Positive deviance, big data, and development: A systematic literature review

Article *in* Electronic Journal of Information Systems in Developing Countries · October 2018

DOI: 10.1002/isd2.12063

2 authors:

Basma Albanna
The University of Manchester

Richard Heeks
The University of Manchester

Received: 4 October 2018 | Accepted: 4 October 2018
DOI: 10.1002/isd2.12063

RESEARCH ARTICLE

WILEY

Positive deviance, big data, and development: A systematic literature review

Basma Albanna | Richard Heeks

Centre for Development Informatics, University of Manchester, Manchester, UK

Correspondence
Basma Albanna, Centre for Development Informatics, University of Manchester, Manchester, UK.
Email: basma.albanna@manchester.ac.uk

Abstract

Positive deviance is a growing approach in international development that identifies those within a population who are outperforming their peers in some way, eg, children in low-income families who are well nourished when those around them are not. Analysing and then disseminating the behaviours and other factors underpinning positive deviance are demonstrably effective in delivering development results. However, positive deviance faces a number of challenges that are restricting its diffusion. In this paper, using a systematic literature review, we analyse the current state of positive deviance and the potential for big data to address the challenges facing positive deviance. From this, we evaluate the promise of "big data-based positive deviance": This would analyse typical sources of big data in developing countries—mobile phone records, social media, remote sensing data, etc—to identify both positive deviants and the factors underpinning their superior performance. While big data cannot solve all the challenges facing positive deviance as a development tool, they could reduce time, cost, and effort; identify positive deviants in new or better ways; and enable positive deviance to break out of its current preoccupation with public health into domains such as agriculture, education, and urban planning. In turn, positive deviance could provide a new and systematic basis for extracting real-world development impacts from big data.

KEYWORDS

big data, developing countries, machine learning, mobile data, positive deviance, systematic literature review

1 | INTRODUCTION

Many development practitioners continue to use a traditional "needs-based" approach to development, involving top-down identification of needs and problems, and the external imposition of solutions that meet those needs. This type of approach can work well in addressing specific technical challenges. But it works much less well where development requires learning and behavioural change by beneficiary groups, something which necessitates much greater knowledge of and engagement with beneficiary communities (Nel, 2018; Pascale, Sternin, & Sternin, 2010; Saïd Business School, 2010; Singhal, 2011). As a result, more bottom-up "asset-based" approaches have come into existence, which capitalize on a community's inherent assets and capabilities—including knowledge—in solving development problems. Positive deviance (PD) is one such asset-based approach. It is based on the observation that in every group or community, a few individuals use uncommon practices and behaviours to achieve better solutions to problems than their peers who face the same challenges and barriers (Pascale et al., 2010). Those individuals are referred to as "positive deviants" (PDs), and adopting their solutions on a wider basis is referred to as the PD approach.

The term "positive deviance" was first used in 1976 to describe a practical strategy for the design of food supplementation programmes in Central America, a strategy that was derived endogenously rather than exogenously through identifying dietary practices developed by mothers in low-income families who had well-nourished children (Wishik & Van Der Vynckt, 1976). The results of this study were not widely publicized, limiting uptake. It was not until the 1990s that PD started to be seen as a credible strategy for nutrition research and action, based on an accumulation of evidence of impact (Sternin, Sternin, & Marsh, 1997; Sternin, Sternin, & Marsh, 1998; Zeitlin, 1991). The 1990s also witnessed its first large-scale adoption in international development by Save the Children, which used PD as a strategy to reduce malnutrition in Vietnam, rehabilitating an estimated 50 000 malnourished children in 250 communities (Sternin, 2002). But it has only really started to attract attention in the 2000s, when Sternin and collaborators promoted PD more broadly as an asset-based approach for social change and demonstrated how it can be operationalized across a variety of development domains (Sternin & Choo 2000; Sternin 2002). Since the early 2000s, PD has been applied across multiple development domains, with public health being the most prominent.

As will be discussed in further detail below, PD tends to rely on in-depth primary data collection in identifying PDs and then community mobilization in disseminating and scaling successful practices. Identification is therefore time and labour intensive, with costs proportional to sample size (Felt, 2011; Lapping, Marsh et al. 2002; Marsh, Schroeder, Dearden, Sternin, & Sternin, 2004). As a result, PD has traditionally made use of relatively small-scale samples. Statistically and practically, this can make it harder to identify positive deviants, given their relative rarity (Marsh et al., 2004). It also limits the ability to accurately generalize the identified practices to larger populations (Marsh et al., 2004). Path dependency has also been evident in uptake of the PD approach, in terms of geographical distribution and domain of application, with most studies concentrated in a few countries of Asia and in addressing malnutrition: the region and domain where it was initially introduced and practised by Sternin et al. (1997).

Given these and other challenges and limitations, there are obvious opportunities for innovation in positive deviance. Our particular interest here is in the innovative opportunities offered by big data: the increasing amounts of data about what we are and what we do and what we say, generated from digital devices, which provide an opportunity to gather insights into human behaviour. If big data can provide insights into behaviour, then big data analytics could identify patterns of “abnormal behaviour”: variances from the average collective behaviour of observed units which could include the behaviours of those which a PD approach would define as positive deviants.

In this paper, we therefore investigate the potential of “big data-based positive deviance” (BDPD). Our particular interest flows from the line of argument above that there are challenges for the traditional positive deviance approach which big data might be able to address. But there is also a converse interest that positive deviance might represent a new approach to the extraction of development value from big data.

To investigate this potential, we have undertaken a systematic literature review (SLR) of the empirical applications of positive deviance and of big data in developing country contexts, in order to answer three questions. First, how is PD currently being applied in development? In particular, we seek to identify from the literature challenges in that application which new approaches might seek to address. Second, how is BD currently being applied in development? We investigate this particularly in light of the challenges to positive deviance identified earlier; but we also extract challenges in use of big data. Third, what development value might result from the combined use of BD and the PD approach? Here, we combine the findings of both literature reviews to address the interests expressed above: not only how big data can address PD challenges but also how PD might be a valuable approach to the use of big data in development.

The paper begins with a brief description of the method used in conducting the literature review, followed by three sections that answer each of the questions in turn: a presentation of the findings from our PD and then BD literature review before concluding with a discussion about big data-based positive deviance.

2 | SYSTEMATIC LITERATURE REVIEW METHODOLOGY

2.1 | Literature search and selection

A systematic review of the literature was conducted using an adaptation of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) protocol (Moher, Liberati, Tetzlaff, & Altman, 2009). The review included academic, peer-reviewed, English-language literature that reported empirical results using secondary or primary data sources from developing countries. The literature search was implemented using Google Scholar because (1) it is free and easy to access, making the SLR reproducible; (2) both PD literature and BD literature are multidisciplinary so it was important to use a nondisciplinary comprehensive base of literature; and (3) Google Scholar has the widest coverage of academic articles in comparison to other search engines and databases (Khabsa & Giles, 2014). The utilized search strings and strategy are summarized in Table 1.

To retrieve relevant studies, we used the **intitle** operator, which ensures that the title of the retrieved articles would include the words following the operator. We also used **AND** and **OR** Boolean operators to ensure the existence of key terms in the text of the articles, thus reducing the time required in screening irrelevant sources. For example, in the PD literature search, the words “positive deviance,” “positive deviant,” or “positive deviants” were used with the “intitle” operator, thereby targeting articles that have PD as the central theme. The words “study,” “empirical,” “practice,” “experimental,” “survey,” or “fieldwork” were used for the in-text search to restrict articles not providing empirical evidence. The same search strategy was used to retrieve the BD literature but with a simpler search string found to deliver the corpus of literature

TABLE 1 Google Scholar search strategy used for the literature search

Positive Deviance Literature	
Search String[a]	Str1: intitle:"Positive deviance" (study OR empirical OR practice OR experimental OR survey OR fieldwork) Str2: intitle:"Positive deviants" (study OR empirical OR practice OR experimental OR survey OR fieldwork) Str3: intitle:"Positive deviant" (study OR empirical OR practice OR experimental OR survey OR fieldwork)
Time period	1970-2017
Exclude	Patents, citations, and non-English results
Big Data Literature	
Search String	intitle:"Big data" ("developing countries"[b] AND application)
Time period	1998-2017[c]
Exclude	Patents, citations, and non-English results

[a]The first part of the search strings looks for terms in the title of a paper, the terms in brackets search within the text of the paper.

[b]The term "developing countries" was not used in the PD search as initial investigation suggested that the majority of studies were in those countries, and we could then manually exclude those that were outside scope. Conversely, it was used in the BD search as the majority of BD literature was not in developing countries, and so the term was seen to be a useful means to quickly narrow the search to more-relevant items. As noted below, however, the term itself was in practice rather narrow and served to omit a number of relevant studies which then had to be manually identified and included via backward snowballing.

[c]To the best of our knowledge, the term "big data" was first introduced in 1998 (Mashey, 1998), thus setting the boundary for the search period.

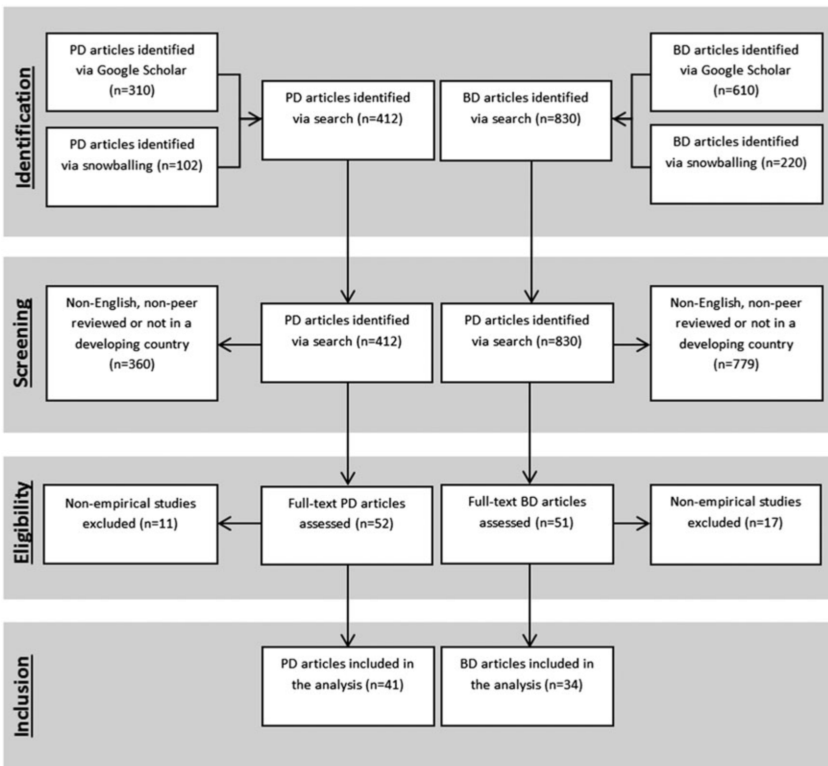

FIGURE 1 Flow diagram for identification and selection of positive deviance (PD) and big data (BD) articles (adapted from the PRISMA protocol)

suitable for analysis. To ensure that key studies were not excluded, backward snowballing[1] of relevant articles was employed, and it led to the identification of one additional PD article and 22 additional BD articles.[2] A total of 75 articles were included in the final corpus of analysis: 41 PD articles and 34 BD articles. Figure 1 reports on the identification and selection protocol.

[1]Backward snowballing involves screening the reference lists of relevant literature review articles to look for additional literature. It is considered an effective technique for conducting complex systematic literature reviews (Greenhalgh & Peacock, 2005).

[2]The number of BD articles identified by backward snowballing was much greater because many relevant big-data-for-development articles either do not specifically use the words "big data" in their title (eg, they use "mobile data" or "satellite images") or do not specifically use the words "developing countries" in their text (eg, they use the name of a particular country or a development goal).

2.2 | Content analysis

NVivo was used for the qualitative and quantitative content analysis of the selected articles. For each article in the PD and BD corpus, the following attributes were identified and used for classification: title, year of publication, research methodology, research approach, types of data used, sample area (ie, rural or urban), sampling unit, country, region, and study duration (if stated). Those attributes were derived based on a mix of commonly used data fields in systematic literature review and an iterative process of attribute selection depending on what arises as an important variable for the topic of analysis (Okoli, 2015; Petticrew & Roberts, 2005). Additionally, articles were coded into several nodes based on the areas covered in the qualitative analysis: those areas identified iteratively based on our overall purpose of understanding the potential development value of combining big data- and positive deviance-based analysis. Those areas can be summarized into (1) challenges and limitations, (2) benefits, (3) conceptual frameworks, (4) methods and data, (5) research findings, and (6) research opportunities. The following sections do not report all content analysis but only those main elements seen as relevant to the purpose of this paper.

