Record Linkage in Statistical Sampling: Past, Present, and Future

Benjamin Williams

Abstract Record linkage is a useful tool to match records across datasets when the datasets lack a unique identifier. In this chapter, we examine the past, current, and present uses of probabilistic record linkage with a specific interest in its use in statistical sampling. For example, given the rise in interest and use of non-probability data within sampling, many researchers seek to augment a non-probability sample with a probability sample. Record linkage is a useful method for doing such combining. This chapter will examine the ways record linkage has been used and is currently being researched and implemented, with an emphasis on its current and future use for statistical sampling. The chapter concludes with open research questions for record linkage in the context of sampling, where the questions center around the idea of creating a total error framework for linked data.

1 Introduction

Analysts broadly use the term *record linkage* to define the matching of records existing in two or more datasets. Record linkage is also used for data deduplication, but that is not the focus of this chapter. Here, record linkage encompasses other commonly used terms for data matching, including but not limited to entity resolution, data blending, data combination, document linkage, and record matching. Originally describing the process of combining specific life event records (e.g., birth, graduation, marriage) in a person's "Book of Life" (Dunn, 1946), record linkage has grown in breadth over the past 75 years and is an active area of statistical research. From its humble roots, record linkage has been mathematically formalized, implemented with machine learning, and employed at numerous public and private agencies (Herzog et al., 2007; Christen, 2019; Dong & Srivastava, 2015).

B. Williams (✉)
University of Denver, Denver, CO, USA
e-mail: benjamin.williams@du.edu

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_9

Record linkage is of use when two or more data files refer to the same entity yet lack a unique identifier common among all sources. In this chapter, without loss of generality, assume there are two files to link; call these files A and B. Record linkage relies on comparing *linking variables*, variables present in both A and B which should be equivalent for matching records. Newcombe et al. (1959) note two issues arising from comparing linking variables: (1) two records that *do not* refer to the same entity may have equivalent linking variable values (e.g., Ben Williams and Ben Leonard have equivalent first names, but may be different people), and (2) two records that *do* refer to the same entity may have different linking variable values (e.g., Benjamin Williams and Ben Williams could be the same person, but have different recorded first names). Record linkage can mitigate these issues.

Record linkage has two primary forms: deterministic and probabilistic (Herzog et al., 2007). A deterministic program links records across datasets via strict, predetermined rules concerning linking variables. An example is as follows: only link two entities if the recorded last names are equivalent and the recorded dates are within 2 days of each other. Deterministic record linkage can work well if there are few or no errors in the datasets. Probabilistic record linkage relies on the distribution of the linking variables to determine the likelihood two records match. Probabilistic record linkage is a powerful tool when there are possible errors in the datasets. Errors such as misspellings or incorrect recording of dates are quite common, making probabilistic record linkage popular. For the rest of this chapter, *record linkage* will refer to probabilistic record linkage.

In 1959, Newcombe et al. developed a linking score aggregating estimates of the log-odds that the values of the linking variables agree for each potential link between A and B (Newcombe et al., 1959). Their work was formalized in Fellegi and Sunter (1969). The Fellegi-Sunter implementation is the classic method of record linkage. They derived the linkage score for a pair of potential links by using the probabilities of observing agreement patterns in true matching and non-matching pairs of records. The expectation-maximization (EM) algorithm (Dempster et al., 1977) is often used to estimate the parameters for the score.

Potential links with a score above an upper threshold are called matches, potential links with a score below a lower threshold are called non-matches, and potential links with a score between the upper and lower thresholds are called potential matches. The thresholds, along with prespecified false-positive and false-negative rates, comprise a linking rule. Fellegi and Sunter proved this rule is optimal in the sense that it minimizes the probability a possible link is classified as a potential match as opposed to a match or a non-match. The rigorous method of combining datasets introduced by Fellegi and Sunter opened a new research context for record linkage: statistical sampling.

