A Mixed Methods Systematic Scoping Review of the Use of Big Data in Early Intervention

Research: Ethical and Practical Implications

Sondra M. Stegenga MS, OTR/L, M.Ed.
University of Oregon College of Education
901 E. 18th Ave. Eugene, OR 97405
sondras@uoregon.edu
616-405-5220

Kelley F. Munger MA, LPC
University of Oregon College of Education
901 E. 18th Ave. Eugene, OR 97405
kmunger@uoregon.edu

Jane Squires, Ph.D.
University of Oregon College of Education
901 E. 18th Ave. Eugene, OR 97405
jsquires@uoregon.edu

Daniel Anderson Ph.D.
University of Oregon College of Education
5262 University of Oregon
daniela@uoregon.edu

**Note.* Authors do not have any conflicts of interest to disclose.

Abstract

Big data holds immense potential for innovation and new understanding in research, evaluation, practice, and policy related to young children and their families. Although big data is a relatively new concept, particularly in early intervention systems (EI), recent pushes for data systems alignment in EI and education have propelled the use of large-scale integrated data systems in recent years (U.S. Department of Health and Human Services & the U.S. Department of Education, 2016). This combined with a plethora of new and rapidly increasing data sources has created a new data world. In response, research methodology, ethics, and tools need to be examined to ensure developmentally appropriate and ethical practices in research. A mixed methods systematic scoping review was conducted to gain a foundational understanding of the literature on big data use in EI settings. Strengths, challenges, systems-level needs, and implications for researchers, administrators, and policy makers are included.

Keywords. big data, early intervention, 619, Part C of IDEA, data ethics, underrepresented populations

Introduction

The concept of *big data* is relatively new. The term is used to describe data for which the volume, variety, and/or velocity is of such immense proportions that it cannot be stored or analyzed with traditional methods (e.g. Ahmadi, Dileepan, & Wheatly, 2016). Only recently technology has evolved to the point where these vast data can be not only stored, but analyzed and explored in meaningful ways (Baro, Degoul, Beuscart, Chazard, 2015). Smart phones and some of the most widely used social media sites, such as Facebook and Twitter, have only been in use for a little over a decade, but have led to massive growth in data, and have fundamentally changed everyday life. Big data continues to grow and evolve from these and other sources into one of the most valued resources and hotly discussed topics in our culture (Columbus, 2017; Montgomery, 2015).

Articles and news coverage have depicted the impact of big data in recent political elections, with outcomes linked to the use of data for strategically identifying and targeting voters (Winston, 2016). Businesses are also heavily recruiting and employing data scientists to track, evaluate, and predict sales, buying patterns, and consumer behaviors (Montgomery, 2015). Big data is also being utilized in health care (Zheng, Liu, & Hsieh, 2013). For example, researchers recently used social media data to predict mental health status with high rates of accuracy (Hao, Li, Li, Zhu, 2013). In the health care fields, there have been calls to examine and use big data to improve practices and health outcomes (e.g. Murdoch & Detsky, 2013; Roski, Bo-Linn, & Andrews, 2014). In California, researchers used a new methodology that harnesses the use of big data to examine pre and perinatal practices impacting mortality rates, ultimately allowing them to identify patterns that would otherwise be undetectable with smaller sample sizes (Mitchell, Lawton, Morton, McCain, Holtby, & Main, 2014). Specifically, through linking

public health records, utilizing a case selection algorithm within the medical records which incorporated risk factors, and conducting a final review by a multidisciplinary committee of experts, the health care systems were able to make improvements and decrease maternal mortality rates. The field of education has also begun to integrate use of big data, particularly related to online education and personalized learning, employing machine learning to identify individual needs (e.g. West, 2012). The implications of big data for early intervention (EI) are currently unclear, but hold promising potential for new and deeper investigations.

In countries with aligned data systems and big data use, concerns have arisen about the potential impacts of big data use with young and vulnerable populations (Wrigley & Wormwell, 2016). Trepidations of “schoolification” of young children have arisen, where school age practices are inappropriately pushed down into younger age groups (Wrigley & Wormwell, 2016). Specifically, Wrigley & Wormwell (2016) note concerns of increased testing, data collection, and in-seat time for young children in EI, as well as a potential for lifelong labels and tracking. If conclusions from large scale data analysis are applied inappropriately, outcomes for our youngest and most vulnerable learners might be misrepresented.

Recent controversy over racial bias in predictive analytics also complicates the use of big data (Dressel, & Farid, 2018), as well as the potential for reidentification of personally identifiable and sensitive information (Capatosto, 2017; Crawford & Schultz, 2014). These and other issues have given rise to debates about how and if predictive algorithms should be used with young and vulnerable children and families in child welfare systems (Zarate, 2018). Yet, there is a relative dearth of literature on the use in EI populations. Most information to date has focused on data governance and the linkage of early childhood administrative data systems (e.g. DaSy Center, 2014), which is only a beginning consideration. The concept of big data extends to

geographic, economic, social media, global positioning system (GPS), wearable technology, home-based technology monitoring and assistant systems, and health monitoring data—to name only a few (Montgomery, 2015). While these newer types of data are not yet being systematically aggregated in a widespread manner for the purpose of improving practices, they represent an untapped potential for improving both individual and systems performance at a time when children are especially susceptible to change and growth (i.e. the rapid and crucial developmental period during early childhood). Given these factors and the unique developmental needs of young children, it is imperative that we broadly examine the ethics and implications of big data use specifically to EI, not only within the traditional purview of administrative practices but also within the broader scope of the lives of young children and families.

Review of the Literature

What is Big Data?

Lohr (2013) credits the term big data to computer scientist John Mashey, who first used the term in a seminal presentation in 1998 (Mashey, 1998). However, big data has only appeared within the scholarly literature for the education and health-related sciences since 2008, and in the developmental literature since 2012 (as identified by this review). Although there continues to be a discrepancy about the definition of big data, authors have attempted to define and quantify the concept, with most recent definitions including not only data size but a broader combination of factors often referred to as the “3 Vs” of volume, variety, and velocity (Ahmadi, Dileepan, & Wheatly, 2016; Baro, Degoul, Beuscart, & Chazard, 2015).

Volume. The volume of the data refers to the overall size or amount of data collected. Volume has grown to such an extent that new terms are continually being added to describe it moving from megabytes, to gigabytes, terabytes, petabytes, and exabytes (See Table 1.) McAfee

and Brynjolfsson (2012) provide an informative example of the vast amount of data gathered by a well-known U.S. company, Walmart, which gathers 2.5 petabytes (PB) of data every hour. This amount of data is roughly equivalent to filling 20 million filing cabinets of text every hour of every day. Since this example was provided in 2012, data gathering has only continued to grow with new tracking devices emerging, such as smart clothing (Chen, Ma, Sang, Lai, Hu, 2016) and many home-based devices with smart technology, ranging from refrigerators to home security also known as the Internet of Things (IoT) (Ritz & Knaack, 2017). Our lives are becoming immersed in data collection devices.

Velocity. The velocity of data is the rate or speed at which the data are collected. The vast increases in velocity seen in recent years are due to advances in the speed of the internet, accessibility and prevalence of data collected, and large data sources available. Data is being collected not only more rapidly but across an increasing variety of sources vastly increasing the velocity of data coming in. Increased velocity of data holds the potential not only for retrospective or summative assessment but real time formative assessment and findings. In business, velocity is allowing for real time decision making. For example, one fast food company was able to see a multiplying increase in fried pickle sales which allowed them to make a decision to keep it as an ongoing menu item before feedback was even back from site managers, ultimately allowing them to be ahead on planning and ultimately profits (Dykes, 2017 June 28).