3 | POSITIVE DEVIANCE

According to the Positive Deviance Initiative (Springer, Nielsen, & Johansen, 2016), the successful application of PD has been reported in more than 60 countries across the globe with a total outreach of more than 30 million individuals for the period between 1990 and 2016. Applications include reducing childhood malnutrition, enhancing school retention, eliminating neonatal mortality, limiting HIV transmission, improving salesforce productivity, fighting against female genital cutting, enhancing health care services, reducing transmission of antibiotic resistant bacteria in hospitals, and enhancing pregnancy outcomes. The central premise of the PD approach is that it harnesses the inherent wisdom of individuals existing within a community to develop solutions to their own problems. And since solutions come from the people, they take into account contextual and cultural variables, making them less vulnerable to social rejection. PD is also considered an efficient approach within international development, because it reduces reliance on aid and external expertise and instead capitalizes on local resources and know-how. It can also generate local engagement in identifying and disseminating practices and is seen as creating self-efficacy (individuals' belief in their capacity to execute behaviours necessary to achieve a desired objective), often considered to be a key influencer in the adoption of recommended behaviours (Babalola, 2007; Babalola, Awasum, & Quenum-Renaud, 2002).

Much of the positive deviance literature has a—forgiving the pun—rather positive, even proselytizing tone. Balancing this, there are some more critical insights with three particular concerns being raised.[3] First is a concern that—compared with its practical application—the ideas of positive deviance lack conceptual clarity, with papers using different definitions and with limited theorization of positive deviance (Herington & van de Fliert, 2018). Second is a concern that positive deviance does not always work in practice. Problems have included difficulties in identifying PDs (Marsh et al., 2004) and/or their differential characteristics and behaviours (Bradley et al., 2009; Felt, 2011) and inability to scale the PD solutions across a community and, particularly, between communities[4](LeMahieu, Nordstrum, & Gale, 2017). These concerns overlap significantly with material on the third area of concern: practical challenges to the implementation of positive deviance, a topic discussed further below as an outcome of the SLR. In sum, although, one may conclude that there has yet to be a weight of critique sufficient to discredit positive deviance as a development approach or to identify aspects necessary and inherent to PD that would undermine it. Conversely, there is a growing weight of evidence demonstrating beneficial development outcomes emerging from its application.

That application generally follows the five steps of the PD methodology, which can be outlined as follows (Positive Deviance Initiative, 2010):

1. Defining the problem and determining desirable outcomes;
2. Discovering PDs, ie, individuals or other social entities who unexpectedly achieved the desired outcomes;
3. Determining the underlying practices that led to those outcomes (this is known as positive deviance inquiry [PDI]);
4. Designing interventions to enable others to access and practice new behaviours; and
5. Monitoring and evaluating the PD intervention.

Building from this background on positive deviance, the systematic literature review begins with a timeline showing the volume of PD literature over the last two decades followed by a thematic classification of the literature. We then analyse the secondary data sources used in previous PD studies as these may share characteristics with attempts to use big data in PD. We then discuss the different units of analysis used in the literature before presenting the identified challenges of the PD approach, those challenges presenting potential opportunities for big data to make a contribution.

[3]Specifically within the field of social psychology, there has been opposition to the idea of positive deviance by those working on deviant behaviour, who wish to solely assign a negative connotation to deviance. However, such arguments do not transfer beyond the specific field of deviant studies and have, in any case, been fairly well refuted (Shoenberger, 2017).

[4]With an "orthodox" view of PD being that it should not seek to transfer solutions between communities, but develop them within each community (LeMahieu et al., 2017). Even if one accepts this view, it clearly depends on where one sets the definition and boundary of "community".

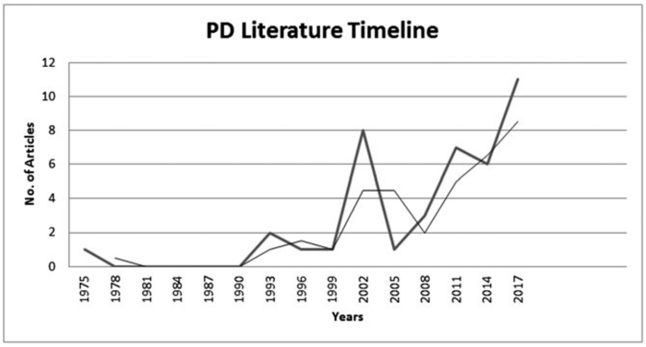

FIGURE 2 Timeline of reviewed positive deviance (PD) literature for period 1976 to 2017. The thick line represents the actual number of studies whereas the thin line represents a projection of the trend. The bold line represents the actual number of studies whereas the thin line represents a projection of the trend

3.1 | PD literature timeline

The publication timeline for work on positive deviance is shown in Figure 2, beginning with the first empirical PD study, published in 1976 (Wishik & Van Der Vynckt, 1976). This article published the impending methodology but did not publish results, and it was not until the early 1990s that the approach started to gain attention due to the book (1990) and study (1991) published by Zeitlin (1990) which provided extensive observations on the PD approach in nutrition with a strong emphasis on impact. There was a peak in the early 2000s, which we can owe to Sternin who operationalized the PD approach and published the results of its application in the Save the Children project which reduced malnutrition in Vietnam by 65% to 80% in 2 years (Sternin & Choo, 2000). This led to an extensive and rigorous evaluation of the PD strategy in solving child malnutrition, revealing positive results that supported its uptake in this field at this time (Hendrickson et al., 2002; Lapping, Schroeder, Marsh, Albalak, & Jabarkhil, 2002; Mackintosh, Marsh, & Schroeder, 2002). From the mid-2000s, there has been steady growth, with particular expansion in recent years: A possible explanation could be that in 2016, three international PD-focused conferences were held for the first time. Whatever the particular reason, it suggests growing interest and activity around positive deviance in developing countries, encouraging further work in this domain.

3.2 | PD research approaches

Four research approaches to positive deviance were identified from the 41 reviewed articles, normal PD (25 studies), comparative PD (seven studies), programmatic PD (six studies), and PD evaluation (three studies). Below is a summary of each detailed here so that the reader may understand better the type of development activity to which positive deviance has so far been applied and in what manner:

3.2.1 | Normal PD

This is the most common approach that applies the PD approach to a single group. Most studies of this type stop at the PD inquiry stage (ie, step 3 of the PD methodology outlined above), where the uncommon, successful practices of PDs are identified, without going further into designing interventions to promote those practices and monitoring progress. For instance, in Lackovich-Van Gorp (2017), a study was conducted to investigate strategies that could prevent marriage by abduction in Ethiopia. PDs were girls over 18 years old coming from very poor households who were still not married. The intervention applied only the first three steps of the PD methodology to identify PDs and the strategies they employed to protect themselves from marriage by abduction. The average duration of such studies is 8 months. A total of 68% of those studies use mixed methods, 21% use quantitative methods, and 11% used qualitative methods. The normal PD approach covered studies that tackled issues including health care–associated infections (de Macedo et al., 2012; Marra et al., 2013), enhancing health outcomes of women in disadvantaged circumstances (Long et al., 2013), cancer prevention (Vossenaar et al., 2009; Vossenaar, Bermúdez, Anderson, & Solomons, 2010), child marriage (Lackovich-Van Gorp, 2017), child rearing (Aruna, Vazir, & Vidyasagar, 2001), infectious disease control (Babalola, 2007; Babalola et al., 2002; Nieto-Sanchez, Baus, Guerrero, & Grijalva, 2015), improving pregnancy outcomes (Ahrari et al., 2002), counselling for family planning (Kim, Heerey, & Kols, 2008), child malnutrition (Aday, Hyden, Osking, & Tomedi, 2016; Bolles, Speraw, Berggren, & Lafontant, 2002; Guldan et al., 1993; Kanani & Popat, 2012; Merchant & Udipi, 1997; Merita, Sari, & Hesty, 2017; Roche et al., 2017; Sethi, Kashyap, Seth, & Agarwal, 2003; Shekar, Habicht, & Latham, 1991; Shekar, Habicht, & Latham, 1992; Wishik & Van Der Vynckt, 1976), neonatal mortality (Marsh et al., 2002), and managing medico-social problems through self-care (Gidado, Obasanya, Adesigbe, Huji, & Tahir, 2010).

3.2.2 | Comparative PD

Studies in this research approach compare the results of two methodologies each applied on a different group, having PD as one of the methodologies. It includes control trial study designs where a PD intervention is applied to one group, and the outcomes are compared with an equivalent control group that was not exposed to the PD intervention. As an example, in one of the reviewed studies (Lapping, Schroeder et al. 2002), a PD inquiry was compared with a case-control study to identify factors associated with nutritional status of Afghan refugees in Pakistan (concluding PD to be at least as good if not more effective than control study in factor identification). Research in such studies is mainly mixed methods and sometimes it is only quantitative. Study durations are 7 months on average. The comparative PD approach includes studies that tackled: health care–associated infections (Escobar et al., 2017; Marra et al., 2010), malnutrition (Hendrickson et al., 2002; Ndiaye, Siekmans, Haddad, & Receveur, 2009; Nishat & Batool, 2011), and clinical performance in medical schools (Zaidi et al., 2012).

3.2.3 | Programmatic PD

Studies belonging to this approach aim at understanding why a few individuals (PDs) respond to a development intervention programme better than their peers who are targeted by the same intervention. The PD inquiry is used to identify reasons behind the successful responses of the PDs, and the findings are used to inform intervention strategies and to increase overall adoption. For instance, in Garrett and Barrington (2013), a qualitative study was conducted to investigate barriers that prevent Honduran women from engaging in a cervical cancer screening programme. PDs were women that engaged in the uncommon but beneficial practice of screening. The PD intervention was designed to identify those women and the factors that led to their uncommon behaviour. Those factors (eg, self-love and self-support) were to be used in future screening promotion efforts. Research in such studies is either quantitative or mixed methods. Study durations are on average 3-month long. And since programmatic studies' main interest is just in post hoc identification of reasons for deviants to adopt or engage with the focal programme, they usually end at the third stage of the PD methodology. Example studies include programmes concerned with malnutrition (D'Alimonte, Deshmukh, Jayaraman, Chanani, & Humphries, 2016; Levinson, Barney, Bassett, & Schultink, 2007; Sethi, Sternin, Sharma, Bhanot, & Mebrahtu, 2017), farmer training (Tekle, 2015), and livestock feed technology adoption (Birhanu, Girma, & Puskur, 2017).

3.2.4 | PD evaluation

This is the least prevalent research approach that aims at evaluating the sustainability and impact of a PD intervention. Studies are also the longest with an average duration of 18 months and rely mainly on mixed methods in evaluation. The reviewed literature included three evaluative studies in malnutrition (Anino, Were, & Khamasi, 2015; Lapping, Schroeder et al. 2002; Mackintosh et al., 2002) and one study in infectious disease control (Marra et al., 2011).

3.3 | Sources of data

All the reviewed studies used primary data for PD identification and inquiry except for two studies that used secondary data. The first of this latter group was an exploratory study (Long et al., 2013) that investigated the factors associated with positive health outcomes among rural women in West Bengal. It used previous data from a randomized control trial conducted in a rural population, on 2227 consenting women and adolescent girls. Using quantitative analysis only, it was possible to examine the characteristics of PDs and factors affecting better health outcomes. However, there was limited ability to examine other possible factors affecting the targeted outcome, since the tool used to collect data for the previous study was not designed for the same purpose as this later study. The second study (Birhanu et al., 2017) was also an exploratory study that aimed at investigating the factors leading to better adoption of livestock feed technologies in Ethiopia. It used a previous household survey that included 603 farm households and aimed at identifying successful cases of improved livestock feed technologies and factors underpinning this success. Since the original study had the same purpose as the PD study, the collected data were able to unveil all possible factors affecting the desired outcome through quantitative analysis. These studies indicate the potential to undertake positive deviant identification without a need for primary research, thus signalling not only the potential for big data-based PD studies but also the challenge of repurposing datasets not specifically gathered for PD purposes.