When a representative sample is drawn at random from a population, inference regarding the population can be made from inspection of the sample (Lohr, 2010). This is a foundational tenet of statistics. However, given the pervasive availability of big data, are large samples drawn not at random (non-probability samples) more useful than small probability samples? See Meng (2018) for a further discussion of this question. Indeed, large non-probability samples are easier than ever to collect,

but often at the cost of representativeness and theoretical formulae for sampling variability (Baker et al., 2013). Wiśniowski et al. (2020) examine the trade-offs between non-probability samples and probability samples. They argue combining a small probability sample with a larger non-probability sample allows one to harness the advantages of both. In this, record linkage becomes immensely valuable.

Integrating two samples may require records to be matched between them. If the probability sample adds auxiliary information, records from one sample likely need to be matched to records on the other. One example of this is a capture-recapture framework used to combine the non-probability and probability samples. If the initial capture sample is a non-probability sample and the recapture sample is a probability sample, the records from each sample must be matched for valid estimation (Liu et al., 2017; Stokes et al., 2021). In such cases one may use record linkage for matching. Another example of this is at the US Census Bureau, where smaller secondary samples are gathered after the census which are linked to the original data for additional inference.

In the US Census example, one of the datasets for linking is quite large, the US Census. Since the census is much larger than the second sample, and is nearly a complete register of the population, linking is easier as there is a high probability that respondents to the second sample exist in the census data. If one or both of the data files to be linked are small, relative to the population size, then the likelihood of finding units existing in both samples could be quite small rendering record linkage impractical and not useful.

However, given the pervasive nature in the world today, big data and datasets nearing the size of populations of interest are becoming more common. In cases where one or more of the datasets are relatively large, record linkage is most useful since the probability of a sizeable overlap is higher. The overlapping units are often where the benefit of combining samples comes from. For a treatment of identifying the overlap between a big data source and a smaller probability sample, see Kim and Tam (2021). Record linkage is an important tool to augmenting samples, be they non-probability or probability. This is a critical area of future research in statistical sampling.

This chapter examines the past and current uses of record linkage, along with opportunities for the method in the future. We pay particular attention to the use of record linkage in statistical sampling, especially in the sections on current and future uses. In the coming years, record linkage will play a key role in the analysis of non-probability samples, and open research questions exist which deserve careful consideration. This chapter will thus conclude by laying out these questions, discussing their critical nature, and offering paths toward solutions.

2 Past Uses of Record Linkage

Historically, record linkage has been primarily used to link records of people, businesses, or addresses (Fellegi, 1999). Often the linking variables are comprised

Data File A					Data File B				
Name	City	Birth Year	Marital Status	...	Name	City	Birth Year	Number of Children	...
Ben Williams	Denver	1991	Y		Ben Williams	Dallas	1989	1	
Brian William	Dallas	1990	N		Ben William	Denver	1991	0	
...					...				

Fig. 1 Example of two files to link some variable names which are the same across the files

of words (or strings). An example of two files to link is in Fig. 1. File *A* and *B* share the variables *Name*, *City*, and *Birth Year* and those are the linking variables. Suppose it is of interest to combine the files to determine the relationship between *Marital Status* (only in File *A*) and *Number of Children* (only in File *B*).

In Fig. 1, a human analyst could reasonably determine the first entry in File *A* (linking variable values: Ben Williams, Denver, 1991) matches the second entry in File *B* (linking variable values: Ben William, Denver, 1991) by observing the misspelling of Williams in the File *B* entry. In this toy example, the values of the *Birth Year* and *City* linking variables are exactly equivalent, but how can the differences in the *Name* linking variable be expressed? String comparator metrics are now well-known, and some resulted from the need to compare strings for matching purposes. Jaro (1989), Jaro (1995), and Winkler (1990) are seminal works which produced the Jaro-Winkler comparator, a metric producing a value between 0 and 1 to determine how similar two strings are. A thorough examination of the Jaro-Winkler comparator is in Herzog et al. (2007), and a deeper examination of more string comparators is in Cohen et al. (2003).