Variety. Variety focuses on the types of data gathered. As technology evolves, so too does the complexity of the type of data available. As mentioned, many types of data are being gathered every second. These are often referred to as digital trace data and include text, pictures, audio, video, GPS, and sensor data (Connelly, Playford, Gayle, & Dibben, 2016, Salganik, 2016). This is different from typical educational or EI data collections, which generally consist

of structured data sets, often limited to text or numeric data in spreadsheet-like formats (Mead & Mitchel, 2016). However, many new sources of data are unstructured, causing difficulty with searching, storing, and analyzing data using traditional methods (Baro et al., 2015; Connelly, Playford, Gayle, & Dibben, 2016). Social media data, for example, are difficult to search according to pre-defined categories because of the varying types and sources of information embedded within them. Advances in electronic curriculum formats and communication modes within education, alongside heightened use of and access to technology has led to increasing numbers and variety of data being stored in the educational portfolio (Mead & Mitchel, 2016).

Table 1

Common Data Sizes and Examples

Size/Volume	Data in Numbers	Equivalence Examples
Megabyte (MB)	1,000,000 bytes	1MB=Small Novel 5MB=Complete works of Shakespeare 500MB = Hard Disk of common PCs
Gigabyte (GB)	1,000,000,000 bytes	1GB=Pickup Truck filled with paper 100GB=A floor of academic journals
Terabyte (TB)	1,000,000,000,000 bytes	1TB=All the X-Ray films in a large hospital 10TB=Printed collection of US Library of Congress
Petabyte (PB)	1,000,000,000,000,000 bytes	1PB=7,812 iPhones (128GB) 2PB=All US academic research libraries 200PB=All printed material in the world
Exabyte	1,000,000,000,000,000,000 bytes	5EB=All words ever spoken by humans to date

Note. Information adapted from Bunn (2012).

Overall, defining big data is a complex task, and no one definition has been fully agreed upon across fields (e.g. Baro et al., 2015, Connelly, Playford, Gayle, & Dibben, 2016).

Therefore, for the purposes of this paper, the term big data is used to describe situations where the magnitude and complexity of the data demonstrate such an immense scale that it requires special considerations for storage and analysis due to any one or combination of the 3V's. This definition encompasses both "found data" as Connelly, Playford, Gayle, & Dibben (2016)

described, as well as data that are “created” through large scale research and administrative data systems (Baro et al, 2015).

Current Data Practices in EI

Recent federal legislation has included mandates for using educational data to drive practices and decision-making, with clauses specifying services to be “scientifically based to the extent practicable” in early intervention 34 C.F.R. §303.112 and based on “peer-reviewed research to the extent practicable” 34 C.F.R. §300.320(a)(4) in preschool age and beyond (Individuals with Disabilities Education Act, 2004; Every Student Succeeds Act, 2015). In addition, having and using high quality data has been cited as a core component of successful EI and early childhood related programs (Lucas, Hurth, & Kasprzak, 2010). This push for data-based decision making as part of a broader system of accountability in educational systems has led to the use of not only individual and classroom level data, but to data collection points at the state and federal levels as well, ultimately leading to vast increases in the amount of administrative data collected (DaSy Center, 2014). However, these changes are not without their difficulties.

First, most data systems in EI and early childhood education were created in “silos,” without any linkage to other systems or age groups, making analysis across systems time-intensive, if not impossible (U.S. Department of Health and Human Services & the U.S. Department of Education, 2016). This lack of linkage led to recent calls for better alignment of data systems in EI and early childhood (e.g. DaSy Center, 2014). Despite tremendous shifts in the use of data and its impact on all levels of practice in EI and early childhood, few publications, research, and/or discussions surrounding the use of big data are available related to this population.

In this paper, we aim to present a foundational understanding of big data use in EI including general characteristics and uses of big data in research, as well as an understanding of the strengths, challenges, knowledge/skills, and systems-level supports needed to optimize the use of big data. This review extends beyond typical early childhood data resources which have mostly focused on data alignment and governance issues related to traditional administrative data. Specifically, it expands the focus to encompass the broader context of big data, including sources outside of large-scale administrative data. Additionally, no systematic review to date has focused on the needs and issues of big data specific to the EI population. This is an important focus because of the vulnerable population with high developmental risk. Therefore, in this review, four main focus areas are examined to build upon the current knowledge base.

First, due to the non-typical nature of sources and uses of big data, information on the topic is almost non-existent within scholarly peer-reviewed journals. Therefore, we aim to identify where information on big data related to EI is being published and summarize this information to gain a foundational understanding on where researchers and systems leaders can look to find pertinent information. Second, as with any methodological innovation, cautions and considerations need to be identified. Third, we identify tools, methods, and skills needed to analyze, understand, and use big data ethically within EI research to provide a starting point for future research.

Methods

Given the aforementioned gaps in the literature, the following research questions were examined:

- 1) What are critical characteristics of research being published on big data use in EI?

(e.g. publication locations and search sources, empirical versus non-empirical, tools

for analyzing outcomes, and methods used to analyze big data in empirical studies that include developmental outcomes with young children)?

- 2) What are the themes within the non-empirical literature related to the use of big data in EI with focus on strengths, challenges, necessary knowledge/skills, and systems level supports?

Rationale for Methodology

Best practices in research and systematic review indicate that research questions and theory should drive methodology and research process (Plano Clark & Ivankova, 2017). Hence, careful attention was taken to identify the appropriate methodology to match the research questions and focus of the review. The topic of big data related to EI populations is so new that little has been published on the topic, creating concerns about conducting a traditional systematic review of empirical studies. Therefore, a scoping review was identified as the optimal choice for this review. Although historically scoping reviews have often lacked the clarity of methodology seen in traditional systematic reviews (Tricco et al., 2016), recent work by Arksey and O'Malley (2005), Levac, Colquhoun, and O'Brien (2010), Colquhoun et al. (2014), and Peters et al. (2015) has facilitated immense progress in systematizing scoping review methodology. In fact, even *Systematic Reviews* journal recently published an article acknowledging the need for and appropriateness of scoping reviews within the family of systematic review (Moher, Steward, & Shekelle, 2015). This combined with the expected release of PRISMA-ScR quality guidelines for scoping reviews in 2018 (Colquhoun, 2016), makes this methodology arguably comprehensive and the best fit for the targeted research questions.

Levac, Colquhoun, and O'Brien (2010) state that scoping reviews are often used to “clarify complex topics” and are “particularly relevant to disciplines with emerging evidence”

(p.1). This is especially pertinent for investigating big data use in the field of EI. Often, framework development is taken on without a methodology for searching and understanding the literature due to the complexity of topics, potential mix of study designs, complex locations of literature sources, and the need for involvement of gray literature (Tricco et al, 2016). Therefore, scoping reviews hold the potential to be particularly useful in bringing structure to this knowledge-building process when creating a new framework for understanding and as a guide for future research. Overall, the steps for conducting a scoping review developed by Arksey and O'Malley (2005), refined by Levac, Colquhoun, and O'Brien (2010) as well as Colquhoun et al. (2014), and clarified by Peters, Godfrey, McInerney, Soares, Hanan, & Parker (2015) were utilized in this review, including the six stage framework. The stages include: 1) Identifying and clarifying the research question(s), 2) Identifying relevant studies while balancing feasibility and comprehensiveness, 3) Study selection completed as part of an iterative team approach, 4) Charting the data including both quantitative counts and descriptives as well as qualitative themes, 5) Collating and summarizing the results including application to policy and/or practice, 6) Consultation with experts for verification of results (Peters et. al, 2015).

