. (2013), who reported that the two primary commercial databases used in VC research are Thomson Reuters’ Eikon and VentureSource from Dow Jones. Hence, we will focus on these two databases in the remainder of this paper.

VentureSource provides information for 36,000 CVC investors and offers data points for about 101,000 PE- and VC-backed companies (VentureSource 2018). In comparison, the “private equity screener” of Eikon comprises information on about 22,000 investors with 51,000 funds and a total number of 133,000 PE- and VC-backed companies (Thomson Reuters 2018). To gather information both databases rely on extensive quarterly surveys of investors in the VC industry; surveys that grant access to sensitive information that is not presented in official deal statements. Additionally, VentureSource uses its Factiva database and a web-crawler to identify information from press releases and investors’ homepages (personal communication, September 2017). Likewise, Eikon draws on government filings, public news releases, and on PE newsmakers including the European Venture Capital and Private Equity Journal (personal communication, September to October 2017; Thomson 2008, 2010).

3. Data sample

To develop a common data-cleaning process for the given CVC definition, we rely on the two primary databases: Eikon and VentureSource. For each database we construct two samples, one for US-based CVCs and one for CVC vehicles headquartered in Europe.\(^4\) As described by Gompers and Lerner (2000) CVC activities are recurrent and strongly related to the general economic condition. In order to cover a full boom-bust cycle, we draw on an extensive dataset ranging from January 2000 to December 2015.\(^5\) In addition, we do not restrict the country of origin of the investees, thus allowing for cross-country investments. We are well aware of the fact that particularly the VC market in Europe is highly diverse in terms of institutional attractiveness (Groh

---

\(^3\) Because some articles discuss several databases simultaneously, the counts cannot be interpreted as mutually exclusive. Some empirical studies also rely on unique, mostly hand-collected data alongside the use of databases.

\(^4\) Note that Europe also includes the non-EU countries Iceland, Norway, Russia, Switzerland, and Turkey.

\(^5\) According to CB Insights (2017) corporate venture capital amounted to $16.8 bn. in 2000 before it substantially fell off after the dot-com bubble burst. It then began to increase again (before dipping during the financial crisis) and reached its current maximum of $28.4 bn. in 2015.
et al. 2010). However, using Europe as a subsample makes it possible to demonstrate the data-cleaning procedure in two geographical areas that are commonly used to describe VC-, PE-, and CVC-related phenomena.

In both databases, the search criteria were set to an appropriate minimum, reducing the risk of omitting a CVC unit owing to incorrect classification in the databases. Accordingly, in addition to using geographical settings, we predefine “Corporate Venture Capital” as an investor type in VentureSource and “Corporate PE/Venture” as a firm type in Eikon. For the predefined period of sixteen years we found 629 investors, 9,602 investees and a total of 19,077 investment rounds (Europe: 282 investors, 2,737 investees, 4,540 investment rounds) for the US-based Eikon sample. For VentureSource our initial data set comprised 235 investors, 4,532 investees and a total number of 7,719 investment rounds (Europe: 171 investors, 2,026 investees, 3,283 investment rounds).

The previously specified samples serve as a starting point for the subsequent data-cleaning process.

4. Data-cleaning process
The proposed data-cleaning procedure comprises seven steps and results in a generic definition for a CVC unit. The underlying methodology of the data-cleaning procedure is shown in Figure 1.

[Insert Figure 1 here]

In the following section, we introduce each step of the procedure and discuss how the underlying samples from both databases are affected. Table 1 offers an overview in numbers of the excluded investors, investees, and investment rounds for both data providers and for each continent separately based on the applied criteria.

[Insert Table 1 here]

Undisclosed investors. Building on the initial step of retrieving the raw data from the databases, we drop all CVC units where only information on the investee but not on the investors was available. This only affected the Eikon sample, in which unknown investors are categorized as “Undisclosed Investors” in the US data and as “Undisclosed Firm” or “Other UK Investor(s)” in
the European data. The omission of such investors led to the exclusion of 3,101 (5) investees in the US (Europe) sample. This step eliminates one third of the hits from the US. Eikon indicates the investors are inactive or unknown. Eikon does not have the full information for these cases, but using a manual double-check with other sources for each investment might be appropriate for some research questions. This, however, exceeds the scope of this paper.