In an early implementation of computer-based record linkage, Newcombe et al. (1959) compared strings using the Russell Soundex Code, which breaks words into phonetic codes of numbers and letters. Those authors used record linkage to determine if health and fertility were affected by exposure to low levels of radiation. Since exposure, marriage, births, and illness information were contained in different files, there was a need to link them with variables common to all files. This is perhaps the earliest example of using computers to implement record linkage, marking a seismic shift in the ability to link large data files, since linking could be done automatically and not solely by hand. Indeed, the advent of computer technology is a key reason for the interest generated for record linkage beginning in the 1960s (Fellegi, 1999).

The work of Newcombe et al. (1959) was a motivator for the formative Fellegi-Sunter method discussed in the Introduction. After the establishment of their method, record linkage surged in popularity. Early use cases included matching insurance claims to medical statistics (Bell et al., 1994), immigration record matching (Copas and Hilton, 1990), and matching records for the Census Bureau (Mulry et al., 2006), to name but a few. If the two files to be linked are not complete enumerations of the populations they represent, inference resulting from

the linkage falls under the purview of sampling. For example, if the goal is to examine the relationship between marital status and number of children, as in the toy example from Fig. 1, because there is no complete list of everyone in the world along with their marital status, the files represent samples of people. When inference is made from the matches, the analyst is engaging in estimation resulting from samples. If the files are representative samples, then the inference is valid and well supported. Indeed, most statistical inference results from samples of data, so this is not necessarily an issue for record linkage. However, early record linkage literature lacks discussions regarding the assumption of representativeness in the datasets to be linked.

Another assumption often implicitly made in early record linkage papers is that errors in matching, e.g., false-positive and false-negative matches, do not affect the results of subsequent analyses. In the current research of record linkage, some effort is spent examining how these errors can affect the final analyses. Next, we discuss this along with current research and uses of record linkage.

3 Current Research and Uses of Record Linkage

Record linkage is currently used in medicine (Hallifax et al., 2018) and insurance (Boudreaux et al., 2015), at the Census Bureau (Abowd et al., 2019), and for big data fusion in general (Dong & Srivastava, 2015). Christen (2019) gives a useful and concise treatment of record linkage and includes additional current applications for further reading. Some of these applications have been studied since the inception of record linkage, but over time, research continues to expand the field.

One way the literature is expanding is in the methods used for record linkage, namely, via the introduction of machine learning techniques. The continued improvement in computing power combined with statistical techniques has allowed machine learning methods to be employed across industries and disciplines. Record linkage is no exception, as evidenced by Jurek et al. (2017) who introduced an ensemble learning method for unsupervised record linkage and Christen (2008) who developed a classification technique for record linkage involving support vector machines. There are many examples of machine learning used for record linkage since it can be distilled to a classification problem (match or non-match), a common use for machine learning. In addition to machine learning, Bayesian methods have also been introduced to record linkage. For example, Dalzell and Reiter (2016) took a Bayesian approach and derived a method to concurrently find matches and estimate the regression model.

In another avenue of current work, scholars are studying how the randomness associated with probabilistic linkage affects subsequent analyses. This was discussed in Neter et al. (1965), and it continues to be an area of active research. Recently, Chambers and Diniz da Silva (2020) noted (citing Harron et al., 2016) analysts' abilities to rigorously account for various biases and errors in linked data cannot keep pace with the inception of such datasets. Given the prevalence and

availability of big data, this is an important issue for study. Chambers and Diniz da Silva (2020) suggest using paradata (data about the linkage process) to correct for biases resulting from linkage errors.

An important paper regarding analyses done with linked data is Lahiri and Larsen (2005). These authors investigated how errors in linkage affect regression analysis done using the linked data. By handling linking errors as measurement errors, they proposed an unbiased bootstrap regression estimator for use when there are matching errors. Chipperfield and Chambers (2015) similarly derived a parametric bootstrap method for evaluating categorical variables from linked datasets. Chambers (2009) examined ways to remove bias in regression analysis resulting from linking errors and took a specific look at logistic regression as well. Additionally, Zhang and Tuoto (2021) developed a regression approach in the presence of linkage errors and offered a diagnostic hypothesis test for examining assumptions about the linkage errors. Chipperfield (2020) approaches this problem by using bootstrap methods to replicate the linkage procedure in each replicate, along with estimating equations, to make inference in the presence of linkage errors. In both Briscolini et al. (2018) and Salvati et al. (2021), the authors investigate several methods to handle linkage errors when the context is small area estimation. Last, Kim and Chambers (2012) develop ways of correcting for the bias due to linkage errors, including incomplete or missed links, when employing regression after linking sample data to a register (dataset of the entire population), which was discussed in Sect. 1.