Inclusion Criteria, Exclusion Criteria, and Search Strategy

The population of interest for this review was early intervention (EI), including infants, toddlers, and young children from birth to kindergarten with or at risk of disabilities. Due to the emerging nature of the topic of big data and EI, years and country were not limited in the search. In addition, traditional peer-reviewed journal articles and first- and second-tier gray literature were included due to the limited literature base. First-tier gray literature is defined by Adams, Smart, and Huff (2017) as items with significant credibility and includes sources such as books, journals, government reports, and publications by think tank organizations. Second-tier gray

literature, as defined by Adams et al. (2017), includes a wide range of materials from blog posts to presentations. For the purposes of this review, second-tier gray literature was limited to peer-reviewed conference presentations that were available in written format (e.g. printed slides or posters) in order to capture up-and-coming information that was directly related to the topic and demonstrated credibility.

Inclusion criteria, exclusion criteria, search terms, and databases relates to Stage 2 of the Scoping Review Framework (Peters et al., 2015). Inclusion criteria included: big data, children ages birth to kindergarten. Articles were included if they had a main focus on education and included early childhood education as part of their definition of education. Exclusion criteria included articles that (a) only focused on school age children or older, (b) focused on education but did not include early childhood as part of their definition and/or taxonomy, (c) were only medical in nature/did not list developmental and/or educational outcomes, or (d) did not state explicitly the use of big data. Caution was taken to conduct the most accurate searches possible by consulting with experts, conducting pilot searches, and hand searching key words of relevant articles.

The search terms utilized for the scoping study included the phrase “big data” plus any of the following: “early intervention”, “Part C”, “early childhood”, “preschool”, “infant*”, “toddler*”, “infant*” + “disabilit*”, “toddler*” + “disabilit*”. The sources of the search were chosen according to direct relevance to the field of EI and included 1) traditional databases relevant to EI and early development including ERIC, Psychnet, and Medline 2) Google Scholar, 3) Google search, and 4) hand/snowball searches to leverage reference sections of previously identified literature identified in this review. Searches were refined as an iterative process from peer-reviewed literature into tiers of gray literature. In addition, due to the massive amounts of

data produced from Google searches, a data saturation approach was taken. In Google Scholar and Google, only the search terms of best match, big data + “early childhood” were utilized in order to attain the highest possible search accuracy. The first author conducted the main search to identify articles for inclusion. In addition, a second reviewer conducted reliability for article inclusion on 20% of the articles with 100% agreement on inclusion of the studies. Searches were conducted in December 2017. See Figure 1 for further detail on inclusion/exclusion criteria according to current PRISMA Guidelines/Flowchart. Study selection is part of Stage 3 of the Scoping Review Framework (Peters et al, 2015).

Figure 1. PRISMA Flow Chart. Adapted from Moher, Liberati, Tetzlaff, & Altman (2009).

Data Coding and Extraction

Data coding and extraction are part of Stage 4 of the Scoping Review Framework (Peters et al, 2015). The first author and an additional reviewer with expertise in EI, qualitative analysis, and quantitative analysis, independently extracted and charted data from the included articles with the first author conducting coding on all articles and the second reviewer conducting coding on 20% of the articles. All studies were coded according to country, literature class, and type (empirical versus non-empirical). Literature class was coded into four categories including: (a) Peer-reviewed journals, (b) Professional journals where peer review status was unknown, (c) Tier 1 gray literature, and (d) select Tier 2 gray.

The empirical studies were coded according to set criteria in Research Question 1. The non-empirical articles were coded using deductive/pragmatic reasoning to identify and code content related to Research Question 2 through use of a predetermined coding template according to recommended methodology for qualitative content analysis (QCA) within systematic review by Finfgel-Connett (2014). Although the newest scoping review recommendations by Colquhoun et al. (2014) recommend assessing study quality, we make the case that rather than necessitating study quality as a component of a scoping review methodology, this decision should be determined by the research questions. This is due to the recent concerns raised by Wolgemuth, Hicks, and Agosto (2017) regarding study inclusion and exclusion criteria in systematic review. Specifically, important information may be missed if studies are eliminated due to quality ratings when addressing research questions in limited fields of research or with vulnerable populations (Wolgemuth, Hicks & Agosto, 2017), such as this review. This perspective that the use of quality assessment should be determined by the research question is also supported by recent guidelines for scoping review by Peters et al. (2015).

An Excel document was used for recording coded data from the included articles and stored in a secure cloud-based storage to provide a clear audit trail and collaboration between article reviewers. R Studio version 1.1.290 and R version 3.4.2 were utilized for frequency calculations and data visualizations. Qualitative Content Analysis (QCA) was utilized to chart the themes of the non-empirical articles. QCA has been recommended as a method for identifying themes in reviews involving qualitative analysis (Finfgel-Connett, 2014) and is acknowledged as being particularly beneficial when new knowledge and framework development is involved as part of Stage 4 and Stage 5 of the scoping review framework developed by Arksey & O'Malley (2005). A deductive approach was utilized with the QCA to analyze overall themes of the articles according to the aforementioned research questions.

Threats to validity were addressed through use of recommended practices including member checking/consultation with experts, bias reduction, memo-ing, diagramming, and reflection (Creswell & Poth, 2017; Finfgel-Connett, 2014). Consultation with experts not only served as member checking as part of the QCA process but also relates to Stage 6 of the Scoping Study Framework (Peters et al, 2015).

Results

Quantitative analyses (frequencies and characteristics) were used to answer Research Question 1, while qualitative analysis (themes) was used to answer Research Question 2. We report results separately by methodology/research question according to the scoping review reporting recommendations (Peters et al., 2015).

Quantitative Analysis (Frequencies and Characteristics)

The 28 studies that met final inclusion criteria for the review were from 5 countries and ranged in publication date from 2012 to 2017 with a gradual increase in articles each year except

for 2017 but it is noted that the review was conducted two weeks prior to the end of December 2017, potentially accounting for the slight decrease in articles for 2017 (See Figure 2).

Figure 2. Number of Articles by Year and Country

Seven articles were empirical and 21 were non-empirical. The articles represented 11 different fields of study including early intervention, economics, education, environment, law, mental health, pediatrics, policy, psychology, social work, and speech (See Figure 3).

Figure 3. Number of Articles by Field of Study and Type

Articles were from four different grades of literature, as shown in Figure 4, with 19 (68%) from peer-reviewed journals, 2 (7%) from professional journals, 5 (18%) from Tier 1 gray literature, and 2 (7%) from Tier 2 gray literature. All empirical studies identified for the review utilized quantitative methods and none used mixed or qualitative methods. Approximately 43% of empirical studies listed the tools/software utilized for analysis with 29% reporting the use of SAS, 14% reporting the use of DAGitty, and the remaining 57% percent not reporting the software/tools utilized. All empirical studies used at least two linked data sets and 29% reported data linkage rates. Approximately 29% of empirical studies reported procedures related to data privacy. All empirical studies listed analysis and methods used, although terminology and thoroughness of procedure description varied. For example, some studies discussed model fitting procedures and tests/assumptions whereas others only listed analysis used.

Figure 4. Number of Articles by Literature Grade and Search Source

Qualitative Analysis (Themes)

To answer Research Question 2, several themes emerged from the non-empirical study coding. The coding identified themes within the deductive/pre-determined areas of strengths, challenges, governance/systems, and knowledge/skills/tools to answer the research questions.