3.4 | PD unit of analysis

The majority of the reviewed PD studies had individuals (infants, children, mothers, patients, students, health care workers, etc) as their primary unit of analysis, except for three studies that investigated positively-deviant farmer training centres (Tekle, 2015), farm households (Birhanu et al., 2017), and (disease-resistant) houses (Nieto-Sanchez et al., 2015). None of the studies conducted aggregation analysis, eg, identifying community-level deviance instead of individual-level deviance. This can be attributed to the small sample size in terms of number of communities covered that would not permit the identification of this type of deviance, a limitation that larger-scale datasets might not suffer.

3.5 | PD challenges

Analysis of the literature on positive deviance reveals a series of challenges or limitations arising from work to date, challenges which we will later interrogate to see if big data might have some response:

3.5.1 | Time and cost

The application of the PD approach is time-consuming (Felt, 2011; Lapping et al., 2002; Marsh et al., 2004). As can be seen from the data in Section 3.2, it takes months to complete the phases sequentially. Alongside concerns about the time requirements are also concerns that the quality of implementation may be compromised due to time constraints. For instance, one of the studies reported that the desired large sample size was not obtained because of time limitations (Nishat & Batool, 2011). Since PD depends typically on primary data collection, community participation, face-to-face interviews and observation, the cost of PD interventions also tend to be high. As with the time constraint, cost is also a function of sample size that can encourage smaller samples. In addition, collecting primary data from some high-risk areas brings with it additional time, cost, and complexity in order to mitigate the risks (Shekar et al., 1991).

3.5.2 | Positive-deviant identification

Within any given population, positive deviants are relatively rare. Based on those reviewed studies that provide the necessary data, we can calculate an average prevalence rate of 11%. This is slightly higher than, but not completely out of line with, earlier estimates that PDs typically form 0% to 10% of a population (Marsh et al., 2004). Whatever the exact figure, PDs are statistical outliers, and sample size thus plays a role. As the sample size increases, the more representative of the population it becomes, and thus, the likelihood/prevalence of positive deviants becomes greater (Osborne & Overbay, 2004).

Hence, there is a statistical pressure to undertake large sample size studies in order to identify a sufficient sample size of PDs. However, given the time- and labour-intensity of PD just noted, with costs proportional to sample size, there is a counter-pressure to keep overall sample sizes small. For example, in the comparative study of Afghan refugees in Pakistan (Lapping, Schroeder et al. 2002), the compared groups were 8 and 50 strong. Another study in Egypt that addressed factors associated with successful pregnancy outcomes reported that the information gained from PDs was limited; this can be attributed to their very small sample size (n = 11) (Ahrari et al., 2002). Similarly, in the Honduras study examining women who overcame barriers to cervical screening, the sample size (n = 8) was seen as not large enough to achieve full saturation of relevant factors. The use of very small samples for PDI not only potentially misses important aspects of PD behaviour but would also have less statistical power to identify valid associations. Additionally, as previously mentioned in Section 3.4, small sample size limits the ability to identify deviance at different levels of aggregation. One potential solution would be the use of large secondary datasets, which could be analysed at low cost while not compromising the number of PDs identified.

Moreover, PD primary data collection—often due to its time and cost—is undertaken via a cross-sectional not longitudinal design. It provides a single snapshot of the population since it depicts the behaviours of the analysed units at a certain point of time; hence, deviance becomes static and could be referred to as *point anomaly* in statistics (Goldstein & Uchida, 2016). What it cannot do is identify the dynamics of deviance such as *contextual/conditional anomalies* that arise due to the particular condition of a context, with those conditions potentially differing over time. For example, one of the reviewed studies sought to identify preventive measures to control Chagas disease; PDs were bug-free houses throughout the period of inspection. However, an identified limitation of the study was that the houses selected are not necessarily bug-free throughout the year, since the entomological searches happened during the summer, and natural factors could have affected the results (Nieto-Sanchez et al., 2015). Hence, a few of those PDs might have been false positives: appearing as a point anomaly attributed to the individual house but in fact a contextual/conditional anomaly. Again, one potential cost-efficient solution would be large secondary datasets, in this case, where the data were collected longitudinally.

3.5.3 | Methodological risk

Alongside the practical risks of PD given the need for large-scale primary fieldwork in developing countries, we were able to identify two methodological risks associated with use of the PD approach. First, there is a PD behaviour identification risk. For example, some of the studies (D'Alimonte et al., 2016; de Macedo et al. 2012; Marra et al., 2010) that used observational methods in PD inquiry reported the potential for a *Hawthorne effect*: an alteration of the behaviour of the subjects of a study due to being observed. Another risk is the inability to extract successful strategies and behaviours practised by the positively deviant individuals. Positive deviance methods presume the willingness of PDs to share their strategies and best practices. However, this might compromise what the deviants see as a competitive advantage over others resulting from their outlier behaviour, leading them to be unwilling to share (Felt, 2011). For example, in one of the reviewed studies (Zaidi et al., 2012), positive deviance was used to try to identify and disseminate the strategies employed by successful medical students, in order to improve the clinical performance of their peers. There is a potential risk that the high performers would refrain from sharing their best practices when interviewed. In both cases, analysis of behaviour via secondary/remote observation could help to avoid these risks.[5]

Second, there is a risk of not being able to establish a cause and effect relationship between PD interventions and achieved results (step 4 of the PD methodology). Some studies (eg, Nieto-Sanchez et al., 2015; Nishat & Batool, 2011) noted that results could not be attributed to the PD intervention alone, since the targeted population might have been exposed to other interventions and external factors that might have contributed, partially, to the desired outcome. This challenge of attribution (and also issues of time and cost) may be one explanation behind the limited number of evaluative studies of PD interventions: ie, those that moved to step 5 of the PD methodology (Ndiaye et al., 2009; Roche et al., 2017). Another explanation may be the lack of guidance on how to apply credible monitoring and evaluation techniques (Lapping, Schroeder et al. 2002; Felt, 2011). As noted in Section 3.2, only 7% of the reviewed studies were evaluative: very low considering the importance of understanding the development impact and value of PD approaches. Being able to demonstrate the lasting success of a PD intervention could support its wider adoption and the wider adoption of PD more generally.

3.5.4 | Scalability

There are two challenges underlying the scale-up of PD interventions. The first challenge is in scaling practices within a community. PD relies heavily on community engagement to promote the adoption and mobilization of the identified practices and to achieve behavioural change through self-efficacy. For instance, 25% of the reviewed studies employed the *PD Hearth* (Wollinka, Keeley, Burkhalter, & Bashir, 1997), a nutrition education framework designed to empower mothers to enhance the conditions of their malnourished children. It requires mothers of PD children to host neighbouring mothers and their malnourished children for 12 consecutive days, where they prepare meals and feed their children together (Felt, 2011; Lapping, Marsh et al. 2002; Marsh et al., 2004; Pascale et al., 2010). With this level of engagement, PD proved successful in small-scale adoption but made large-scale adoption a challenging task given both the growing complexity and the challenge of a strong enough pre-existing social fabric to ensure cooperation of the PD mothers. Zeitlin (1991) notes a similar point that—for certain communities and domains of action—there could be resistance to making everyone a “top performer”, and that moves towards that might disrupt and disintegrate system dynamics.

The second challenge is the scaling of practices across communities. An issue with PD is the inability to generalize practices and behaviours inferred from one community to another. In the majority of the reviewed studies, PD interventions targeted small-scale communities, and the inferred practices were particular to the circumstances of this community making it difficult to replicate in other communities (Saïd Business School, 2010). If broader, cross-community data could be accessed—identifying PDs and their behaviour on a wide scale—then this challenge of limitations on generalization could be reduced to some extent; although of course this would likely assume/require the presence of noncommunity-specific behaviours underlying positive deviance.

3.5.5 | Narrow domain/geographic scope

There is a current skew in the domain and geographic focus of PD applications in developing countries, summarized in Figure 3. Regarding domain coverage, we found that the vast majority—89%—of the reviewed studies were in public health, with 41% focused specifically on malnutrition. Put another way, there were only four non-health studies: two on agriculture (Birhanu et al., 2017; Tekle, 2015), one on child protection (Lackovich-Van Gorp, 2017), and one on education (Zaidi et al., 2012). As for the geographic coverage, there are nearly 150 developing countries (OECD, 2017), but PD studies identified by the review encompassed only 20, with just four countries (India, Brazil, Pakistan, and Ethiopia) responsible for almost 50% of studies, and only two countries having hosted studies from more than two domains. There is also a within-country geographic concentration, with 83% of studies being undertaken with rural communities, significantly out of kilter with population distribution in developing countries.

This domain and geographic concentration can be attributed to a form of path dependency in positive deviance. The first use of PD (Wishik & Van Der Vynckt, 1976) was for a nutrition-based intervention in a rural community. Then, early adopters of PD who set the foundation for the field (Sternin et al., 1997; Zeitlin, 1991; Zeitlin, Ghassemi, & Mansour, 1994) including development of an operationalizable framework for PD (Sternin et al., 1998) all undertook work on malnutrition in rural areas. Despite applicability of PD across many domains, subsequent PD actions have often followed suit, leaving a gap of domains and locations that have been largely ignored by PD to date.

That PD has relevance to other countries, and other domains can readily be seen from its application in the global North, eg, to public sector reforms (Andrews, 2015), enhancement of prison conditions (Awofeso, Irwin, & Forrest, 2008), organizational scholarship (Mertens, Recker, Kohlborn, & Kummer, 2016), and waste management (Delias, 2017). But PD for developing countries needs encouragement to spread further than its current narrow path.

[5]Although noting as per the point raised in Section 3.3 that relying on quantitative analysis of secondary data to infer practices might limit the ability to test the effect of other potential factors influencing the desired PD outcome, especially in studies where the instruments used to collect the data were not designed to measure the desired outcome (Long et al., 2013).

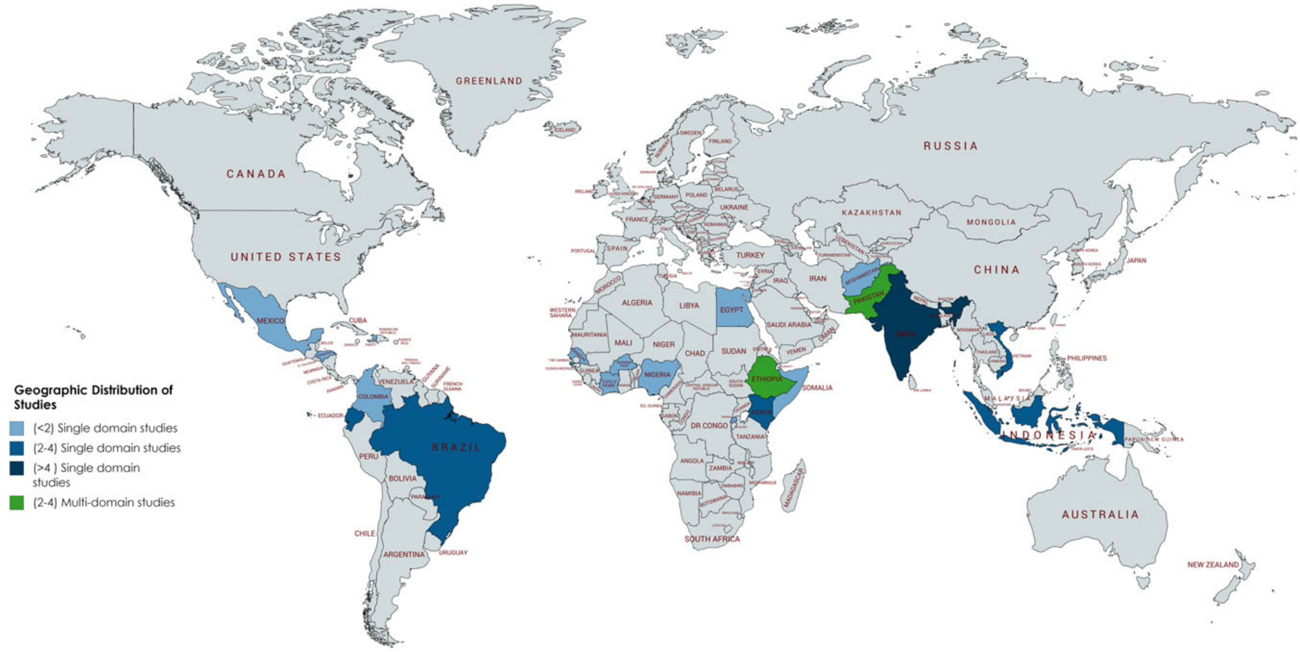

FIGURE 3 Geographic distribution of the domains of positive deviance (PD) literature relating to developing countries

4 | BIG DATA FOR DEVELOPMENT

The evolution and diffusion of digital infrastructures has led to a proliferation of data, often referred to as "big data" (BD). There has also been an increasing ability to make use of it, characterized in enhanced processing and storage capacities, which has provided an opportunity to convert these data into information and knowledge that feeds decisions and actions. The main characteristics of BD derived from Gartner's definition (Gartner, 2013) are (1) volume: huge amounts of data generated from the rapid diffusion of mobile phones, social media, and other online services, plus the growing use of sensors and satellite imagery; (2) velocity: the growing speed and currency of data production and enabling decisions and actions to be taken in a timely manner; and (3) variety: the combined availability and potential use of structured data, having a predefined structure (eg, mobile transactions), and unstructured data, not having a predefined structure (eg, email, video, and audio), to extract insights.