Most work in this stream focuses on regression analyses of linked data. However, there are other inferential methods which use linked data, such as sampling estimation. Zhang (2021) recently developed several generalized regression estimators (GREG) (see Särndal et al., 1992) for estimating totals when the sample and the auxiliary information, used in GREG estimators, cannot be perfectly matched. Their work builds on research from Breidt et al. (2017) who examined a difference estimator (type of GREG estimator) when matching between samples is imperfect.

Stokes et al. (2021) similarly attempt to examine the effect of matching errors on estimates of total. In their work, the authors employed capture-recapture methodology where the capture sample was electronic self-reports of fish catch (non-probability sample) and the recapture sample was a randomized dockside intercept sample of anglers (probability sample). Record linkage was used to link the two samples, and then estimates of total were made from the linked data. The authors developed a theoretical model for the probability of linking specific records and derived an expression for the approximate relative bias of an estimator as a function of various levels of matching error (including false-positive and false-negative errors). The works of Stokes et al. (2021), Zhang (2021), and Breidt et al. (2017) discussed here represent a bridge to the future of record linkage in survey sampling.

4 Future Uses of Record Linkage and Open Questions

A bright future of record linkage in survey sampling exists in the combination of non-probability samples with probability samples. As noted in Wiśniowski et al. (2020), the benefits of blending a non-probability sample with a probability sample are substantial. Elliott and Haviland (2007) did this by combining estimators from a probability sample with a web-based non-probability sample. They note the probability sample must be large for useful estimation. Recently, Sakshaug et al. (2019) offered a Bayesian approach for analyzing data from a smaller probability sample blended with a larger non-probability sample. They used the non-probability samples to construct priors for the model and show their approach worked well to reduce mean square error in estimates even when bias was present in the non-probability samples, a usual concern when investigating non-probability samples. These papers, however, do not link specific observations across datasets (samples) but seek to harness the information from both samples to improve the overall estimation.

Often, for inference, the non-probability sample is adjusted or weighted to have similar characteristics as the target population or to be used as auxiliary information (Elliott, 2009; Brus & Gruijter, 2003; Valliant & Dever, 2011). Another framework is to link actual records appearing in two samples, one a probability sample and one a non-probability sample. This occurs if the non-probability sample and the probability sample are subsets of the same population with increased overlap between the two as the non-probability sample size grows.

Specifically, call the population of interest U, the set of observations comprising the probability sample s_p, and the set of observations comprising the non-probability sample s_{np}. Then $s_p \in U$ and $s_{np} \in U$ and as $|s_{np}| \to |U| \Rightarrow P(s_p \cap s_{np}) = \emptyset) \to 0$. By examining the overlapping observations between the two samples, inference can be improved. This is how Liu et al. (2017) approached the problem of estimating fish catch in the Gulf of Mexico when they combined a voluntary sample of captains' fishing reports with a random intercept of boats returning to the dock. The overlapping trips, trips both reported and intercepted, provide auxiliary information, namely, measurement error estimates, which is incorporated into the estimator. This is an example of combining samples via matching and is a great application for record linkage.

While Liu et al. (2017) operate in a capture-recapture framework, using record linkage to combine a non-probability and a probability sample need not exist in such a setting. Examining the overlap, the matched set of entities between the samples, can provide accurate and useful auxiliary information to be used along with current non-probability sampling methods such as pseudo-weights or propensity scores. As data from non-probability samples become more available in ever-increasing sizes, linking them to existing or new probability samples will become more and more feasible. Regardless of the final use, record linkage certainly has a role to play.