Strengths. Eight core themes related to potential strengths of big data use in areas of development and early childhood were identified. These included: availability, timeliness/efficiency, cost effectiveness, decreased subject risk, new insights, increased strength of conclusions, broad applicability, and personalized feedback (See Table 2).

Table 2

Thematic Results – Strengths of Big Data Use in EI

Strength	Articles	Brief Description of Theme
Availability	Blair (2016); Beckman (2017); Coulton, Goerge, Putnam-Hornstein, & de Haan (2015); Payakachat (2016); Piety (2014); Pugh & Foster (2014); Zeide (2015)	Many datasets already available
Timeliness/efficiency	Blair (2016); Boone-Heinonen et al. (2017); Coulton et al. (2015)	Decreased time due to no subject recruitment period
Cost effectiveness	Blair (2016); Falster et al. et al., (2015); Mader (2015)	Data often can be obtained at no cost/low cost; savings due to time and efficiency/no recruitment
Decreased subject risk	Blair (2016); Currie (2013)	No new subject exposure/use of pre-existing data
New insights	Blair (2016); Bone et al. (2015); Brundage et al. (2016); Falster et al. (2015); Fischer, Anthony, & Dorman (2014); Payakachat (2016)	New insights may be possible with big data/combined data sets due to due to: 1) ability to detect small effects, 2) identification of effects on subgroups of individuals, 3) ability to examine long term effects

Increased strength of conclusions	Blair (2016); Cohen & Nonacs (2016); Coulton et al. (2015); Currie (2013); Falster et al. (2015); Fischer, Anthony, & Dorman (2014); Mader (2015)	With large-scale data there is potential for increased controls, better representation from population level data, decreased risk of bias due to attrition and selection, and increased level of control/covariates; particularly when datasets are combined
Broad applicability	Mader (2015); Terzeieva (2015); Mead & Mitchel (2016); Pugh & Foster (2014); Wrigley & Wormwell (2016)	Big data available with potential uses for research, practice, program improvement, and policy
Personalized feedback	Montgomery (2015); Terzevia (2014)	Potential for personalized feedback on use of data through use of machine learning

Challenges. Seven core themes related to potential challenges of big data use in EI were identified, which included data quality, complexity, limited scope, historical lack of a data-driven culture, ethical challenges, legal issues, and structural problems (See Table 3).

Table 3

Thematic Results – Challenges of Big Data Use in EI

Challenge	Articles	Brief Description of Theme
Data quality	Beckman et al. (2017); Coulton et al. (2015); Mader (2015); Mead & Mitchel (2016); Piety (2014); Pugh & Foster (2014)	Quality issues in EI administrative data
Data complexity	Blair, (2016); Bone et al. (2015); Coulton et al. (2015); Fischer, Anthony, & Dorman (2014); Gilmore (2016); Mead & Mitchel (2016); Payakachat et al. (2016)	Big data requires special considerations and knowledge for locating, storing, understanding, and utilizing
Limited scope	Blair (2016); Cohen & Nonacs, (2016); Gilmore (2016); Mader (2015)	Big data is primarily limited to observational data, therefore limiting its ability to identify “what works”
Lack of data driven	Blair (2016); Coulton et al. (2015); Fischer, Anthony, & Dorman (2014); Mead & Mitchel (2016)	Overall difficulty with data literacy on all levels in EI, lack of training and supports for local EI systems to access and use big data, and

history/culture in EI		lack of readily available big data pertinent to EI
Ethical challenges	Beckman et al. (2015); Coulton et al. (2015); Gilmore (2016); Currie (2013); Mader, (2015); Montgomery (2015); Petrilla (2014); Wrigley & Wormwell (2016)	Ethical challenges include: 1) Privacy/reidentification, 2) Informed consent difficulties, 3) Risk of datafication and schoolification, 4) Risk of tracking or grouping children by ability groups
Legal considerations	Blair (2016); Currie (2013); Gilmore (2016); Petrilla (2014); Zeide (2015)	Legal considerations potentially complicating big data use and access include: 1) Differing state laws regarding access to administrative data, 2) Conflicting Federal laws governing data access and privacy - FERPA HIPPA, 3) Data ownership issues (business versus public) limiting access to particular groups, 4) Privacy laws do not thoroughly address current day needs in a digital society, 5) the need to maintain balance between privacy protections and data access (Blair, 2016)
Structural problems	Derrington, & Cochenour (2015); Fischer, Anthony, & Dorman (2014); Gilmore (2016); Mader (2015); Mead & Mitchell (2016); Pugh & Foster (2014)	Core systems level issues related to EI big data use include: 1) Siloed EI data systems, 2) Lack of funding to support data systems, 3) Current focus of EI administrative data falls heavily on accountability versus usability

Knowledge and Skills. Five core themes related to knowledge and skills emerged which included knowledge and training in data science, research methodology, domain knowledge, collaboration, and knowledge of data sharing and privacy laws.

Table 4

Thematic Results – Knowledge and Skills Needed for Big Data Use in EI

Knowledge/Skills Needed	Articles	Brief Description of Theme
Data science	Bone et al. (2015); Coulton et al. (2015); Gilmore (2016); Terzieva, Todorova, & Kademova-Katzarova (2015); Petrilla (2014)	Machine learning, educational data mining (EDM), text mining, data extraction, ability to combine large scale data sets, understand multiple formats of big data, data cleaning and wrangling, data visualization, data programming/analysis tools, knowledge of advanced statistics

Research methodology & methods	Falster et al. (2015); Mader (2015); Montgomery (2015); Terzieva, Todorova, & Kademova-Katzarova (2015)	Understanding of quantitative, qualitative, and mixed methods to ensure accurate and ethical application to developmental data
Domain knowledge in area of study	Bone et al. (2015); Coulton et al. (2015); Gilmore (2016); Mader (2015); Montgomery (2015); Wrigley & Wormwell (2016)	Understanding of development and specialized area in EI is key to ensuring understanding variables and ethical conclusions
Collaboration	Bone et al. (2015); Coulton et al. (2015); Derrington, & Cochenour (2015); Falster et al. (2015); Fischer, Anthony, & Dorman (2014); Piety (2014); Pugh, & Foster (2014); Wrigley & Wormwell (2016)	Collaboration with other agencies and professionals is necessary to ensure domain knowledge, privacy, and data understanding
Data sharing and privacy law	Petrilla (2014)	Ability to understand and correctly apply sharing and privacy laws related to big data use is needed due to access to individual level data

Systems/Structural Level Needs. Seven core themes emerged related to necessary systems level/structural supports, including linked data, privacy safeguards, interdisciplinary/interagency collaboration, funding, integrated data systems (IDS)/data repositories, support/technical assistance for data use, and federal policy needs.