The applications of "big data for development" (BD4D) span a wide variety of domains and leverage new sources of data and new analytical tools. It is argued that big data can fundamentally shift the way we pursue social change as it is capable of providing snapshots of the wellbeing of populations at high frequency, high degree of granularity, and from a wide range of angles, narrowing both time and knowledge gaps (UNGP, 2012). Sitting alongside concerns about and critiques of BD4D (eg, Taylor & Broeders, 2015), big data are therefore also argued to offer new opportunities for development for reasons that include the following:

1. Low cost: Digital traces produced from digital platforms provide a low-cost alternative to traditional sources of data (eg, censuses and surveys); in some instances by replacing variables of interests with correlated proxies (Hilbert, 2016). For instance, mobile call duration and frequency have been correlated to income or education levels in a geographic region (Frias-Martinez & Virseda, 2013) and could then substitute for survey and similar data gathering.
2. Real-time feedback and awareness: Through monitoring populations, BD makes it possible to understand where policy and programme interventions are succeeding or failing in real time, in order to make adjustments in a timely manner.
3. Broad sampling: With a global average penetration of 95% and a 75% penetration among base of the pyramid populations (Cartesian, 2014), mobile phones are coming close to sampling the universe N instead of sampling n of the universe N (Hilbert, 2016).
4. Detail and insight: The ability to merge and use different sources of data reflecting a certain event or reflecting the behaviours of an individual, community, or an organization provides a real-time, cross-validated, fine-grained picture of reality.
5. Big data analytics: Advanced analytics techniques—those which perform particularly well when applied to huge datasets—enable big data to be used for better decision-making. An example is machine learning, a subfield of artificial intelligence, which gives computers the ability to learn from data without being explicitly preprogrammed on the knowledge they will extract.

Given this potential value of big data to development, the review of literature outlined in Section 2 was undertaken and is reported next. This section begins with a timeline showing the volume of BD literature over the last decade followed by a thematic classification of the literature and the geographic distribution of its application domains. We then discuss the other forms of data that were combined with BD, the different units of

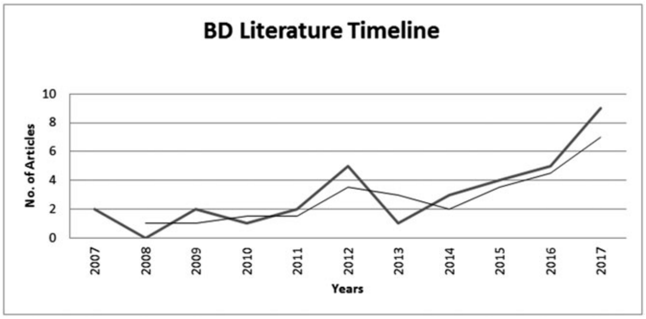

FIGURE 4 Timeline of reviewed big data for development (BD4D) literature for period 2007 to 2017. The thick line represents the actual number of studies whereas the thin line represents a projection of the trend. The bold line represents the actual number of studies whereas the thin line represents a projection of the trend

analysis, and the employed BD analytics techniques before outlining the challenges of BD4D at the end. All this provides the knowledge of big data necessary to understand how it might be applied to action research on positive deviance, as then discussed in Section 5.

4.1 | BD4D literature timeline

Figure 4 shows that the BD4D research area is relatively new, given that all the identified literature was published since 2007. There is some growth in the number of studies during the later years and, similar to PD studies, there is an accelerated growth towards the very end of the period under review. As with the PD literature, this suggests a growth in activity and interest, encouraging further work on big data for development.

4.2 | BD4D research approaches

Utilising the taxonomy of BD applications proposed by Hilbert (2016), we classified the reviewed literature into four main approaches based on the elements being tracked: locations (12 studies), words (four studies), nature (six studies), and economic activity (12 studies). This gives a sense of the general scope of big data application, for example, its potential application in positive deviance analysis.

4.2.1 | Tracking locations

This approach contains applications that analyse human and object mobility data. This typically comes from mobile phones in the form of de-identified call detail records (CDRs), which usually provide the time and associated cell tower of text messages and calls. This is the most common data type in the reviewed BD4D studies, notwithstanding the concerns of mobile phone operators about releasing such data: either that it has commercial sensitivity and could give competitors an unfair advantage or that techniques could be used to de-anonymize the data and uncover individual subscribers' identities. There are also inherent biases in CDRs due to socio-economic and geographic variations in phone ownership, but evidence still suggests that CDRs provide the best description to date of population movement in low- and middle-income countries (Wilson et al., 2016). CDRs have thus been used to analyse travel and migration patterns of mobile users to understand the spread of infectious diseases in low-income settings (Bengtsson, Lu, Thorson, Garfield, & von Schreeb, 2011; Buckee, Wesolowski, Eagle, Hansen, & Snow, 2013; Tatem et al., 2009; Wesolowski et al., 2014; Wesolowski et al., 2015) and to identify population displacement following disasters (Bengtsson et al., 2011; Lu, Bengtsson, & Holme, 2012; Wilson et al., 2016) and migration patterns in climate-stressed regions (Lu et al., 2016).

But CDRs are not the only location-related data that has been used. For example, data from a web mapping service application (Baidu.com) were used to calculate average travel distance to healthy food outlets in order to identify urban areas with limited food accessibility in China (Su, Li, Xu, Cai, & Weng, 2017). Car GPS data have been used to map the spatio-temporal distribution of pollution emissions from traffic (Huang, Cao, Jin, Yu, & Huang, 2017; Luo et al., 2017). There are also studies that combine mobility data with other sources of data for better representation, cross-validity, data enrichment, and for covariance analysis. For instance, in Tatem et al. (2014), physical data in the form of satellite images, climate, and topographic data were combined with CDR data to understand the spread of malaria, specifically, the seasonality of movements including movement across borders.

4.2.2 | Tracking words

This approach contains applications that analyse actions, activities, and events based on words, which typically come from social media. It usually faces the challenge of representational validity in terms of the demographic, socio-economic, and geographic profile of contributors given skews in terms of those who do and do not use social media. There is also the challenge of potential differences between digital and real behaviour such as self-censorship or presenting a false image. One advantage, although, is that in many cases, data sources are readily accessible because they are "open data"[6] in nature; and they also benefit from enabling mapping of behaviour in real time (Pfeffer, Verrest, & Poorthuis, 2015). Examples of applications include the use of Google search word trends to compare the demand for massive open online courses (MOOCs) between different countries (Tong & Li, 2017), applying co-word analysis to map the research themes of Indonesian scholars' publications (Surjandari, Dhini, Lumbantobing, Widari, & Prawiradinata, 2015), analysing protest activity using twitter data during the Egyptian January 25, 2011 revolution (Wilson, 2011), and revealing geographical and social patterns of tweets pertaining to flooding and criminal activity in Caribbean cities (Pfeffer et al., 2015).

4.2.3 | Tracking nature

This approach contains applications that use data to observe environmental and natural phenomena to mitigate risk, improve emergency response, or to optimize performance (Hilbert, 2016). Satellite imagery is the most common BD type used by those applications; this is due to its increasing availability at global and lower scales, often via open or other no-cost access. Growth in datasets over time is also allowing use for time series analysis. The reviewed studies falling under this approach used satellite imagery to detect illegal deforestation activities (Burgess, Hansen, Olken, Potapov, & Sieber, 2012), to monitor coal fires (Jiang, Jia, Chen, Deng, & Rao, 2017), to map temporal water surfaces (Haas, Bartholomé, & Combal, 2009), and to model crop growth (Tesfaye et al., 2016).

Other studies collect environmental data via sensors. For example, sensor networks were used to monitor the spatio-temporal distribution of greenhouse gas emissions in China (Tang, Yang, & Zhang, 2014). There is also a study (Zhang, Chen, Chen, & Chen, 2016) that combined sensor data, satellite images, and meteorological data with social media for the analysis of urban waterlogging disasters (where drainage systems are unable to cope). Physical data were used to observe and understand waterlogging, and social media data (ie, tracking words) were used to identify qualitative features of waterlogging incidents.

4.2.4 | Tracking economic activity

This approach contains applications that use data that reflect the economic situation of the analysed units. Satellite images and CDRs are the two most popular BD sources that are used for this purpose. For instance, CDRs were used to predict wealth of individuals by tracking their history of mobile use represented in the intensity, volume, time or direction of calls (Blumenstock, Cadamuro, & On, 2015), or through phone ownership (Blumenstock & Eagle, 2012). Similarly, satellite images were used to predict poverty and estimate economic growth in a number of studies. For instance, daytime satellite images were used to predict socio-economic well-being by analysing visible household assets (Jean et al., 2016). They were also used to map population settlements (Tatem, Noor, von Hagen, Di Gregorio, & Hay, 2007) and classify slums (Kohli, Sliuzas, Kerle, & Stein, 2012).

Other studies used satellite nightlight images as a proxy for electricity consumption levels which, in turn, can be seen as a proxy for levels of economic activity (Doll & Pachauri, 2010; Henderson, Storeygard, & Weil, 2012; Sutton, Elvidge, & Ghosh, 2007). However, nightlight images have difficulty distinguishing between poor densely populated areas and wealthy sparsely populated areas (Jean et al., 2016). Hence, Njuguna and McSharry (2017) complemented satellite nightlight data with CDR data to build a stronger poverty proxy by incorporating the level of mobile usage. Other examples include the use of data from a leading online retail platform in China (sofang.com) to analyse spatio-temporal features of housing prices (Li, Ye, Lee, Gong, & Qin, 2017), and the use of data from China's industrial enterprise database and customs import/export trade database to examine the extent of "greening" of global value chain enterprises (Song & Wang, 2017).

4.3 | BD4D studies domain/geographic distribution

Table 2 summarizes the overall domains and specific application topics of the reviewed BD4D studies. We can see that economics, public health, and environmental studies were the most common application domains, within which the most common applications were infectious disease (16%) and poverty measurement (16%). In infectious disease studies (Bengtsson et al., 2011; Tatem et al., 2009; Wesolowski et al., 2014; Wesolowski et al., 2015), mobile data and health data are combined to identify risk areas. As for poverty measurement (Blumenstock et al., 2015; Jean et al., 2016; Njuguna & McSharry, 2017), satellite nightlight and mobile data are used as a proxy for economic activity at temporal and geographic scales for which traditional data are of poor quality or unavailable.

Figure 5 shows the geographic distribution of BD studies by domain, indicating that China had the biggest share of BD4D studies that span multiple domains. In addition, there were three studies not presented on this map since they were applied across multiple countries. The first study (Doll & Pachauri, 2010) used night-time satellite imagery to estimate rural populations without access to electricity in developing countries. The second study (Henderson et al., 2012) used satellite data to augment official income growth measures of coastal areas in sub-Saharan Africa.

[6]Data "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control" (WP, 2018).

TABLE 2 Classification of BD4D studies by domain and application[a]

Domain	Application	No. of studies
Economics	Predicting poverty	6
	Measuring economic growth	3
	Mapping population distribution	1
	Analysing housing prices over time	1
Public Health	Infectious disease control	6
Environment	Analysing traffic pollution	2
	Monitoring deforestation activities	1
	Monitoring changes in small water surfaces	1
	Monitoring greenhouse gas emissions	1
	Coal fire suppression efforts	1
Disaster Response	Population displacement post disasters	4
Food & Agriculture	Accessibility of healthy food stores	1
	Drought tolerant crops	1
Energy	Predicting electricity demand load	1
	Green technology adoption	1
Education	Quantifying MOOC demand	1
	Research theme mapping	1
Urban Governance	Urban waterlogging	2
	Security	1
Politics	Analysing twitter activity in revolutions	1

[a]The total number of studies listed in the table is 37 although the reviewed BD studies were only 34. This is because two studies presented applications in multiple domains.