In the future, assuming record linkage takes an increasing role in non-probability sample inference, there are several research questions which should define the next

era of record linkage literature. We present a few open questions which should steer future research regarding record linkage in survey sampling.

The main research question of interest is: "what is the total error framework for linked data?" This question is closely linked to the idea of a total survey error (TSE) framework; see Groves and Lyberg (2010) for a thorough discussion of the TSE framework. The TSE framework decomposes the sources of error and bias when making inferences from surveys. This idea was recently extended in Amaya et al. (2020) for big data. They proposed a total error framework (TEF) for analyzing big data which has specific differences from the usual TSE framework. The authors discuss how certain errors manifest differently when applied to big data, such as coverage error, non-response error, and measurement error, to name a few (Amaya et al., 2020). Meng (2018) adopts a similar framework for making inferences from non-probability samples. He derived a formula to describe the difference between the population and sample averages as the product of measures of data quality, data quantity, and the problem difficulty (standard deviation of the variable of interest). Such previous research informs a TEF for linked data.

When analyzing linked data, a new source of randomness is introduced into the estimation which comes from linking errors. When considering a TEF for linked data, the linkage errors form a new component in the framework. The framework can be expressed as *Total Error* = *Sampling Error* + *Non-Sampling Error* + *Linkage Error*. Previous work has been done to examine both sampling error and non-sampling error in both the traditional, big data, and non-probability settings (Groves & Lyberg, 2010; Amaya et al., 2020; Meng, 2018). These three sources of error are broad and encompass many errors within them, e.g., *non-response error* is a subset of non-sampling error. Though these subsets have been investigated for sampling error and non-sampling error, there needs to be a partitioning of linkage error to build the TEF for linked data.

Stokes et al. (2021) started down this path by deriving a model for the effect linking errors have on the approximate relative bias of estimates made from linked data. Their model considers response rates and the discrepancies in the measurements when records are incorrectly linked. The model is generalizable and used to examine the effect of linking errors on the bias when estimating a total. Their work should be extended and further generalized to understand the effect of linkage errors within a total error framework. Linkage errors are especially difficult to partition because each linking scenario is different (Bell, 2017). Additionally, the magnitude of the effect of different linking errors will differ depending on various factors such as the amount of measurement error existing among matched records and if various errors can balance each other out (e.g., false-positive errors vs false-negative errors). Another source of linkage error that deserves further research is coverage error resulting from false-negative or unmatched links. That is, because some records are not linked, error arises. But this error is unique in such a context because the probability of linking two records can depend on the linkage algorithm (e.g., one-to-many linkage or one-to-one linkage) as well as the likelihood that other records link to each other.

A secondary question within the TEF for linked data has to do with estimating matching error if one lacks training data or the ability to perform clerical review. Training data offers a set of true links on which a record linkage algorithm can be tested. Clerical review is the term for manual inspection of potential links to determine if they match or not. Clerical review is usually the gold standard way to evaluate links if the entities refer to people or addresses, such as in the example from Fig. 1.

An example of when clerical review might be impossible is if an analyst links health data from wearable electronic devices to a census probability sample. In that case, manual review of links may prove too difficult to confidently mark links as false-positives, false-negatives, true-positives, or true-negatives. This might be true if the variables used for linking are error prone or if human judgment does not do a good job at determining true match status. Human judgment might also not be useful if no names or strings are used as linking variables, but instead identification numbers or usernames comprise the linking variables. In these settings, a sensitivity analysis for different levels of matching error will prove useful. In the future, a rigorous framework for such sensitivity analyses or methods of expressing confidence in the link states (match vs non-match) deserves careful thought as part of a TEF for linked data.