Table 5

Thematic Results – Structural/Systems Level Needs for Big Data Use in EI

Systems Needs	Articles	Brief Description of Theme
Linked data	Brundage et al. (2016); Currie (2013); Coulton et al. (2015); Derrington, & Cochenour (2015); Fischer, Anthony, & Dorman (2014); Mader (2015); Payakachat et al. (2016); Wrigley & Wormwell (2016)	In order to be able to examine longitudinal and broad issues, data must be linked across time and systems requiring: 1) Common measures, 2) Unique IDs (UIDs), 3) Well-defined variables, 4) Child development considerations built into system, 5) Culture and community considerations, 6) Continuous data collection
Privacy safeguards	Coulton et al. (2015); Currie (2013); Falster et al. (2015); Montgomery (2015); Zeide (2015)	Privacy safeguards must address: 1) Guidelines from local to federal levels, 2) State level IRBs in addition to research institutions, 3) Consent considerations to avoid reidentification, 4) Creation and use of centers for linking

		population level data, 5) Utilization of data custodians to ensure separation principle, 6) Data linkage and data merging should be two separate processes, 7) Contractors should be minimized to decrease risk of privacy breaches and ensure quality.
Interdisciplinary and interagency collaboration	Bone et al. (2015); Coulton et al. (2015); Derrington, & Cochenour (2015); Falster et al. (2015); Fischer, Anthony, & Dorman (2014), Piety (2014); Pugh, & Foster (2014); Wrigley & Wormwell (2016)	Three main areas related to collaboration included: Stakeholder groups, research practice partnerships, transdisciplinary or dually trained in EI and data science
Funding	Fischer, Anthony, & Dorman (2014); Mead & Mitchel (2016); Petrila (2014)	Adequate funding is needed for big data storage and for trained staff to analyze big data
Integrated data systems (IDS)/repositories	Beckman, Plummer, Munson, & Reidy (2017); Brundage, Corcoran, Wu, & Sturgill (2016); Coulton et al. (2015); Payakachat, Tilford, & Ungar (2016)	IDS and repositories must involve consistent and mechanized data collection to optimize accuracy and ensure ongoing growth of data
Data support/technical assistance	Mead & Mitchel (2016)	Data support and technical assistance is needed to ensure adequate gathering, sharing, and use of data
Federal Policies	Currie (2013); Coulton et al. (2015); Mead & Mitchell (2016); Petrila (2014); Piety et al. (2014); Zeide (2015)	Federal policies must require: 1) Data collection focused on meaningful outcomes, 2) Procedures for research related data access in all states, 3) Balance of risk/benefit when developing procedures and policies

Discussion

Many critical discussion points pertinent to the use of big data in EI emerged from the scoping review related to not only critical characteristics of the data but also the core themes related to strengths, challenges, necessary knowledge/skills, and systems-level supports.

Discussion of Quantitative Results

Search source, literature class, and fields of study. Results indicate that information on big data use in EI is being published across many classes of literature, underscoring the importance of extending searches across all literature classes when examining an emerging topic (See Figure 2). Perhaps the most interesting finding related to where information is being published on big data use in EI involved the search engines used to locate information. Google Scholar was the main source of identification of all classes of literature that met inclusion for this review (See Figure 3). This is important to note because most systematic reviews do not utilize Google Scholar and primarily focus on traditional database searches. If that approach had been taken in this search, only fifty percent of the articles would have been included in this review, significantly limiting the findings and strength of this review. Overall, this finding supports the assertion that research questions should drive the decisions for all aspects of review, including examination of which search engines should be used. Because missed information is a source of bias in studies, this should also be a consideration when conducting a systematic review.

The types of articles identified in this review demonstrated wide variation and included both empirical and non-empirical articles. Articles included traditional peer-reviewed empirical articles, conceptual works, policy papers, and law articles. Most notable was the relatively limited amount of empirical research, with only 25% of the included articles being empirical research (See Figure 3). In addition, despite the growing recognition that the use of mixed methods may optimize internal and external validity (Glimore, 2016; Mader, 2015), no empirical studies identified in this review employed mixed methods. Clearly, there is room for continued growth and development of methods and design when using big data in EI. Communities of

support and innovation, particularly from interdisciplinary fields, will be needed to solve some of the biggest challenges and to drive innovation in this area.

Finally, few studies reported data privacy procedures or linkage rates (29%). In addition to issues related to protecting subjects, low linkage rates essentially represent missing data, which is a potential source of bias if not properly accounted for. Therefore, it is recommended that both data privacy procedures and linkage rates become necessary reporting pieces in scholarly work utilizing big data.

Tools for Analyzing and Storing. Less than half of the empirical studies listed the software programs and/or tools utilized (43%). Using the right tools in data analysis and storage, as with any type of tool use, ensures the most accurate and efficient analysis and outcome. We also argue that using up to date, open-source data tools (e.g. R, Python) and platforms (e.g. GitHub), can increase collaboration and replication practices—both of which are practices being recommended in science and research (e.g. McNutt, 2014, Miguel et al., 2014). Therefore, in order to practice replicable research, and to ensure accurate and efficient data analysis, descriptions of the tools and programs for analysis and research should be included in all articles. In addition, programming skills and coding collaboration have not historically been a focus of EI and educational research programs. The lack of noted statistical programming tool use within this review exemplifies this point. As we head into a new age of data, it will become imperative that these are foundational skills for future researchers in EI and education.

Discussion of Qualitative Results.

Strengths. There were many strengths identified in the themes related to the use of big data in EI systems as noted in the results (See Table 2). Perhaps the most notable theme for discussion included the decreased risk and exposure of subject groups. This is particularly

pertinent in vulnerable populations such as young children and individuals with disabilities. Ultimately, analysis of existing data should be considered before exposing new subject groups through traditional research. Perhaps this could be part of external-funding or IRB criteria; specifically, funders and review boards could ask applicants whether the research question(s) could be answered through use of, or in combination with, big data that are already available in order to decrease subject risk.

Another notable theme identified in the literature was the capability of big data to capitalize on larger sample sizes through combining datasets. Small sample sizes are a common barrier in EI research due to the individualized nature of interventions and specialized developmental issues investigated (Blair, 2016). Therefore, having capacity to examine key questions on a larger scale opens new doors to innovation and intervention. Specifically, a larger sample size or use of population-level data allows for improved ability to detect effects, to undergo analysis specific to specialized groups or areas, and to gain greater control of covariates (Blair, 2016; Currie, 2013; Falster et al., 2015; Fischer, Anthony, & Dorman, 2014; Mader, 2015). This ultimately opens the door to begin exploring key questions, such as what intervention holds potential to impact certain groups and under what conditions, as has been called for in the EI literature (Center on the Developing Child at Harvard University, 2016).

Lastly, a strength of big data in EI is that there are already integrated data systems (IDS) and repositories available with information specific to early development and children with disabilities. Specifically, speech and autism researchers have begun to develop an IDS (Beckman, Plummer, Munson, & Reidy, 2017; Brundage, Corcoran, Wu, & Sturgill, 2016; Payakachat, Tilford, & Ungar, 2016). EI could learn from these systems and consider the expansion of data repositories related to issues or areas of development that impact young

children and families to address issues across a broader range of developmental needs and diagnoses. The frameworks for aligned data terminology set forth by the Federal Department of Education on Common Education Data Standards (CEDS) could be built upon to help facilitate the alignment and usability of these repositories and IDS.

Challenges. Several challenges were identified related to use of big data in EI. (See Table 3) Perhaps the most notable challenge, given recent efforts to link and align data system in EI, are related to data linkage. Although in recent years there has been increased attention to and support for aligning administrative data systems in EI and education as a whole (e.g. DaSy Center, 2014; Institute of Educational Sciences, 2017), there remains a significant amount of work to be done in this area. In order to advance big data exploration, there needs to not only be data alignment in EI and educational systems but also a paradigm shift to regularly include birth to three age groups in educational taxonomies and data linkage. Currently, it is common to include only preschool age and beyond when discussing data linkage in early education systems (e.g. Piety, Hickey, & Bishop, 2014), ultimately leaving out one of the youngest and most vulnerable groups in our educational system. The aforementioned CEDS data terminology alignment efforts, for example, start with preschool instead of including the EI system and infants and toddlers from birth, stating that their mission is to “streamline the understanding of data within and across P-20W institutions and sectors” (retrieved from: <https://ceds.ed.gov/whatIsCEDS.aspx>).