FIGURE 5 Geographic distribution of the domains of big data for development (BD4D) literature

The third study (Haas et al., 2009) used remote sensing data to map temporary water bodies in regions of western sub-Saharan Africa. For within-country studies, some are rural-focused, many cover both urban and rural areas, and some are solely urban-focused.

While the geographic coverage of big data studies in terms of countries is, as yet, not much better than that of PD studies, there is clearly ready potential for a much broader scope given the universal presence of at least some types of big data, including scope for more urban-focused work. And, in relation to domains, big data have already shown application to a wider range of topics than positive deviance.

4.4 | Sources of data

A total of 37% of the reviewed studies complemented big data with other sources of primary and secondary data. For instance, in Sutton et al. (2007), a model was developed using GDP and population data and night-time satellite imagery to predict GDP at subnational levels. Similarly, Jean et al. (2016) used daytime satellite images, annotated with geo-referenced household consumption survey data, to develop a transfer learning model which trained the data in survey-rich countries to predict consumption and assets in survey-poor countries, using daytime satellite images alone. Wesolowski et al. (2014) demonstrated that community surveys can complement mobile data to approximate travel patterns of non-subscribers in rural areas. Secondary data are also used for cross validation, for example, in Wesolowski et al. (2015), population-based surveys were used to validate the results of mobile data analysis that measured the magnitude of population displacements.

A number of studies combined health data, having geo-referenced disease cases, with mobility data to develop risk maps for infectious diseases (Bengtsson et al., 2011; Tatem et al., 2009; Wesolowski et al., 2014; Wesolowski et al., 2015). Spatially referenced primary survey data have been used to support satellite imagery and mobile phone data in understanding the reality and the impacts of socio-economic or socio-spatial differences. For example, Blumenstock et al. (2015) was able to derive insights on the degree of wealth of individuals by supplementing mobile phone history big data with data from an anonymized mobile phone survey. Hence, it was possible, through training models, to predict wealth using only mobile phone use history for individuals not included in the survey sample. We may conclude from the above examples that —notwithstanding the potential for big data to provide a faster, cheaper, and a more granular alternative to traditional data sources—greater value may be captured when BD is combined with those data sources instead of simply replacing them. In particular, big data can be validated in locations where comparator data exists and then applied alone in locations where comparators are absent, an especially helpful approach in developing countries where comparator data may be thin on the ground.

4.5 | BD4D unit of analysis

In the reviewed BD4D studies, the unit of analysis ranged from individuals (eg, mobile users and twitter users), through geographic areas (eg, grid areas located via satellite imaging), to regions and even countries. The majority of studies applied aggregation on different scales to analyse and visualize patterns, events, and spatial relations. Conversely, BD4D studies also provided a disaggregation opportunity. For example, in poor countries, data about economic growth are often available only at high levels of geographic aggregation (eg, national level) because it is collected using sample surveys instead of location-disaggregatable census surveys (Chandy, Hassan, & Mukherji, 2017). Due to the finer granularity of big data represented in user CDRs or night-time satellite images, it was possible in a number of studies (Blumenstock et al., 2015; Blumenstock & Eagle, 2012; Doll & Pachauri, 2010; Henderson et al., 2012; Sutton et al., 2007) to use those forms of data as proxies for economic growth at subnational level. Compared with typical PD data, big data therefore may provide much greater aggregation and disaggregation potential.

4.6 | BD4D analytics

Discussion of big data in development can tend to focus on inherent qualities of the data such as the 3Vs: volume, velocity, and variety. But greater value—including value for positive deviance—may vest in the advanced analytics techniques that are being applied to big data, and to the use of these techniques for improved development decision making. The reviewed BD4D literature utilized two types of analytics:

1. **Descriptive analytics** provides information about the past and present. It uses data aggregation and data mining techniques to summarize historical data and answer the questions, “What has happened?” or “What is happening?”. BD visualization is the essence of this type of analytics, creating a new face to standard descriptive statistical methods. For example, Pfeffer et al. (2015) used descriptive analytics to geo-map the word frequency of twitter data relating to two Caribbean cities which referred to crimes and flooding in order to better understand those phenomena.
2. **Predictive analytics** uses statistical models and forecasting techniques to answer the question, “What will happen?”. It encompasses two types: inference and forecasting. Inference models predict the value of a certain variable of interest based on its association with variables in another data source. For example, Jean et al. (2016) used daytime satellite images and households surveys, covering a specific area, to train a predictive model that was able to estimate the economic well-being, in another area, using only satellite imagery. This approach was able to overcome the data sparsity issue, through training models in data-rich areas to make predictions in data-poor areas. Inference models were also used to predict water precipitation and potential for waterlogging based on climate data and road and terrain maps (Zhang et al., 2016). On the other hand, forecasting models utilize trend analysis and pattern recognition techniques to predict what will happen in the future, based on what happened in the past. For example, in Ifaei, Karbassi, Lee, and Yoo (2017), multivariate dynamic models were used to forecast power consumption using previous data on exported and imported power and quantity of stored power over a period of time.

Big data also enable the use of intelligent data analytics techniques that perform better when applied to huge amounts of data. As noted above, an example is machine learning (ML). From the reviewed BD4D literature, ML techniques can be grouped into two main categories: supervised and unsupervised learning. In the former, ML is applied on a training dataset where each input X is labelled to a class or output Y and the primary objective of the learning algorithm is to develop a mapping function $Y = f(X)$, so that when you have any input x, you can predict its output *y*. For example, supervised learning was used to predict poverty levels from mobile phone data (Blumenstock et al., 2015) and from satellite imagery (Jean et al., 2016). In unsupervised learning, ML is applied on a dataset where you only have input data *X* and no corresponding output *Y*. The primary objective of the learning algorithm is to discover underlying similarities between the input data points and create clusters of data based on the perceived similarity. It is also capable of allocating new inputs into the appropriate cluster. For example, in one of the studies, unsupervised learning was used to find the interrelationship among academics' research approaches and cluster them, based on the co-occurrence of the publications' keywords (Surjandari et al., 2015).

4.7 | BD4D challenges

While there are broader challenges relating to the use of big data in development—such as associated shifts in power between different groups (Sengupta, Heeks, Chattapadhyay, & Foster, 2017; Taylor & Broeders, 2015)—there are also a set of more practical challenges that emerged from the literature reviewed, which would need to be taken into account if using big data to research positive deviance.

Absence of Theory: The majority of BD applications do not use theory-driven models, especially in cases of predictive analytics where they depend mainly on past data to predict what will happen in the future. However, attribution analysis (cause and effect studies) investigating why outcomes change in response to variations in inputs will need a theoretical framework. Employing a theory of change can guide the identification of explanatory variables (inputs) and indicators for outputs, outcomes, and impact (Bamberger, 2016).

Proof of Concept Skew: As might be expected given the relatively formative nature of big-data-for-development, most of the BD4D literature represents a proof of concept rather than use of data for actual development-related decision-making and implementation. As just one example, Pfeffer et al. (2015) demonstrate what (relatively little) tweets might tell us about location of urban flooding and crime but without any engagement with real-life urban planning decisions.

Representational Validity: Mobile phone ownership is skewed towards certain population groups based on income, gender, or age, leaving specific groups and geographic areas under-represented in mobile-based sources of big data (Bengtsson et al., 2011; Tatem et al., 2014; Wesolowski et al., 2014; Wilson et al., 2016). This is even more of a challenge for social media data, in terms of demographic and socio-economic profiling and the geographic spread of the content generators (Pfeffer et al., 2015). BD sources were not particularly produced to investigate, assess, or measure any of the presented development-related applications; they are rather a side effect (Pfeffer et al., 2015). They provide one or more aspects of the studied issue but they might overlook other important aspects requiring on the ground, targeted inquiry (Lu et al., 2016; Pfeffer et al., 2015). This explains the recent debates (Graham & Shelton, 2013; Pfeffer et al., 2015) around the combined use of BD and other sources of data for better representational validity.

Human Capacity: Both researchers and practitioners typically lack the necessary technical skills needed to clean up data sources, to link different data sources, to analyse big data, to identify emerging patterns from big data, and so on (Pfeffer et al., 2015). There are also highly unstructured data types, like satellite imagery, the analysis of which requires knowledge of advanced analytics and machine learning tools (Jean et al., 2016). Those missing skills and local conditions, especially in developing countries, limit the exploitation of this valuable data source, creating new digital divides (Batty et al., 2012).

Data Accessibility: Most big data are not open and easily accessible. Data gatekeepers, such as mobile operators and public institutions, are not always willing to share their data, often because they consider it to be a source of commercial or political advantage (Pfeffer et al., 2015; Jean et al., 2016a).

Privacy and Legal Issues: It can be difficult to link and analyse different data sources while respecting privacy, eg, of individuals who produced the data (Blumenstock et al., 2015; Sutton et al., 2007). This can be particularly challenging in developing countries, where there is an absence of legal frameworks protecting citizens (Pfeffer et al., 2015). As noted above, one reaction of data providers, like mobile operators, is to restrict access to datasets. Another reaction is to anonymize CDRs by removing and aggregating some attributes. While understandable, this can reduce the developmental value that can be captured from the provided datasets.

In summary, and despite the demonstrated value of using big data in a variety of development-related applications, it is important to note the challenges associated with its use. Of particular relevance for this paper are challenges that could affect the significance of its use in positive deviance, like privacy and accessibility. For instance, BD depicting human behaviour is the most relevant data for PD; however, if these data might compromise the security or privacy or undermine competitive advantage of data owners, its accessibility and usage would require strict rules and principles backed by adequate tools and systems to ensure “privacy-preserving analysis” (UNGP, 2012).

5 | DISCUSSION

Positive deviance has been shown to be effective as a problem-solving approach in certain development domains but it faces challenges that have so far limited its uptake. Big data could potentially address some of those challenges and/or in other ways enhance current approaches to positive deviance, providing there was an adapted PD framework that could guide its use. Conversely, positive deviance appears to provide an approach to the detection of socio-economic anomalies that might broaden the application of big data in development. Alongside the growing interest in and practice of both positive deviance and big data in development, this creates an opportunity for "big data-based positive deviance" (BDPD). In this section, we will examine how BD could address some of the aforementioned PD challenges and we then propose a means to operationalize their combined use.

5.1 | BD as a response to PD challenges

While big data cannot address all of the positive deviance-related challenges identified in Section 3.5, it has potential in relation to most of them:

5.1.1 | Time and cost

PD studies mainly use primary data collection both for identification of positive deviants and for PDI: the inquiry into what causes the deviant outcomes. As noted above, primary data collection involves significant time, cost and risk, and use of other forms of data collection could therefore offer important advantages. In light of this, a few studies have made use of traditional secondary data—such as that from surveys—but this brought its own challenges; for example, it is hard to identify positive deviants from such datasets as they are often anonymized or out-of-date by the time they become publicly accessible; or it may be difficult to explain causes of positive deviance as important factors are missing from the survey. In addition, while the cost of reuse of survey data may be low, the actual financial costs of the original data-gathering are very high—particularly for census data.

In comparison, big data brings with it not just the gains of reduced time and cost common to reuse of secondary datasets but the more foundational reduction that the costs of initially gathering big data tend to be very low since it often makes use of already existing "data exhaust" from digital processes. In part thanks to low cost, there are also – thinking of satellite imaging and social media data – increasing sources of real-time big data. These avoid the problem of time lag (something particularly challenging with, say, census data which is often only gathered every ten years). Finally, the lack of cost constraints means that big datasets often have a much greater geographical scale than other forms of secondary data.

While big data does still suffer the secondary data shortcoming that it has been created for purposes other than PD analysis, it is increasingly present in locations—such as poorer countries or communities—where survey data either tend to be lacking, or based on very small samples, or inaccurate; these problems themselves sometimes deriving from the high cost and time requirements of surveys and the lack of resources for these locations (Letouzé & Jütting, 2015; Mügge, 2014). Initiatives have already demonstrated the ability of big data—such as satellite imaging or CDRs—to fill these data gaps and act as proxies for socio-economic indicators (Henderson et al., 2012; Jean et al. 2016; Njuguna & McSharry, 2017).

Big data therefore show significant potential to help address the time, cost, and risk constraints faced by current positive deviance studies, including the constraints associated with using traditional secondary data sources for PD analysis.