Another secondary question in this framework manifests when more than two files are to be linked. As stated earlier, the methodologies for linking two files extend to linking three or more files. However, it is likely that the data structures will differ for the different datasets. Each may have distinct and possibly different error sources. It may be that when linking three files (A, B, and C), a record $a \in A$ may be a false-positive link to record $b \in b$ but be a false-negative match to record $c \in C$. If records from one dataset are allowed to link to multiple records from the other datasets (not uncommon in record linkage), the errors and their effects can quickly build up. The implications of linking multiple data files, which likely will be more common in the big data climate of the day, must be considered and included in the TEF for linked data. This issue is under consideration, as seen in Kim and Chambers (2015).

This total error framework is critical for record linkage in survey sampling. Record linkage as a method continues to grow and has its own set of questions deserving inspection, such as issues of privacy (see Vatsalan et al., 2017) and how record linkage can fit into artificial intelligence programs, but we leave those questions to others since that is not in the scope of this chapter.

To conclude, record linkage is a technique which despite being in existence for 75 years continues to thrive. The ubiquitous nature of non-probability data in our world demands rigorous methods to analyze it. In the overlap between big data, non-probability samples, and statistical sampling lies record linkage. This is an exciting time to research record linkage as it will play an important role in statistical sampling in the future.

References

Abowd, J. M., Abramowitz, J., Levenstein, M. C., Mccue, K., Patki, D., Raghunathan, T., Rodgers, A. M., Shapiro, M. D., & Wasi, N. (2019). Optimal probabilistic record linkage: Best practice for linking employers in survey and administrative data. Center for Economic Studies Working Paper Series Working Paper Number CES-19-08.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Baker, R., J. M. Brick, Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). *Report of the AAPOR task force on non-probability sampling*. American Association for Public Opinion Research. www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Bell, R. M. (2017). *Diverse applications of probabilistic record linkage: Schucany lecture series.* Southern Methodist University.

Bell, R. M., Keesey, J., & Richards, T. (1994). The urge to merge: Linking vital statistics records and Medicaid claims. In *Medical care* (pp. 1004–1018).

Boudreaux, M. H., Call, K. T., Turner, J., Fried, B., & O'Hara, B. (2015). Measurement error in public health insurance reporting in the American community survey: Evidence from record linkage. *Health Services Research, 50*, 1972–1995. https://doi.org/10.1111/1475-6773.12308

Breidt, F. J., Opsomer, J. D., & Huang, C.-M. (2017). Model-assisted survey estimation with imperfectly matched auxiliary data. In: *TES 2018: Predictive econometrics and big data, studies in computational intelligence*.

Briscolini, D., Di Consiglio, L., Liseo, B., Tancredi, A., & Tuoto, T. (2018). New methods for small area estimation with linkage uncertainty. *International Journal of Approximate Reasoning, 94*, 30–42. https://doi.org/10.1016/j.ijar.2017.12.005

Brus, D., & Gruijter, J. D. (2003). A method to combine non-probability sample data with probability sample data in estimating spatial means of environmental variables. *Environmental Monitoring and Assessment, 83*(3), 303–317. https://doi.org/10.1023/A:1022618406507

Chambers, R. (2009). *Regression analysis of probability-linked data*. Official statistics research series (Vol. 4). Statistics New Zealand. oCLC: 908449516.

Chambers, R., & Diniz da Silva, A. (2020). Improved secondary analysis of linked data: A framework and an illustration. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 183*(1), 37–59. https://doi.org/10.1111/rssa.12477

Chipperfield, J. (2020). Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms. *Statistica Neerlandica, 74*(2), 96–111. https://doi.org/10.1111/stan.12189

Chipperfield, J. O., & Chambers, R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics, 31*(3), 397–414. https://doi.org/10.1515/jos-2015-0024

Christen, P. (2008). Automatic training example selection for scalable unsupervised record linkage. In *Advances in knowledge discovery and data mining, 12th Pacific-Asia conference PAKDD* (pp. 511–518).

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review* https://doi.org/10.1162/99608f92.84deb5c4

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 International Conference on Information Integration on the Web* (pp. 73–78).