**Unknown investors.** Third, we merge all investor specific information with data from the Capital IQ platform of Standard & Poor’s. This made it possible to triangulate our data by drawing on an extensive pool of more than 4 million companies (S&P Global 2018). Compared to other databases, Capital IQ offers a broad coverage of both private and public companies. The database provides information on the investors’ business descriptions and information related to the company affiliation. We exclude all investors where we could not find a fitting investor profile in the Capital IQ database. Doing so ensured data consistency and simultaneously provided a solid and reliable foundation for the subsequent steps. This step led to the exclusion of 44 US-based investors appearing in the Eikon sample (25 in Europe) and 11 appearing in the VentureSource sample (7 in Europe).

**Geographical overlap.** The fourth step includes the analysis of the investors’ position within an existing corporate network. This is important because knowledge typically flows from the investor to the corresponding corporate mother (e.g., Gupta and Govindarajan 2000). Hence, the corporate mother determines the geographical affiliation. However, previous articles in the field of CVC limit their empirical analysis to one geographical area (Röhm 2018), thus making them vulnerable to excluding external factors such as cultural aspects or institutional settings. Therefore, authors need to clarify if their selected CVC units are still suitable to their research question. In order to elucidate the ownership status and thus determine the geographic affiliation, we draw on the Capital IQ database to identify potential corporate mothers for each investor. Accordingly, we use the business descriptions in conjunction with the corporate tree function of Capital IQ to clearly match the investor to a corporate mother. Although we excluded non-US and non-European investors from our sample construction, we could still identify a large number of investors with a corporate mother from an excluded geographical region. For instance, German-based companies such as BMW and Bertelsmann operate investment vehicles in the USA. Both databases classify
these CVC units as US-based, although the corporate mother is from Europe. Accordingly, we omit all CVC units with a corporate mother from a different region. This procedure resulted in the exclusion of 80 investors from the US sample of Eikon (7 in Europe) and 50 from VentureSource (4 in Europe).

Alternative investors. Based on the business description, we omit business associations, NGOs, universities, regional development vehicles, advisory firms, independent VCs, and several other non-CVC investment vehicles such as hedge funds, PE investors, and business angels. These investor types were initially recorded as CVC units in the databases but do not meet the definition due to missing corporate parents or because of their own descriptions of the unit in the Capital IQ database. Including them would therefore risk skewing the empirical analysis. Accordingly, between eight and twenty-five percent of the remaining investors were removed.

CVC governance. The sixth step includes the deep analysis of the remaining corporate investment vehicles. Following Dushnitsky (2006), corporations can structure their venturing activities in three ways: first, they can act as a limited partner (LP) in already existing funds of independent VCs (IVC). Second, the investments can be organized through an operating business unit responsible for the venturing strategy (also called direct investments). In practice, it is mainly R&D or business development units that are responsible for such transactions (Bertoni et al. 2013). Third, CVC units can also be organized as wholly-owned subsidiaries within corporate boundaries. The problem, however, is that investments made through IVCs cannot be assigned to a specific corporate LP and are therefore not observable in the databases. There are also challenges involved in clearly matching direct CVC investments, because commercial databases only provide information about the existing corporate entities but not on the business unit level. Consequently, only wholly-owned subsidiaries were considered in the further analysis using the corporate trees in Capital IQ. In line with Dushnitsky and Lenox (2005), we also exclude corporate pension trusts and comparable investment schemes. This step led to the exclusion of 54 percent of the investment vehicles in the US-based sample of Eikon (17 percent in Europe) and 21 percent of the VentureSource investors (14 percent in Europe).
Outside LPs. In contrast to the proper sense of CVC, some corporate venture units act as a general partner (GP) for external investors. In this case, LPs such as insurance firms, can invest in a fund organized and run by a CVC and benefit from the knowledge and experience of the GP. However, this investment practice is accompanied by a risk of sharing knowledge with actual or potential competitors through a knowledge outflow. Therefore, we excluded CVC vehicles with external LPs. This results in the exclusion of 22 investors appearing in the Eikon US sample (Europe 17), including prestigious CVCs such as Intel Capital. Looking at the VentureSource samples we excluded investors of similar magnitude resulting in a drop of 8 (9) percent of the investors in the US (European) sample.

CVC definition. The process deployed above yields the specified CVC vehicles. Of 629 (282), we consider 179 as CVCs in the Eikon sample (Europe 139). In VentureSource, we identify out of 235 (171) listed CVCs, 116 (115) for the USA (Europe). All other firms cannot be considered a CVC because they are funded by financial companies, partnerships, or multiple corporate parents, or have a foreign or unknown parent (Chemmanur et al. 2014).