5.1.2 | Positive deviant identification

As mentioned earlier, the use of primary data collection to identify positive deviants has three main drawbacks: low sample power, inability to identify dynamic anomalies, and limited aggregation analysis. Use of big data sources offers a potential to overcome those challenges as follows:

1a) Sample Power: Big data are produced passively at marginal or no additional cost whereas traditional data sources are produced actively with a cost that is proportional to the size of the sample. As a result, BD sources tend to have a much larger coverage of populations. Given that positive deviants are relatively rare, the larger samples from big datasets will enable the identification of a larger number of PDs. Accordingly, the risk of overlooking important factors will be reduced and the ability to generalize practices to larger populations will be improved.

1b) New Identification Techniques: Although not recognized in the PD literature as a challenge, there is an opportunity offered via BDPD that is not available to traditional positive deviance identification. This is the application of machine learning which, as noted above, works most effectively with large datasets (Hilbert, 2016). Machine learning–based approaches for anomaly detection outperform simple statistical models for various types of anomaly (Chandola, Banerjee, & Kumar, 2009; Goldstein & Uchida, 2016), thus providing the potential for better PD identification than currently possible. Supervised machine learning can also be used for predictive analysis, by first being trained to analyse a small sample set of big data in tandem with ground truth data: for example, satellite images combined with survey data. The survey data already identifies the positive deviants, and machine learning then develops the ability to identify PDs within the corresponding satellite image data. The analytical algorithms can then be applied solely to big data sources and will identify positive deviants from those sources on a much wider scale.

2) Dynamic Anomalies: Where traditional data sources typically provide a static, cross-sectional view of behaviour, big data can often provide a dynamic picture of the targeted population over time. Hence, as discussed in Section 3.5, BDPD can be better at identifying and potentially eliminating contextual, conditional anomalies as explanations for positive deviance.

3) New Levels of Aggregation: The majority of PD studies have individuals as the primary unit of analysis, whereas in BD studies, the unit of analysis can range from individuals to communities and regions. Hence, the use of big data could provide PD with the ability to identify deviance at different—ie, higher—levels of aggregation than just individuals.

5.1.3 | Methodological risk

Big data are in almost all cases gathered within the explicit intervention of, or tangible visibility to, the subject populations. As such, risks arising from populations knowing they are being observed and questioned, such as the Hawthorne effect or refusal to share practices, are avoided. In addition, where BD sources incorporate outcome indicators of the positive deviant behaviours, then those sources can be used for ongoing monitoring of the effects of a PD intervention. This might help address the challenge that the lack of credible monitoring and evaluation techniques limits uptake of the PD approach, especially in new contexts and settings.

5.1.4 | Scalability

As discussed earlier, PD faces two scalability challenges: scaling practices within a community and scaling practices across communities. Both of these issues are partly rooted in socio-behavioural factors that big data are unlikely to be able to address. But one can hypothesize some potential added value.

For example, for the first challenge, unsupervised machine learning could be applied to a big dataset to cluster the PD intervention population (based on machine-inferred similarities). Then intervention could target only those clusters with socio-economic determinants similar to those of the deviants for practice dissemination and adoption. This could reduce both the time and cost required for scale-up in comparison to the traditional methods of practice dissemination that rely heavily on community mobilization. Although noting two potential limitations: first that, by definition, positive deviants are sought within populations with similar socio-economic determinants; and second that mobilization and incentives are always likely to be important in any type of PD implementation. There is also a small possibility that BD-based, statistically verified evidence might prove more convincing to potential adopters of PD behaviours than the more qualitative findings typical of traditional PD.

The second challenge could be mitigated if cross-community big data sources are available. These would enable the identification of PDs and their behaviour on a broader scale making generalizations possible.

5.1.5 | Narrow domain/geographic scope

Despite the effectiveness of PD as a problem-solving approach for international development, its uptake by developing countries in domains other than public health has been very limited, and its application has been concentrated in rural areas of just a few countries. While geographic coverage of BD4D in the literature to date has also been concentrated, that literature already illustrates more application in urban areas and application in several other domains. Big data can thus expand the scope of PD, enabling it to break from its current path dependency. Outline domain examples where big data-based positive deviance could operate include the following:

- Infectious Disease Control: A number of studies (Nieto-Sanchez et al., 2015; Tatem et al., 2014; Wesolowski et al., 2014; Wesolowski et al., 2015) used CDRs to map the travel patterns of individuals who are members of disease "sources"(areas with many reported disease cases), in order to identify areas vulnerable to transmission, known as "sinks" (areas having high inflows of individuals from source areas). The aim of such studies was to prioritize source and sink areas for disease control. PDs in this case would be "sink" areas with very high potential for transmission due to high travel inflows from source areas, yet having only a small number of reported cases. *Understanding and identifying the measures and the factors that were behind this deviance could provide valuable insights into successful disease control for other infected areas*.
- Urban Resilience/Planning: In Zhang et al. (2016), factors affecting waterlogging in one city were used to predict waterlogging in another city using satellite imagery, precipitation meteorological data, terrain data, and road maps. *Positive deviance could be used to investigate why certain areas (PDs) within the same city experience less frequent waterlogging than others, and using those factors (eg, infrastructure and road networks) for better urban planning*.
- Academic Research: In Surjandari et al. (2015), Indonesian scholars' publications indexed in Scopus were analysed to map their primary research themes and advise on a nationwide research roadmap. *One could aggregate those publications by department, and identify departments with exceptionally high research publication quality (PDs) as proxied by citation indicators in Scopus or equivalent sources, eg, using the average h-index*[7] *of the publishing authors. Understanding factors leading to better publication quality would provide insights into departmental-level good practices that could be adopted in other departments*.

[7]Where *h* is the highest number where a scholar has *h* papers that have been cited at least *h* times.

- Deforestation: In Burgess et al. (2012), satellite images were used to identify forested districts in Indonesia that are practising illegal logging. *Positive deviance could be applied to identify districts with minimal illegal logging activities (PDs) and then investigate the measures and practices in place within those districts that are linked to that minimization*.
- Agriculture: In Tesfaye et al. (2016), crop, soil, and climate data were used to assess the performance of new drought-tolerant crop varieties. *These data could be used to identify those smallholder farms with high productivity (PDs), ie, high output levels from drought-tolerant crops. Using these or other big datasets, or survey data, one could infer good practices that could be adopted by neighbouring farms facing the same social, economic, and environmental constraints*.

5.2 | PD as an opportunity for BD4D applications

The primary interest of this paper is that outlined in the previous section: to identify current challenges to positive deviance action research and to identify ways in which big data-based positive deviance might address those challenges. But we can also reverse the polarity of the investigation and ask what positive deviance might offer the subfield of big data for development. Reviewing Section 4.7, there is little or nothing that positive deviance can do to address issues of representational validity, human capacity, data accessibility, or privacy and legal issues. Instead, these represent constraints that BDPD would have to contend with.

But those working on big data do identify a "need to develop methodologies to characterize and detect socioeconomic anomalies in context" (UNGP, 2012). Use of positive deviance to analyse big datasets and to detect anomalies in context cannot be said to provide a theoretical foundation in an academic sense, but it does offer a conceptual frame and a systematic methodology—a theory of change—that can link big data to development outcomes, something which has typically been missing to date. And at least if all five steps of the positive deviance approach were undertaken, then BDPD helps big data for development move beyond just proof of concept, by creating a real-world impact from the analysis of big data.

5.3 | Towards big data-based positive deviance analysis

In summary, we have a two-sided argument in favour of a big data-based positive deviance approach. Particularly, using big data instead of—or in conjunction with—traditional primary data sources can potentially address many of the challenges currently faced by positive deviance: reducing time, cost, and effort; identifying positive deviants in new or better ways; and enabling positive deviance to break out of its current path dependencies. And, conversely, positive deviance provides a systematic basis for extracting real-world development impacts from big data by putting knowledge about anomalies into action.

We can summarize big data-based positive deviance as follows:

> *The BDPD approach* is a problem-solving asset-based approach that uses big data sources to identify objects (positive deviants) performing unexpectedly well in a specific outcome measure that is digitally recorded, mediated, or observed. The primary objective of the BDPD approach is to identify the behaviours, strategies, and factors employed by the positive deviants and develop interventions to facilitate the dissemination and adoption of those strategies.
>
> *BDPD objects*—the positive deviants—could be individuals, communities, entities, areas, or countries whose uncommon behaviours and strategies, in a specific context, can be translated into a performance measure that is digitally recorded, mediated or observed.

We end with Table 3, which compares the PD and BDPD approaches in terms of the data sources used, the type of anomalies detected, the possible units of analysis, and the employed research methods and techniques.

We will be taking forward work applying BDPD, and we hope that other development researchers and practitioners may be encouraged to do the same.

TABLE 3 Comparison between the positive deviance (PD) and the big data for development (BDPD) approach

	Data Sources Used	Type of Positive Deviants	Unit of Analysis	Research Methods	Data Analysis Methods
PD	Surveys, focus groups, interviews, observations	Point anomalies	Individuals and entities	Qualitative, quantitative or mixed	Statistical methods
BDPD	Government data, online incl. social media data, mobile data, physical sensor incl. remote sensing data, offline and online surveys, focus groups, interviews, observations	Point and contextual (time and spatial) anomalies	Individuals, entities, communities, regions, countries, etc.	Quantitative or mixed	Advanced analytics and statistical methods

ORCID

Basma Albanna http://orcid.org/0000-0002-6884-1610

Richard Heeks http://orcid.org/0000-0002-4551-2208

REFERENCES

Aday, J., Hyden, A., Osking, J., & Tomedi, A. (2016). Hygiene, sanitation, and behaviors that produce positive deviant outcomes in childhood growth in rural eastern Kenya: A qualitative positive deviant investigation. *Annals of Global Health*, *82*(3), 437. https://doi.org/10.1016/j.aogh.2016.04.212

Ahrari, M., Kuttab, A., Khamis, S., Farahat, A. A., Darmstadt, G. L., Marsh, D. R., & Levinson, F. J. (2002). Factors associated with successful pregnancy outcomes in upper Egypt: A positive deviance inquiry. *Food and Nutrition Bulletin*, *23*(1), 83–88. https://doi.org/10.1177/156482650202300111

Andrews, M. (2015). Explaining positive deviance in public sector reforms in development. *World Development*, *74*, 197–208. https://doi.org/10.1016/j.worlddev.2015.04.017

Anino, O. C., Were, G. M., & Khamasi, J. W. (2015). Impact evaluation of positive deviance hearth in Migori County, Kenya, African. *Journal of Food, Agriculture, Nutrition and Development*, *15*(5), 10578–10596.

Aruna, M., Vazir, S., & Vidyasagar, P. (2001). Child rearing and positive deviance in the development of preschoolers: A microanalysis. *Indian Pediatrics*, *38*(4), 332–339.

Awofeso, N., Irwin, T., & Forrest, G. (2008). Using positive deviance techniques to improve smoking cessation outcomes in New South Wales prison settings. *Health Promotion Journal of Australia*, *19*(1), 72–73.

Babalola, S. (2007). Motivation for late sexual debut in Cote d'Ivoire and Burkina Faso: A positive deviance inquiry. *Journal of HIV/AIDS Prevention in Children and Youth*, *7*(2), 65–87. https://doi.org/10.1300/J499v07n02_05

Babalola, S., Awasum, D., & Quenum-Renaud, B. (2002). The correlates of safe sex practices among Rwandan youth: A positive deviance approach. *African Journal of AIDS Research*, *1*(1), 11–21. https://doi.org/10.2989/16085906.2002.9626540

Bamberger, M. (2016). *Integrating big data into the monitoring and evaluation of development programmes*. New York, NY: UN Global Pulse.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., … Portugali, Y. (2012). Smart cities of the future. *European Physical Journal: Special Topics*, *214*(1), 481–518.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS Medicine*, *8*(8), 1–9.

Birhanu, M. Y., Girma, A., & Puskur, R. (2017). Determinants of success and intensity of livestock feed technologies use in Ethiopia: Evidence from a positive deviance perspective. *Technological Forecasting and Social Change*, *115*, 15–25. https://doi.org/10.1016/j.techfore.2016.09.010

Blumenstock, J. E., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Blumenstock, J. E., & Eagle, N. (2012). Divided we call: Disparities in access and use of mobile phones in Rwanda. *Information Technologies and International Development*, *8*(2), 1–16.

Bolles, K., Speraw, C., Berggren, G., & Lafontant, J. G. (2002). Ti foyer (hearth) community-based nutrition activities informed by the positive deviance approach in Leogane, Haiti: A programmatic description. *Food and Nutrition Bulletin*, *23*(4 Supp), 11–17.