The ethical and legal challenges of big data is another substantial challenge. Specifically, there are privacy concerns due to the potential for re-identification of the individual, given the individual level needs in developmental/EI research (Coulton et al., 2015; Gilmore, 2016; Currie, 2013; Mader, 2015; Petrilla, 2014). Therefore, added procedures and protections should be

considered when combining sensitive data. Here, the field can learn from global partners, such as Australia, who have put steps into place to avoid re-identification. For example, Australian data scientists employ methods such as utilization of specially trained third-party agencies for completion of the data merging process to decrease risk of re-identification while still allowing access to data (Smithers, Searle, Chittleborough, Scheil, Brinkman, & Lynch, 2015). Reporting of these data privacy procedures when using big data should also be considered part of best practice methods in research with big data. Within this review, only 29% of the empirical articles reported procedures related to data protection processes. Privacy is of the utmost concern when it comes to any research or data analysis. At the same time, the balance between privacy protections and data access must be considered (Blair, 2016). There is immense potential for new findings and advancements in treatments and services through analysis of big data (Currie, 2013; Falster et al. et al., 2015; Payakachat, 2016). If we limit data access too much, potentially life-changing findings will go undiscovered and difficult conversations will need to be had surrounding the benefits of data access relative to the risks for participants.

Systems level supports. As noted in the results, seven core themes emerged related to necessary systems level/structural supports. Perhaps most notable was the need for ongoing training and technical assistance to improve data quality and linkages (Mead & Mitchel, 2016). Although some supports are available on the federal level (e.g. DaSy Center, Early Childhood Data Initiative, Zero to Three, Early Childhood Technical Assistance), these efforts have been mostly focused on alignment and governance of administrative data in EI (e.g. DaSy Center, 2014, U.S. Department of Health and Human Services & the U.S. Department of Education, 2016). Therefore, next steps may be to broaden the array of training and technical support not only related to creating aligned systems, but also to provide supports for local service areas to

better access and use large-scale data (Mead & Mitchel, 2016). This will be particularly important if we want to move from mere data accountability in EI to full-scale data usability for practice improvement on all levels.

Next, although administrative data are an important piece of understanding, and there have been many advances in this area in the last decade (U.S. Department of Health and Human Services & the U.S. Department of Education, 2016), administrative data is only one piece of the big data puzzle. As noted earlier, there are many data sources contributing to the massive troves of data available for analysis. However, beginning to link and integrate data from other sources can be complicated, often necessitating interdisciplinary and interagency collaboration. For example, in order to combine datasets from education, health, and business, three different professional languages must be integrated and aligned, requiring cross training and collaboration. In addition, the complexities of differing funding, regulations, policies, and laws for each field, must be understood, again requiring added alignment and problem-solving between disciplines and fields. Overall, as systems continue to progress, so too must our processes and scope when considering systems-level needs related to data use. To advance our understanding of early development and harness the potential of big data in EI, the next step will be to move beyond just the use of administrative data in EI and truly engage all facets of big data.

Knowledge and Skills. Themes in the literature highlight the multifaceted nature of skills needed for the ethical and optimal use of big data in EI. Specifically, specialized knowledge in research, data science, methodology and methods, early development, and EI systems is needed. This will require collaboration between researchers, statisticians, data scientists, and EI domain knowledge specialists to optimize new methods and methodology for

analyzing and answering questions related to early development in EI (Blair, 2016; Coulton et al., 2015; Petrilla, 2014).

Finally, the need for domain knowledge in EI when analyzing big data appeared consistently in the literature (Mader, 2015; Bone et al., 2015; Coulton et al., 2015; Gilmore, 2016; Wrigley & Wormwell, 2016). Authors have noted the risk of young children being put into groups or tracks due to predictions from aligned data, which could impact long-term ability to access resources and obtain jobs if the data are not positive (Gilmore, 2017; Montgomery, 2015). EI professionals have argued for years against the “schoolification” of young age groups while emphasizing the need for developmentally appropriate practices (Division for Early Childhood, 2014). When utilizing big data for analysis of issues related to young children, this again becomes a risk. Therefore, when setting up data systems, considerations should be made to allow time for collaboration, knowledge sharing, and the assurance that domain knowledge in EI is part of all data processes.

Limitations of the Scoping Review

Although care was taken to identify and utilize an optimal methodology, there are some limitations of this review. First, although the Google Scholar search engine produced the greatest number of articles included in the review, thousands of hits were accrued with only a small percentage being relevant. Therefore, a data saturation approach was taken and only the first 100 results were reviewed. This procedure could have potentially resulted in missed information.

Second, as with other studies, consistency of terminology is a limitation in searches. Without consistency of search terms, there is always a risk of missing potentially valuable research and information. Precautions were taken to minimize this risk and provide accurate searches, including consulting with experts in the field, engaging in pilot searches, and hand

searching key words of relevant articles. Yet, a risk of missed studies exists due to inconsistency of key words and search terms, which is particularly notable when examining birth to three Part C of IDEA early intervention services for infants and toddlers with disabilities, a population in which there were no consistent search terms to access literature. Last, although a highly systematic approach was taken in this review, its replicability may be impacted by the snowball and data saturation techniques used to identify articles.

Future Directions

Given the emergent nature of the use of big data in EI, there remain large gaps and needs related to areas of future research in order to increase innovation and ensure ethical practices. In the following, we focus on a few of the most striking gaps that emerged during this review.

Knowledge and Skills. Numerous questions warrant consideration regarding knowledge and skills related to the use of big data in EI. For example, it would be beneficial to gain greater understanding of what skills researchers in EI are gaining both in preservice and as new researchers in the field related to the analysis and use of big data. This may lead to recommendations for training programs and future use of big data in EI. Also, it would be beneficial to understand the ethical requirements of data scientists related to acquiring domain knowledge as part of protocols in big data analysis, particularly related to vulnerable groups such as EI. As the field of data science continues to grow with many universities rapidly adding degree programs focused specifically on the nuances of data science, this may be beneficial knowledge for program development.

Data Tools. New data tools will continue to emerge as technology evolves. Therefore, it is important to examine what tools work best to analyze developmental EI data and what data tools allow for collaboration in EI settings. Additionally, better understanding of what the data

storage options are that meet requirements for confidentiality and yet allow growth for new data sources such as wearable technology and health informatics as technology continues to advance would be helpful for IDS development in EI.

Systems Supports. Perhaps one of the largest areas of needed research is related to data access and quality. As identified in the review, data quality is a challenge in EI. Therefore, it would be helpful to begin delving into this area of research. First, it would be helpful to have a foundational understanding of what data sources are available in EI and where these are located including procedures for access. This access and understanding continues to improve as we speak with efforts such as a new search engine through Google for datasets. It functions similar to how Google Scholar works but for datasets (Google Dataset Search, 2018). This understanding includes identifying not only administrative data sources (e.g. Child Outcomes Summary data, Assessment data, Part C reporting measures) but also other sources of big data (e.g. IDS, research repositories, social media, search data, GPS data, provider lists, pertinent wearable technology data, health data). After all, this information will be necessary for moving from accountability to data use for practice improvement in EI.