Copas, J. B., & Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society Series A (Statistics in Society), 153*(3), 287. https://doi.org/10.2307/2982975

Dalzell, N. M., & Reiter, J. P. (2016). Regression modeling and file matching using possibly erroneous matching variables. arXiv preprint arXiv:160806309.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dong, X. L., & Srivastava, D. (2015). *Synthesis lectures on data management:Big data integration*. Morgan and Claypool. https://doi.org/10.2200/S00578ED1V01Y201404DTM040

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nation's Health, 36*(12), 1412–1416.

Elliott, M. N., & Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey methodology, 33*(2), 211–215. http://www.thewitnessbox.com/10498-en.pdf

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice, 2*(6), 1–7. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.981.4054&rep=rep1&type=pdf

Fellegi, I. P. (1999) Record linkage and public policy—a dynamic evolution. In: *Record Linkage Techniques—1997 Proceedings of an International Workshop and Exposition*. National Academies Press, (pp. 1–12).

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.2307/2286061

Groves, R. M., & Lyberg, L. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hallifax, R., Goldacre, R., Landray, M. J., Rahman, N. M., & Goldacre, M. J. (2018). Trends in the incidence and recurrence of inpatient-treated spontaneous pneumothorax. *JAMA, 320*. https://doi.org/10.1001/jama.2018.14299

Harron, K., Goldstein, H., & Dibben, C. (Eds.). (2016). *Methodological developments in data linkage*. Wiley.

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer. oCLC: ocn137313060.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*, 414–420.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*, 491–498.

Jurek, A., Hong, J., Chi, Y., & Liu, W. (2017). A novel ensemble learning approach to unsupervised record linkage. *Information Systems, 71*, 40–54. https://doi.org/10.1016/j.is.2017.06.006

Kim, G., & Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis, 56*(9), 2756–2770. https://doi.org/10.1016/j.csda.2012.02.026

Kim, G., & Chambers, R. (2015). Unbiased regression estimation under correlated linkage errors: Correlated linkage errors. *Stat, 4*(1), 32–45 https://doi.org/10.1002/sta4.76

Kim, J., & Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review, 89*(2), 382–401. https://doi.org/10.1111/insr.12434

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association, 100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Liu, B., Stokes, L., Topping, T., & Stunz, G. (2017). Estimation of a total from a population of unknown size and application to estimating recreational red snapper catch in Texas. *Journal of Survey Statistics and Methodology, 5*(3), 350–371. https://doi.org/10.1093/jssam/smx006

Lohr, S. L. (2010). *Sampling: Design and analysis* 2nd ed.. Brooks/Cole.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics, 12*(2). https://doi.org/10.1214/18-AOAS1161SF

Mulry, M. H., Bean, S. L., Bauder, D. M., Wagner, D., Mule, T., & Petroni, R. J. (2006). Evaluation of estimates of census duplication using administrative records information. *Journal of Official Statistics, 22*(4), 655–679.

Neter, J., Maynes, E. S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association, 60*(312). https://doi.org/10.2307/2283401

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics, 35*(3), 653–681. https://doi.org/10.2478/jos-2019-0027

Salvati, N., Fabrizi, E., Ranalli, M. G., & Chambers, R. L. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83*(1), 78–107. https://doi.org/10.1111/rssb.12401

Stokes, S. L., Williams, B. M., McShane, R. P. A., & Zalsha, S. (2021). The impact of nonsampling errors on estimators of catch from electronic reporting systems. *Journal of Survey Statistics and Methodology, 9*(1), 159–184. https://doi.org/10.1093/jssam/smz042

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Valliant, R., Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research, 40*(1), 105–137. https://doi.org/10.1177/0049124110392533

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017) *Privacy-preserving record linkage for big data: Current approaches and research challenges*. Springer. https://doi.org/10.1007/978-3-319-49340-4_25

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods American Statistical Association* (pp. 354–359).

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology, 8*(1), 120–147. https://doi.org/10.1093/jssam/smz051

Zhang, L., & Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(2), 522–547. https://doi.org/10.1111/rssa.12630

Zhang, L.-C. (2021). Generalised regression estimation given imperfectly matched auxiliary data. *Journal of Official Statistics, 37*(1), 239–255. https://doi.org/10.2478/jos-2021-0010