A large number of the identified CVCs are present in both databases. In particular, we identify 75 (65) shared CVC investors in the US (European) sample. Overall, it appears that Eikon offers a greater availability of CVC investors. However, a closer look reveals that this is mainly driven by past data points. More recently, VentureSource has caught up, offering similar numbers of CVC investors (see Table 2). Examining the industry groups of the unique investors reveals that Eikon is especially suited for US-based CVCs from the transportation and utilities industries (SIC codes starting with 4). In comparison, VentureSource has a greater availability of European CVCs from manufacturing industry (SIC codes starting with 2 or 3) and US-based CVCs from the service industry (SIC codes starting with 7 or 8). Regarding the covered investment rounds, Eikon systematically offers greater data coverage with one exception: VentureSource covers more investment rounds in the European sample between the years 2011 and 2012. Moreover, we found that the underlying definition of CVC in VentureSource is superior to the definition provided in this article.

Electronic copy available at: https://ssrn.com/abstract=3583310
5. Conclusion

This article seeks to address how CVC activity is measured and in which ways the commonly used databases, namely Eikon and VentureSource can be used to reach a theoretically defined dataset of CVCs. Most published studies provide researchers with insufficient information about the technical definition of CVC or base their empirical work on the definition of the commercial data providers. We propose a data-cleaning procedure to promote future coherence in research. The presented results significantly contribute to the ongoing discussion of CVC. We provide a data-cleaning process allowing researchers to more generically define CVC. This would increase comparability and replicability of the research results. Moreover, we provide a comprehensive analysis of the data coverage in the commonly used databases of Eikon and VentureSource. This can help researchers decide which data provider is better suited for their research question.

References

Alvarez-Garrido E, Dushnitsky G (2016) Are entrepreneurial venture's innovation rates sensitive to investor complementary assets? Comparing biotech ventures backed by corporate and independent VCs. Strategic Management Journal, 37(5):819-834. https://doi.org/10.1002/smj.2359

Bertoni F, Colombo MG, Grilli L (2013) Venture capital investor type and the growth mode of new technology-based firms. Small Business Economics, 40(3):527-552. https://doi.org/10.1007/s11187-011-9385-9

Chemmanur TJ, Loutska E, Tian X (2014) Corporate venture capital, value creation, and innovation. The Review of Financial Studies, 27(8):2434-2473. https://doi.org/10.1093/rfs/hhu033

CB Insights (2017) The history of CVC: from Exxon and DuPont to Xerox and Microsoft, how corporations began chasing ‘the future’. https://www.cbinsights.com/research/report/corporate-venture-capital-history/

Da Rin M, Hellmann T, Puri M (2013) A survey of venture capital research. In: Constantinides GM, Harris M, Stulz RM (eds) Handbook of the Economics of Finance. Elsevier, pp 573-648. https://doi.org/10.1016/B978-0-44-453594-8.00008-2

Dushnitsky G (2006) Corporate venture capital: Past evidence and future directions. In: Classon W, Yeung B, Basu A, Wadeson N (eds) The Oxford handbook of entrepreneurship. Oxford: Oxford University Press, pp 387-431. https://doi.org/10.1093/oxfordhb/9780199546992.003.0015