Relationships. Relationships and collaboration are under-studied in general and to date, there is no known empirical research on collaboration related to big data practices in EI. Although some work on systems governance and alignment has started to emerge related to data systems in EI and early childhood systems (e.g. DaSy Center, 2014; U.S. Department of Health and Human Services & the U.S. Department of Education, 2016), there remain many questions about the collaborative elements of these practices. Therefore, it will be important to answer questions such as, what levels of collaboration in EI service areas and/or state systems are present and what key ingredients are necessary for successful interdisciplinary and interagency

collaboration related to data sharing, linking, and organizing? Similarly, what systems-level supports are needed to promote and sustain data partnerships (e.g. memorandums of understandings, formalized task force, collaborative councils)? What are the cultural implications of current data practices and how do we promote the involvement of and collaboration with representatives from all groups when collecting, analyzing, and drawing conclusions from big data? Finally, in order to have true data translation and alignment across systems, EI needs to be incorporated into all frameworks and efforts. For example, the aforementioned CEDS currently focuses on preschool and beyond. It would be helpful if future research and efforts began to incorporate and align data efforts starting with EI/at birth or prenatally.

Conclusion

We are entering a new phase in practice and data use in EI research and systems with the development of integrated data systems and new sources of large scale data. With this new phase comes both exciting opportunities and considerable challenges. New findings in EI research are possible because of the increased amounts and types of data being combined and analyzed, but these new avenues come with increased potential for data misuse, privacy breaches, and inappropriate application of findings. Therefore, it is imperative for researchers, leaders, and practitioners in EI to continue to grow the knowledge base on big data use in EI to ensure ethical and accurate practices with some of our youngest and most vulnerable populations. Ultimately, it is the hope that the results, discussion, and suggestions for future directions and research provided herein will spur thoughts, conversations, research, and further innovation related to data practices and the use of big data in EI. Big data is the new frontier in our modern society; are we ready in EI to harness its potential?

References

- Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of gray: guidelines for working with the gray literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, 19(4), 432-454.
- Adjerid, I., & Kelley, K. (2018, February 22). Big Data in Psychology: A Framework for Research Advancement. *American Psychologist*. Advance online publication. <http://dx.doi.org/10.1037/amp0000190>
- Ahmadi, M., Dileepan, P., & Wheatley, K. K. (2016). A SWOT analysis of big data. *Journal of Education for Business*, 91(5), 289-294.
- Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19-32.
- Azzam, N., Oei, J. L., Adams, S., Bajuk, B., Hilder, L., Mohamed, A. L., Wright, I.M., & Holland, A. J. (2017). Influence of early childhood burns on school performance: an Australian population study. *Archives of Disease in Childhood*, archdischild-2017.
- Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed Research International*, 2015.
- Beckman, M. E., Plummer, A. R., Munson, B., & Reidy, P. F. (2017). Methods for eliciting, annotating, and analyzing databases for child speech development. *Computer Speech & Language*, 45, 278-299.
- Blair, L. M. (2016). Publicly Available Data and Pediatric Mental Health: Leveraging Big Data to Answer Big Questions for Children. *Journal of pediatric health care: official publication of National Association of Pediatric Nurse Associates & Practitioners*, 30(1), 84-87.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2015).

Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, 45(5), 1121-1136.

Boone-Heinonen, J., Tillotson, C. J., O'Malley, J. P., Cottrell, E. K., Gaudino, J. A., Amofah, A.,

Brickman, A., Mayer, K., McBurnie, M. & Gold, R. (2017). Characterizing a big data cohort of over 200,000 low-income us infants and children for obesity research: The advance early life cohort. *Maternal and Child Health Journal*, 21(3), 421-431.

Brawley, S., & Stormont, M.A. (2014). Investigating reported data practices in early childhood:

An exploratory study. *Journal of Positive Behavior Interventions*, 16(2), 102-111.

Brundage, S. B., Corcoran, T., Wu, C., & Sturgill, C. (2016, May). Developing and using big

data archives to quantify disfluency and stuttering in bilingual children. In *Seminars In Speech And Language* (Vol. 37, No. 02, pp. 117-127). Thieme Medical Publishers.

Bunn, J. (2012, September). How big is a Petabyte, Exabyte, Zetabyte, or Yottabyte? High

Scalability in Databases. Retrieved from: <http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html>

Capatosto, K. (2017). Foretelling the Future: A Critical Perspective on the Use of Predictive

Analytics in Child Welfare. Retrieved from: <http://kirwaninstitute.osu.edu/wp-content/uploads/2017/05/ki-predictive-analytics.pdf>

Carpenter, K. L., Sprechmann, P., Calderbank, R., Sapiro, G., & Egger, H. L. (2016).

Quantifying risk for anxiety disorders in preschool children: A machine learning approach. *PloS one*, 11(11), e0165524.

Center on the Developing Child at Harvard University (2016). *From Best Practices to*

Breakthrough Impacts: A Science-Based Approach to Building a More Promising Future

For Young Children And Families. Retrieved from:

<http://www.developingchild.harvard.edu>

Chen, M., Ma, Y., Song, J., Lai, C. F., & Hu, B. (2016). Smart clothing: Connecting human with clouds and big data for sustainable health monitoring. *Mobile Networks and Applications*, 21(5), 825-845.

Cohen, L. S., & Nonacs, R. (2016). Neurodevelopmental Implications of Fetal Exposure to Selective Serotonin Reuptake Inhibitors and Untreated Maternal Depression: Weighing Relative Risks. *JAMA Psychiatry*, 73(11), 1170-1172.

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., Kastner, M., & Moher, D. (2014). Scoping reviews: time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology*, 67(12), 1291-1294.

Colquhoun, H.L. (2016). Current best practices for conducting reviews [PPT]. Impactful Biomedical Research: Achieving Quality and Transparency. Retrieved from: http://www.equator-network.org/wp-content/uploads/2016/06/Gerstein-Library-scoping-reviews_May-12.pdf

Columbus, L. (2017, March). IBM predicts need for data scientists will grow by 28% by 2020. *Forbes*. Retrieved from: <https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#2aff448f7e3b>

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, 1-12.

Coulton, C. J., Goerge, R., Putnam-Hornstein, E., & de Haan, B. (2015). *Harnessing big data for social good: A grand challenge for social work* (No. 11). Working Paper.

- Crawford, K., Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms, 55 B.C. L. Rev. 93.
- Creswell, J. W., & Poth, C. N. (2017). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.
- Currie, J. (2013). big data versus “big brother”: on the appropriate use of large-scale data collections in pediatrics. *Pediatrics*, 131(Supplement 2), S127-S132.
- DaSy Center. (2014). DaSy data systems framework. Menlo Park, CA: SRI International.
- Derrington, T., & Cochenour, M. (2015, December). Big data for little kids: The importance and utility of early childhood integrated data systems. Poster presentation at Zero to Three National Training Institute. Retrieved from: <http://dasycenter.org/big-data-for-little-kids-the-importance-and-utility-of-early-childhood-integrated-data-systems/>
- Division for Early Childhood. (2014). DEC recommended practices in early intervention/early childhood special education 2014. Retrieved from: <http://www.dec-sped.org/recommendedpractices>
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1).
- Dykes, B. (2017, June 28). *Big data, forget volume and variety, focus on velocity*. Forbes. Retrieved from: <https://www.forbes.com/sites/brentdykes/2017/06/28/big-data-forget-volume-and-variety-focus-on-velocity/#578ad3d26f7d>
- Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).
- Falster et al., K., Jorm, L., Eades, S., Lynch, J., Banks, E., Brownell, M., Kraven, R., Einarsdottir, K., & Randall, D. (2015). What factors contribute to positive early childhood health and development in Australian Aboriginal children? Protocol for a population-based