Bradley, E. H., Curry, L. A., Ramanadhan, S., Rowe, L., Nembhard, I. M., & Krumholz, H. M. (2009). Research in action: Using positive deviance to improve quality of health care. *Implementation Science*, *4*, 25. https://doi.org/10.1186/1748-5908-4-25

Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E., & Snow, R. W. (2013). Mobile phones and malaria: Modeling human and parasite travel. *Travel Medicine and Infectious Disease*, *11*(1), 15–22. https://doi.org/10.1016/j.tmaid.2012.12.003

Burgess, R., Hansen, M., Olken, B. A., Potapov, P., & Sieber, S. (2012). The political economy of deforestation in the tropics. *The Quarterly Journal of Economics*, *127*(4), 1707–1754. https://doi.org/10.1093/qje/qjs034

Cartesian (2014). *Using Mobile Data for Development*. Boston, MA: Cartesian.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chandy, R., Hassan, M., & Mukherji, P. (2017). Big data for good: Insights from emerging markets. *Journal of Product Innovation Management*, *34*(5), 703–713. https://doi.org/10.1111/jpim.12406

D'Alimonte, M. R., Deshmukh, D., Jayaraman, A., Chanani, S., & Humphries, D. L. (2016). Using positive deviance to understand the uptake of optimal infant and young child feeding practices by mothers in an urban slum of Mumbai. *Maternal and Child Health Journal*, *20*(6), 1133–1142. https://doi.org/10.1007/s10995-015-1899-3

de Macedo, R. C., Jacob, E. M., Silva, V. P., Santana, E. A., Souza, A. F., Gonçalves, P., … Edmond, M. B. (2012). Positive deviance: Using a nurse call system to evaluate hand hygiene practices. *American Journal of Infection Control*, *40*(10), 946–950. https://doi.org/10.1016/j.ajic.2011.11.015

Delias, P. (2017). A positive deviance approach to eliminate wastes in business processes. *Industrial Management & Data Systems*, *117*(7), 1323–1339. https://doi.org/10.1108/IMDS-09-2016-0393

Doll, C. N. H., & Pachauri, S. (2010). Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy*, *38*(10), 5661–5670. https://doi.org/10.1016/j.enpol.2010.05.014

Escobar, N. M. O. Márquez IA, Quiroga JA, Trujillo TG, González F, Aguilar MI, Escobar-Pérez J. (2017) Using positive deviance in the prevention and control of MRSA infections in a Colombian hospital: A time-series analysis, *Epidemiology and Infection*, *145*(5), 981–989, DOI: https://doi.org/10.1017/S095026881600306X.

Felt, L. J. (2011) Present promise, future potential: Positive deviance and complementary theory. Unpublished manuscript. http://www.laurelfelt.org/wp-content/uploads/2011/06/PositiveDeviance-CodyMayer.LaurelFelt.Quals_.May2011.pdf

Frias-Martinez, V., & Virseda, J. (2013). Cell phone analytics: Scaling human behavior studies into the millions. *Information Technologies and International Development*, *9*(2), 35–50.

Garrett, J. J., & Barrington, C. (2013). 'We do the impossible': Women overcoming barriers to cervical cancer screening in rural Honduras—A positive deviance analysis. *Culture, Health & Sexuality*, *15*(6), 637–651. https://doi.org/10.1080/13691058.2012.760206

Gartner (2013). *Big Data*. Stamford, CT: Gartner. https://www.gartner.com/it-glossary/big-data

Gidado, M., Obasanya, J. O., Adesigbe, C., Huji, J., & Tahir, D. (2010). Role of positive deviants among leprosy self-care groups in leprosy settlement, Zaria, Nigeria. *Journal of Community Medicine and Primary Health Care*, *22*(1–2).

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255–261. https://doi.org/10.1177/2043820613513121

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *British Medical Journal*, *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Guldan, G. S., Zhang, M. Y., Zhang, Y. P., Hong, J. R., Zhang, H. X., Fu, S. Y., & Fu, N. S. (1993). Weaning practices and growth in rural sichuan infants: A positive deviance study. *Journal of Tropical Pediatrics*, *39*(3), 168–175. https://doi.org/10.1093/tropej/39.3.168

Haas, E. M., Bartholomé, E., & Combal, B. (2009). Time series analysis of optical remote sensing data for the mapping of temporary surface water bodies in sub-Saharan western Africa. *Journal of Hydrology*, *370*(1–4), 52–63. https://doi.org/10.1016/j.jhydrol.2009.02.052

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, *102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Hendrickson, J. L., Dearden, K., Pachón, H., An, N. H., Schroeder, D. G., & Marsh, D. R. (2002). Empowerment in rural Viet Nam: Exploring changes in mothers and health volunteers in the context of an integrated nutrition project. *Food and Nutrition Bulletin*, *23*(4 Supp), 86–94.

Herington, M. J., & van de Fliert, E. (2018). Positive deviance in theory and practice: A conceptual review. *Deviant Behavior*, *39*(5), 664–678. https://doi.org/10.1080/01639625.2017.1286194

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development and Policy Review*, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

Huang, Z., Cao, F., Jin, C., Yu, Z., & Huang, R. (2017). Carbon emission flow from self-driving tours and its spatial relationship with scenic spots—A traffic-related big data method. *Journal of Cleaner Production*, *142*, 946–955. https://doi.org/10.1016/j.jclepro.2016.09.129

Ifaei, P., Karbassi, A., Lee, S., & Yoo, C. (2017). A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data. *Energy Conversion and Management*, *153*(October), 257–277. https://doi.org/10.1016/j.enconman.2017.10.014

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Jiang, W., Jia, K., Chen, Z., Deng, Y., & Rao, P. (2017). Using spatiotemporal remote sensing data to assess the status and effectiveness of the underground coal fire suppression efforts during 2000–2015 in Wuda, China. *Journal of Cleaner Production*, *142*, 565–577. https://doi.org/10.1016/j.jclepro.2016.03.082

Kanani, S., & Popat, K. (2012). Growing normally in an urban environment: Positive deviance among slum children of Vadodara, India. *Indian Journal of Pediatrics*, *79*(5), 606–611. https://doi.org/10.1007/s12098-011-0612-9

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS One*, *9*(5), e93949. https://doi.org/10.1371/journal.pone.0093949

Kim, Y. M., Heerey, M., & Kols, A. (2008). Factors that enable nurse-patient communication in a family planning context: A positive deviance study. *International Journal of Nursing Studies*, *45*(10), 1411–1421. https://doi.org/10.1016/j.ijnurstu.2008.01.002

Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, *36*(2), 154–163. https://doi.org/10.1016/j.compenvurbsys.2011.11.001

Lackovich-Van Gorp, A. (2017). Unearthing local forms of child protection: Positive deviance and abduction in Ethiopia. *Action Research*, *15*(1), 39–52. https://doi.org/10.1177/1476750316679240

Lapping, K., Marsh, D. R., Rosenbaum, J., Swedberg, E., Sternin, J., Sternin, M., & Schroeder, D. G. (2002). The positive deviance approach: Challenges and opportunities for the future. *Food and Nutrition Bulletin*, *23*(4 Supp), 128–135. https://doi.org/10.1177/15648265020234S217

Lapping, K., Schroeder, D., Marsh, D., Albalak, R., & Jabarkhil, M. Z. (2002). Comparison of a positive deviant inquiry with a case-control study to identify factors associated with nutritional status among Afghan refugee children in Pakistan. *Food and Nutrition Bulletin*, *23*(4 Supp2), 26–33.

LeMahieu, P. G., Nordstrum, L. E., & Gale, D. (2017). Positive deviance: Learning from positive anomalies. *Quality Assurance in Education*, *25*(1), 109–124. https://doi.org/10.1108/QAE-12-2016-0083

Letouzé, E. & Jütting, J. (2015) Official statistics, big data and human development: *Towards a New Conceptual and Operational Approach*. Data Pop Alliance. http://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf

Levinson, F. J., Barney, J., Bassett, L., & Schultink, W. (2007). Utilization of positive deviance analysis in evaluating community-based nutrition programs: An application to the Dular program in Bihar, India. *Food and Nutrition Bulletin*, *28*(3), 259–265. https://doi.org/10.1177/156482650702800301

Li, S., Ye, X., Lee, J., Gong, J., & Qin, C. (2017). Spatiotemporal analysis of housing prices in China: A big data perspective. *Applied Spatial Analysis and Policy*, *10*(3), 421–433. https://doi.org/10.1007/s12061-016-9185-3

Long, K. N., Gren, L. H., Rees, C. A., West, J. H., Hall, P. C., Gray, B., & Crookston, B. T. (2013). Determinants of better health: A cross-sectional assessment of positive deviants among women in West Bengal. *BMC Public Health*, *13*, 372. https://doi.org/10.1186/1471-2458-13-372

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576–11581. https://doi.org/10.1073/pnas.1203882109

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., … Bengtsson, L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, *38*, 1–7. https://doi.org/10.1016/j.gloenvcha.2016.02.002

Luo, X., Dong, L., Dou, Y., Zhang, N., Ren, J., Li, Y., … Yao, S. (2017). Analysis on spatial-temporal features of taxis' emissions from big data informed travel patterns: A case of Shanghai, China. *Journal of Cleaner Production*, *142*, 926–935. https://doi.org/10.1016/j.jclepro.2016.05.161

Mackintosh, U. A., Marsh, D. R., & Schroeder, D. G. (2002). Sustained positive deviant child care practices and their effects on child growth in Viet Nam. *Food and Nutrition Bulletin*, *23*(4 Supp), 18–27.

Marra, A. R., Guastelli, L. R., de Araújo, C. M. P., dos Santos, J. L. S., Lamblet, L. C. R., Silva, M., … dos Santos, O. F. P. (2010). Positive deviance: A new strategy for improving hand hygiene compliance. *Infection Control and Hospital Epidemiology*, *31*(1), 12–20. https://doi.org/10.1086/649224

Marra, A. R., Noritomi, D. T., Westheimer Cavalcante, A. J., Sampaio Camargo, T. Z., Bortoleto, R. P., Durao Junior, M. S., … Positive Deviance For Hand Hygiene Study Group (2013). A multicenter study using positive deviance for improving hand hygiene compliance. *American Journal of Infection Control*, *41*(11), 984–988. https://doi.org/10.1016/j.ajic.2013.05.013

Marra, A. R., Reis Guastelli, L., Pereira de Araújo, C. M., Saraiva dos Santos, J. L., Filho, M. A. O., Silva, C. V., … Edmond, M. B. (2011). Positive deviance: A program for sustained improvement in hand hygiene compliance. *American Journal of Infection Control*, *39*(1), 1–5. https://doi.org/10.1016/j.ajic.2010.05.024

Marsh, D. R., Schroeder, D. G., Dearden, K. A., Sternin, J., & Sternin, M. (2004). The power of positive deviance. *British Medical Journal*, *329*(7475), 1177–1179. https://doi.org/10.1136/bmj.329.7475.1177

Marsh, D. R., Sternin, M., Khadduri, R., Ihsan, T., Nazir, R., Bari, A., & Lapping, K. (2002). Identification of model newborn care practices through a positive deviance inquiry to guide behavior-change interventions in Haripur, Pakistan. *Food and Nutrition Bulletin*, *23*(4 Supp), 109–118.

Mashey, J. R. (1998) Big data and the next wave of infrastress, paper presented at *Computer Systems Laboratory Colloquium*, Stanford University, Ca, 25 Feb.

Merchant, S. S., & Udipi, S. A. (1997). Positive and negative deviance in growth of urban slum children in Bombay. *Food and Nutrition Bulletin*, *18*(4), 323–336.