Dushnitsky G, Lenox MJ (2005) When do incumbents learn from entrepreneurial ventures? Corporate venture capital and investing firm innovation rates. Research Policy, 34(5):615-639. https://doi.org/10.1016/j.respol.2005.01.017
| Step       | Excluded investors | Excluded investees | Excluded rounds |
|------------|--------------------|--------------------|-----------------|
| **Step 1: Full sample** | 1 (0%) | 282 (1%) | 0 (0%) |
| Initial investors | 629 | 2,737 | 235 |
| Initial investees | 9,602 | 4,540 | 4,532 |
| Initial rounds | 7,719 | 3,283 | 2,026 |
| **Step 2: Undisclosed investors** | 2 (1%) | 0 (0%) | 0 (0%) |
| Excluded investors | 3,101 (32%) | 5 (0%) | 0 (0%) |
| Excluded investees | 6,332 (33%) | 8 (0%) | 0 (0%) |
| Excluded rounds | 0 (0%) | 0 (0%) | 0 (0%) |
| **Step 3: Unknown investors** | 0 (0%) | 0 (0%) | 0 (0%) |
| Excluded investors | 92 (1%) | 157 (6%) | 24 (1%) |
| Excluded investees | 199 (2%) | 213 (5%) | 14 (1%) |
| Excluded rounds | 43 (1%) | 16 (0%) | 16 (0%) |
| **Step 4: Geographical overlap** | 0 (0%) | 0 (0%) | 0 (0%) |
| Excluded investors | 731 (11%) | 55 (2%) | 571 (13%) |
| Excluded investees | 1,885 (15%) | 69 (2%) | 1,161 (13%) |
| Excluded rounds | 8 (0%) | 0 (0%) | 12 (1%) |
| **Step 5: Alternative investors** | 0 (0%) | 0 (0%) | 0 (0%) |
| Excluded investors | 507 (9%) | 636 (25%) | 86 (2%) |
| Excluded investees | 901 (8%) | 1,089 (26%) | 207 (3%) |
| Excluded rounds | 31 (17%) | 207 (33%) | 510 (16%) |
| **Step 6: CVC governance** | 0 (0%) | 0 (0%) | 0 (0%) |
| Excluded investors | 843 (16%) | 123 (7%) | 276 (7%) |
| Excluded investees | 1,828 (19%) | 155 (5%) | 419 (7%) |
| Excluded rounds | 33 (21%) | 10 (8%) | 91 (3%) |
| **Step 7: Outside LPs** | 0 (0%) | 0 (0%) | 0 (0%) |
| Excluded investors | 1,318 (30%) | 434 (25%) | 362 (22%) |
| Excluded investees | 2,604 (33%) | 732 (24%) | 1,319 |
| Excluded rounds | 3,015 | 1,327 | 2,344 |

**CVC definition**

- Remaining investors: 179 139 116 115
- Remaining investees: 3,015 1,327 2,344 1,319
- Remaining rounds: 5,328 2,274 3,721 1,987

Electronic copy available at: https://ssrn.com/abstract=3583310
Table 2. Comparison of unique CVCs and investment rounds (follow-on rounds excluded) covering the period between 2000 and 2015

|                      | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|----------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| **Unique CVCs**      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| US sample            |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| VentureSource        | 31   | 29   | 24   | 28   | 26   | 29   | 25   | 26   | 35   | 33   | 34   | 30   | 31   | 36   | 43   |      |
| Eikon                | 63   | 58   | 36   | 44   | 37   | 43   | 38   | 38   | 45   | 34   | 35   | 42   | 36   | 38   | 43   | 45   |
| **Investment rounds**|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| US sample            |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| VentureSource        | 288  | 180  | 117  | 109  | 107  | 136  | 147  | 147  | 136  | 176  | 203  | 217  | 252  | 313  | 381  |      |
| Eikon                | 716  | 321  | 166  | 134  | 150  | 161  | 214  | 227  | 234  | 159  | 210  | 244  | 263  | 284  | 354  |      |
| **Unique CVCs**      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| European sample      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| VentureSource        | 23   | 24   | 26   | 25   | 21   | 21   | 22   | 23   | 27   | 24   | 27   | 23   | 24   | 31   | 36   | 38   |
| Eikon                | 26   | 25   | 23   | 19   | 20   | 21   | 14   | 21   | 29   | 19   | 23   | 22   | 29   | 27   | 31   | 38   |
| **Investment rounds**|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| European sample      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| VentureSource        | 78   | 103  | 79   | 94   | 80   | 81   | 78   | 103  | 104  | 77   | 78   | 110  | 173  | 85   | 110  | 122  |
| Eikon                | 113  | 98   | 64   | 81   | 94   | 83   | 77   | 98   | 112  | 82   | 85   | 82   | 102  | 107  | 129  | 138  |
Figure 1. Underlying methodology of the proposed data-cleaning procedure

**Full sample**
All investment information was included.

**Unknown investors**
Investors not covered in the S&P Capital IQ database were excluded.

**Alternative investors**
Alternative investment vehicles (i.e. incubators, accelerator) were excluded based on the S&P Capital IQ definition.

**Outside LPs**
CVC units with outside LP were excluded.

**Undisclosed investors**
Undisclosed investors, e.g. investors that are categorized as a CVC but without mentioning the specific name, need to be excluded.

**Geographical overlap**
Investors with headquarters in the considered continent were excluded if the corresponding corporate mother is headquartered in a different continent.

**CVC governance**
CVC related instruments (e.g. pension trusts or employee share schemes) were omitted based on the S&P Capital IQ description.

**CVC definition**
This definition of CVCs includes all steps performed previously.

Electronic copy available at: https://ssrn.com/abstract=3583310