- cohort study using linked administrative data (The Seeding Success Study). *BMJ open*, 5(5), e007898.
- Finfgeld-Connett, D. (2014). Use of content analysis to conduct knowledge-building and theory-generating qualitative systematic reviews. *Qualitative Research*, 14(3), 341-352.
- Fischer, R., Anthony, B., & Dorman, R. (2014). Big data in early childhood: Using integrated data to guide impact. Presented at Zero to Three National Training Institute.
- Gilmore, R. O. (2016). From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2), 112-126.
- Google (2018). Dataset search. Retrieved from: <https://toolbox.google.com/datasetsearch>
- Grant, C., & Osanloo, A. (2014). Understanding, selecting, and integrating a theoretical framework in dissertation research: Creating the blueprint for your “house”. *Administrative Issues Journal*, 4(2), 4.
- Hao, B., Li, L., Li, A., & Zhu, T. (2013, July). Predicting mental health status on social media. In *International Conference on Cross-Cultural Design* (pp. 101-110). Springer, Berlin, Heidelberg.
- Hebbeler, K., Spiker, D., Bailey, D., Scarborough, A., Mallik, S., Simeonsson, R., Singer, M., & Nelson, L. (2007). Early intervention for infants and toddlers with disabilities and their families: Participants, services, and outcomes. SRI International, Menlo Park, CA.
- Individuals with Disabilities Education Act, 20 U.S.C. §1400 (2004).
- Institute of Educational Sciences (2017). Best practices brief: P-20W + data governance tips from the States. Retrieved from: <https://nces.ed.gov/programs/slds/publications.asp#28>
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science*, 5(1), 69.

Lohr, S. (February, 2013). "The origins of big data." *New York Times*.

Lucas, A., Hurth, J., & Kasprzak, C. (2010). Essential Elements of High Performing, High Quality Part C Systems. NECTAC Notes No. 25. *National Early Childhood Technical Assistance Center (NECTAC)*.

Mader, N. (2015). The big data era and an integrated mode of inquiry for social policy-relevant research. *ISJLP*, 11, 97.

Mashey, J. (1998). Big data and the next wave of infrastress. Retrived from:

http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

McAfee, A., & Brynjolfsson, E. (2012, October). Big data: The management revolution. *Harvard Business Review*. Retrieved from <http://hbr.org/2012/10/big-data-the-management-revolution/ar>

McNutt, M. (2014). Reproducibility. *Science*, 343(6168), 229. DOI: 10.1126/

Mead, S., & Mitchel, A. L. (2016). Moneyball for Head Start: Using Data, Evidence, and Evaluation to Improve Outcomes for Children and Families. *Bellwether Education Partners*.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Green, P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B., Persen, M., Sedlmayr, R., Simmons, J., Siimonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343(6166), 30-31.

Mitchell, C., Lawton, E., Morton, C., McCain, C., Holtby, S., & Main, E. (2014). California pregnancy-associated mortality review: mixed methods approach for improved case identification, cause of death analyses and translation of findings. *Maternal and child health journal*, 18(3), 518-526.

- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7), e1000097.
- Moher, D., Stewart, L., & Shekelle, P. (2015). All in the family: Systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Systematic reviews*, 4(1), 183.
- Montgomery, K. (2015). Children's media culture in a big data world. *Journal of Children and Media*, 9(2), 266-271.
- Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13), 1351-1352.
- Odgers, C. L., Caspi, A., Bates, C. J., Sampson, R. J., & Moffitt, T. E. (2012). Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method. *Journal of Child Psychology and Psychiatry*, 53(10), 1009-1017.
- O'Leary, J. D., Janus, M., Duku, E., Wijeyesundera, D. N., To, T., Li, P., ... & Crawford, M. W. (2016). A population-based study evaluating the association between surgery in early life and child development at primary school entry. *The Journal of the American Society of Anesthesiologists*, 125(2), 272-279.
- Payakachat, N., Tilford, J. M., & Ungar, W. J. (2016). National Database for Autism Research (NDAR): big data opportunities for health services research and health technology assessment. *Pharmacoeconomics*, 34(2), 127-138.
- Peters, M., Godfrey, C., McInerney, P., Soares, C., Hanan, K., & Parker, D. (2015). The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews.
- Petrila, J. (2014). Data Integration for Social Policy: Challenges and Opportunities. *What Counts: Harnessing Data for America's Communities*, 190-204.

- Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014, March). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 193-202). ACM.
- Plano Clark, V., & Ivankova, N. (2016). *Mixed Method Research: A Los Angeles*.
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51-59.
- Pugh, K., & Foster, G. (2014). Australia's National School Data and the 'Big Data' Revolution in Education Economics. *Australian Economic Review, 47*(2), 258-268.
- Ritz, J., & Knaack, Z. (2017). Internet of Things. *Technology & Engineering Teacher, 76*(6).
- Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: opportunities and policy implications. *Health affairs, 33*(7), 1115-1122.
- Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton University Press.
- Smithers, L. G., Searle, A. K., Chittleborough, C. R., Scheil, W., Brinkman, S. A., & Lynch, J. W. (2015). A whole-of-population study of term and post-term gestational age at birth and children's development. *BJOG: An International Journal of Obstetrics & Gynaecology, 122*(10), 1303-1311.
- Stingone, J. A., Pandey, O. P., Claudio, L., & Pandey, G. (2017). Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. *Environmental Pollution, 230*, 730-740.
- Terzieva, Todorova, & Kademova-Katzarova, V., Todorova, K., & Kademova-Katzarova, P. (2015). *Big Data—Opportunities and Challenges for Education*.

- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., Levac, D., Ng, C., Pearson Sharpe, J., Wilson, K., Kenny, M., Warren, R., Wilson, C., Stelfox, H., & Straus, S. (2016). A scoping review on the conduct and reporting of scoping reviews. *BMC medical research methodology*, 16(1), 15.
- Tseng, S., Liu, K., & Wang, W. L. (2011). Moving toward being analytical: A framework to evaluate the impact of influential factors on interagency collaboration. *Children and Youth Services Review*, 33(6), 798-803.
- U.S. Department of Health and Human Services & the U.S. Department of Education (2016). The integration of early childhood data. Retrieved from:
<https://www2.ed.gov/about/inits/ed/earlylearning/files/integration-of-early-childhood-data.pdf>
- Winston, J. (2016, November). How the Trump campaign built an identify database and used Facebook ads to win the election. StartupGrind. Retrieved from:
<https://medium.com/startup-grind/how-the-trump-campaign-built-an-identity-database-and-used-facebook-ads-to-win-the-election-4ff7d24269ac>
- Wrigley, T., & Wormwell, L. (2016). Infantile accountability: When big data meet small children. *Improving Schools*, 19(2), 105-118.
- West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. *Governance studies at Brookings*, 4, 1-0.
- Wolgemuth, J. R., Hicks, T., & Agosto, V. (2017). Unpacking Assumptions in Research Synthesis: A Critical Construct Synthesis Approach. *Educational Researcher*, 46(3), 131-139.

- Zarate, M. (2018, July 13). Congress weighs pilot program for predictive analytics use in child welfare. Retrieved from: <https://chronicleofsocialchange.org/child-welfare-2/congress-pilot-program-predictive-analytics-child-welfare>
- Zeide, E. (2015). Student Privacy Principles for the Age of Big Data: Moving Beyond FERPA and FIPPs. *Drexel L. Rev.*, 8, 339.
- Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1436-1444). ACM.