Merita, M., Sari, M. T., & Hesty, H. (2017). The positive deviance of feeding practices and carring with nutritional status of toddler among poor families. *Jurnal Kesehatan Masyarakat*, *13*(1), 106–112. https://doi.org/10.15294/kemas.v13i1.7919

Mertens, W., Recker, J., Kohlborn, T., & Kummer, T.-F. (2016). A framework for the study of positive deviance in organizations. *Deviant Behavior*, *37*(11), 1288–1307. https://doi.org/10.1080/01639625.2016.1174519

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *British Medical Journal*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

Mügge, D. (2014). Poor numbers: How we are misled by African development statistics and what to do about it, by Morten Jerven. *Review of International Political Economy*, *21*(5), 1131–1133. https://doi.org/10.1080/09692290.2014.953354

Ndiaye, M., Siekmans, K., Haddad, S., & Receveur, O. (2009). Impact of a positive deviance approach to improve the effectiveness of an iron-supplementation program to control nutritional anemia among rural Senegalese pregnant women. *Food and Nutrition Bulletin*, *30*(2), 128–136. https://doi.org/10.1177/156482650903000204

Nel, H. (2018). A comparison between the asset-oriented and needs-based community development approaches in terms of systems changes. *Practice*, *30*(1), 33–52. https://doi.org/10.1080/09503153.2017.1360474

Nieto-Sanchez, C., Baus, E. G., Guerrero, D., & Grijalva, M. J. (2015). Positive deviance study to inform a Chagas disease control program in southern Ecuador. *Memórias do Instituto Oswaldo Cruz*, *110*(3), 299–309. https://doi.org/10.1590/0074-02760140472

Nishat, N., & Batool, I. (2011). Effect of 'Positive Hearth Deviance' on feeding practices and underweight prevalence among children aged 6-24 months in Quetta district, Pakistan: A comparative cross sectional study. *Sri Lanka Journal of Child Health*, *40*(2), 57–62. https://doi.org/10.4038/sljch.v40i2.3093

Njuguna, C., & McSharry, P. (2017). Constructing spatiotemporal poverty indices from big data. *Journal of Business Research*, *70*, 318–327. https://doi.org/10.1016/j.jbusres.2016.08.005

OECD (2017). *DAC list of ODA recipients*. Paris: Organisation for Economic Co-operation and Development.

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the AIS*, *37*, 879–910.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6), 1–8.

Pascale, R., Sternin, J., & Sternin, M. (2010). *The power of positive deviance: How unlikely innovators solve the world's toughest problems*. Boston, MA: Harvard Business Press.

Petticrew, M., & Roberts, H. (2005). *Systematic reviews in the social sciences: A practical guide*. Chichester, UK: John Wiley & Sons.

Pfeffer, K., Verrest, H., & Poorthuis, A. (2015). Big data for better urban life?—An exploratory study of critical urban issues in two Caribbean cities: Paramaribo (Suriname) and Port of Spain (Trinidad and Tobago). *European Journal of Development Research*, *27*(4), 505–522.

Positive Deviance Initiative (2010). *Basic field guide to the positive deviance approach*. Medford, MA: Positive Deviance Initiative, Tufts University.

Roche, M. L., Marquis, G. S., Gyorkos, T. W., Blouin, B., Sarsoza, J., & Kuhnlein, H. V. (2017). A community-based positive deviance/hearth infant and young child nutrition intervention in ecuador improved diet and reduced underweight. *Journal of Nutrition Education and Behavior*, *49*(3), 196–203. https://doi.org/10.1016/j.jneb.2016.10.007

Saïd Business School (2010). *Exploring positive deviance—New frontiers in collaborative change*. Oxford, UK: Saïd Business School, University of Oxford.

Sengupta, R., Heeks, R., Chattapadhyay, S. & Foster, C. (2017) Exploring big data for development: An electricity sector case study from India, Development Informatics Working Paper 66. Manchester, UK: Global Development Institute, University of Manchester.

Sethi, V., Kashyap, S., Seth, V., & Agarwal, S. (2003). Encouraging appropriate infant feeding practices in slums: A positive deviance approach. *Pakistan Journal of Nutrition*, *2*(3), 164–166.

Sethi, V., Sternin, M., Sharma, D., Bhanot, A., & Mebrahtu, S. (2017). Applying positive deviance for improving compliance to adolescent anemia control program in tribal communities of India. *Food and Nutrition Bulletin*, *38*(3), 447–452. https://doi.org/10.1177/0379572117712791

Shekar, M., Habicht, J. P., & Latham, M. C. (1991). Is positive deviance in growth simply the converse of negative deviance? *Food and Nutrition Bulletin*, *13*(1), 7–11.

Shekar, M., Habicht, J. P., & Latham, M. C. (1992). Use of positive-negative deviant analyses to improve programme targeting and services: Example from the Tamil Nadu integrated nutrition project. *International Journal of Epidemiology*, *21*(4), 707–713. https://doi.org/10.1093/ije/21.4.707

Shoenberger, N. A. (2017). Bridging normative and reactivist perspectives: An introduction to positive deviance. In S. E. Brown, & O. Sefina (Eds.), *Routledge handbook on deviance* (pp. 42–51). Abingdon, UK: Routledge.

Singhal, A. (2011). Turning diffusion of innovation paradigm on its head: The positive deviance approach to social change. In A. Vishwanath, & G. A. Barnett (Eds.), *The diffusion of innovations* (pp. 193–205). New York, NY: Peter Lang.

Song, M., & Wang, S. (2017). Participation in global value chain and green technology progress: Evidence from big data of Chinese enterprises. *Environmental Science and Pollution Research*, *24*(2), 1648–1661. https://doi.org/10.1007/s11356-016-7925-1

Springer, A., Nielsen, C., & Johansen, I. (2016). *Positive deviance by the numbers*. Medford, MA: Positive Deviance Initiative, Tufts University.

Sternin, J. (2002). Positive deviance: A new paradigm for addressing today's problems today. *The Journal of Corporate Citizenship*, *5*(Spring), 57–63.

Sternin, J., & Choo, R. (2000). The power of positive deviancy. *Harvard Business Review*, *78*(1), 1–3.

Sternin, M., Sternin, J., & Marsh, D. (1997). *Rapid sustained childhood malnutrition alleviation through a positive-deviance approach in rural Vietnam: Preliminary findings*. Arlington, VA: Partnership for Child Health Care, BASICS.

Sternin, M., Sternin, J., & Marsh, D. (1998). *Designing a community-based nutrition program using the hearth model and the positive deviance approach: A field guide*. Westport, CT: Save the Children.

Su, S., Li, Z., Xu, M., Cai, Z., & Weng, M. (2017). A geo-big data approach to intra-urban food deserts: Transit-varying accessibility, social inequalities, and implications for urban planning. *Habitat International*, *64*, 22–40. https://doi.org/10.1016/j.habitatint.2017.04.007

Surjandari, I., Dhini, A., Lumbantobing, E. W. I., Widari, A. T., & Prawiradinata, I. (2015). Big data analysis of indonesian scholars' publications: A research theme mapping. *International Journal of Technology*, *6*(4), 650–658. https://doi.org/10.14716/ijtech.v6i4.1956

Sutton, P., Elvidge, C., & Ghosh, T. (2007). Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. *International Journal of Ecological Economics & Statistics*, *8*(S07), 5–21.

Tang, H., Yang, X., & Zhang, Y. (2014). Effort at constructing big data sensor networks for monitoring greenhouse gas emission. *International Journal of Distributed Sensor Networks*, *10*(7), 619608. https://doi.org/10.1155/2014/619608

Tatem, A. J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D. K., … Lourenço, C. (2014). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria Journal*, *13*, 52.

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLoS One*, *2*(12), e1298. https://doi.org/10.1371/journal.pone.0001298

Tatem, A. J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S., & Moonen, B. (2009). The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among Zanzibar residents. *Malaria Journal*, *8*, 287. https://doi.org/10.1186/1475-2875-8-287

Taylor, L., & Broeders, D. (2015). In the name of development: Power, profit and the datafication of the global South. *Geoforum*, *64*, 229–237. https://doi.org/10.1016/j.geoforum.2015.07.002

Tekle, L. (2015). Analysis of positive deviance farmer training centers in Northern Ethiopia. *American Journal of Rural Development*, *3*(1), 10–14.

Tesfaye, K., Sonder, K., Cairns, J., Magorokosho, C., Tarekegn, A., Kassie, G. T., … Erenstein, O. (2016). Targeting drought-tolerant maize varieties in Southern Africa: A geospatial crop modeling approach using big data. *International Food and Agribusiness Management Review*, *19*(A), 75–92.

Tong, T., & Li, H. (2017). Demand for MOOC—An application of big data. *China Economic Review*. in press, DOI: https://doi.org/10.1016/j.chieco.2017.05.007

UNGP (2012). *Big data for development: Challenges & opportunities*. New York, NY: UN Global Pulse.

Vossenaar, M., Bermúdez, O. I., Anderson, A. S., & Solomons, N. W. (2010). Practical limitations to a positive deviance approach for identifying dietary patterns compatible with the reduction of cancer risk. *Journal of Human Nutrition and Dietetics*, *23*(4), 382–392. https://doi.org/10.1111/j.1365-277X.2010.01056.x

Vossenaar, M., Mayorga, E., Soto-Méndez, M:. J:., Medina-Monchez, S. B., Campos, R., Anderson, A. S., & Solomons, N. W. (2009). The positive deviance approach can be used to create culturally appropriate eating guides compatible with reduced cancer risk. *Journal of Nutrition*, *139*(4), 755–762. https://doi.org/10.3945/jn.108.100362

Wesolowski, A., Metcalf, C. J. E., Eagle, N., Kombich, J., Grenfell, B. T., Bjørnstad, O. N., … Buckee, C. O. (2015). Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, *112*(35), 11114–11119. https://doi.org/10.1073/pnas.1423542112

Wesolowski, A. et al. (2014) Quantifying travel behavior for infectious disease research: A comparison of data from surveys and mobile phones, Scientific Reports, 4, 5678.

Wilson, C. (2011). Digital media in the Egyptian revolution: Descriptive analysis from the Tahrir data sets. *International Journal of Communication*, *5*, 1248–1272.

Wilson, R. et al. (2016) Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake, *PLoS Currents Disasters*, 24 Feb, 8.

Wishik, S. M., & Van Der Vynckt, S. (1976). The use of nutritional "positive deviants" to identify approaches for modification of dietary practices. *American Journal of Public Health*, *66*(1), 38–42. https://doi.org/10.2105/AJPH.66.1.38

Wollinka, O., Keeley, E., Burkhalter, B. R., & Bashir, N. (1997). *Hearth nutrition model: Applications in Haiti, Viet Nam and Bangladesh*. Arlington, VA: Partnership for Child Health Care, BASICS.

WP (2018) Open Data. Wikipedia. https://en.wikipedia.org/wiki/Open_data

Zaidi, Z., Jaffery, T., Shahid, A., Moin, S., Gilani, A., & Burdick, W. (2012). Change in action: Using positive deviance to improve student clinical performance. *Advances in Health Sciences Education*, *17*(1), 95–105. https://doi.org/10.1007/s10459-011-9301-8

Zeitlin, M. (1991). Nutritional resilience in a hostile environment: Positive deviance in child nutrition. *Nutrition Reviews*, *49*(9), 259–268.

Zeitlin, M., Ghassemi, H., & Mansour, M. (1994). Positive deviance in child nutrition: A discussion. *Ecology of Food and Nutrition*, *31*(3–4), 295–302. https://doi.org/10.1080/03670244.1994.9991371

Zeitlin, M. F., Ghassemi, H., Mansour, M., & United Nations University & Joint WHO/UNICEF Nutrition Support Programme (1990). *Positive deviance in child nutrition: With emphasis on psychosocial and behavioural aspects and amplications for development*. Tokyo: United Nations University.

Zhang, N., Chen, H., Chen, J. & Chen, X. (2016) Social media meets big urban data: A case study of urban waterlogging analysis, Computational Intelligence and Neuroscience, 2016.

AUTHOR BIOGRAPHIES

Basma Albanna is a Development Informatics PhD student at the Global Development Institute, University of Manchester and an assistant lecturer at the faculty of computer science Ain Shams University, Cairo, Egypt. She has been working in development organizations since 2012, particularly in supporting social innovations. She has publications in semantic trajectories and location-based social networks. Basma holds a Master's degree in computer science and another Master's degree in management of technology.

Richard Heeks is Chair in Development Informatics at the Global Development Institute, University of Manchester and Director of the Centre for Development Informatics (http://www.cdi.manchester.ac.uk). He has been consulting and researching on informatics and development for more than 30 years. His book publications include India's Software Industry (1996), Reinventing Government in the Information Age (1999), Implementing and Managing eGovernment (2006), ICTs, Climate Change and Development (2012), and Information and Communication Technology for Development (2018). His research interests are data-intensive development, e-resilience and e-sustainability, digital development, and the digital economy in developing countries. He has a PhD in Indian IT industry development, directs the MSc programme in ICTs for Development, and runs the ICT for Development blog: http://ict4dblog.wordpress.com.

How to cite this article: Albanna B, Heeks R. Positive deviance, big data, and development: A systematic literature review. *E J Info Sys Dev Countries*. 2019;85:e12063. https://doi.org/10.1002/isd2.12063