Journal of Official Statistics
Vol. 4, No. 1, 1988, pp. 3–16
© Statistics Sweden

Record Linkage, Privacy and Statistical Policy

Lawrence H. Cox and Robert F. Boruch[1]

Abstract: Record linkage technology is a valuable statistical tool. It can be used to enhance the quality, completeness and usefulness of data. Record linkage also can be used to check the accuracy of data, demanding fewer resources and imposing less respondent burden than other verification methods. In developing formal record linkage policy or informal policy guidelines, organizations must cope with several competing issues regarding the need for privacy and the benefits and costs of record linkage. This paper discusses recent approaches to using the technology and addressing the problems it generates. Our observations are presented from the perspectives of both data user and data supplier.

Key words: Record linkage; statistical matching; exact matching.

1. Introduction

The objective of this paper is to articulate policy issues for record linkage for research purposes, and to provide suggestions and guidelines for developing and implementing such policies.

In Section 2 we begin by presenting basic concepts and several statistical uses of record linkage. Section 3 describes the principal problems associated with record linkage, including abuses of this technology. This section provides policy makers with guidelines and discusses important aspects of record linkage. This advice is presented from the standpoint of the data supplier. Section 4 discusses policy guidelines that facilitate the availability of data from record linkages for research. This advice is presented from the standpoint of the data user. Concluding comments are presented in Section 5.

U.S. Department of Commerce (1980) provides a general introduction to record linkage. U.S. Department of the Treasury (1985) provides an excellent compendium of current usage and methods of record linkage technology. Policy issues stemming from record linkage are intertwined with the broader issues of access to publicly collected data for research and data sharing. These issues are discussed in Pearson (1986) and Fienberg, Martin, and Straf (1985), respectively.

[1] Lawrence H. Cox is Director, Board on Mathematical Sciences, National Academy of Sciences, Washington, D.C. 20418, U.S.A. Robert F. Boruch is Professor, Department of Psychology and Statistics, Northwestern University, Evanston, IL 60201, U.S.A. Cox's work on this topic has been supported by the U.S. Bureau of the Census where he was Senior Mathematical Statistician. Boruch's work has been supported by the National Science Foundation's Division of Ethics and Values in Science and Technology. The opinions expressed are solely those of the authors and are not intended to represent the policies or practices of the U.S. Bureau of the Census, the National Academy of Sciences or the National Science Foundation. This paper is adapted from Cox and Boruch (1985).

2. Uses of Record Linkage in Statistics

Record linkage is a concept which any two statisticians are likely to define differently. In the broadest sense, a linkage is made between two files when subsets of file A are associated with subsets of file B, and conversely. For example, digitized map files of the same area might be merged to produce a single, more complete file. Before merging, it may be appropriate to link to each point on file B all digitized points on file A within a fixed distance. (Distance here can be simple Euclidean distance; for example, a circle or ellipse that is sufficiently large to contain a predetermined number of points.) Or, one may wish to compare two files drawn from two populations such as working men and working women. Each person is linked to those persons of the opposite sex who best match this person on age, education, etc. In this way the groups can be compared on a variable, such as income, that is common to both files.

Record linkage is useful in statistics. Records may be linked between a post-enumeration survey and a census to evaluate content or coverage. Survey data may be linked to administrative records to benchmark a survey or to edit and correct questionnaire data. Through record linkage, survey data can be augmented by administrative data and respondent burden and survey costs can be reduced. Administrative data can also be used as ancillary information to improve or evaluate estimates. Information collected on the same respondents by different sources can be merged via record linkage. Census files can be used to evaluate proposed survey designs, to create or maintain sampling frames or to stratify samples drawn from these frames. New applications of record linkage methodologies are used to evaluate statistical programs; for instance, when an ongoing survey is coupled with randomized field experiments. These new applications do not exclusively solve problems for policy makers; often new applications are accompanied by new problems.

Record linkage is sometimes referred to as matching. Exact and statistical matching are the two matching techniques used in record linkage. The objective of *exact matching* is to associate each A-record with all B-records that belong to the same owner. In *statistical matching*, an A-record is linked with one or more B-records and the linking is based on characteristics common to both files. The statistically matched B-record(s) need not belong to the owner of the A-record.

Statistical matching may be used as a surrogate for exact matching, especially when exact matching would be too expensive. Statistical matching is usually the method used in applications like imputation. It is also used when, as in our mapping example, the objective is to reduce the resolution of a single map by identifying all points within a small radius of important map features.

Two cautionary notes are necessary. The first is that exact matching is not necessarily easier than statistical matching; it can be more complex. Certainly, if a common owner identification number is available and correctly recorded on both files, then exact matching is easy and inexpensive. This is more likely in a country like Sweden where unique person identification numbers are maintained and in wide use than in countries like the United States or the United Kingdom where such identifiers are lacking. Furthermore, the objective of exact matching is to link records belonging to the same owner, but it is often the case that the matching variables do not correspond exactly between the two files. Names and addresses can be abbreviated or misspelled; the files may not be contemporaneous; or the data may be recorded inaccurately. Consequently, the exactly matching B-record for the target A-record may not be the B-record which most closely resembles the target as it would in a statistical match.

The second cautionary note is that exact and statistical matching are not as similar as they appear. In fact, they differ profoundly in issues like privacy, reliability, and data use. The policy maker must evaluate the effects these differences engender before deciding whether to match and whether the match should be exact or statistical.

2.1. *Statistical applications of exact matching*

Exact matching is an essential tool in census coverage evaluation. For example, file A is drawn from a post-enumeration survey or administrative records and file B is a census file. The objective of an exact match is to estimate population units missed in the census. Exact matching is used here to link the owners of the records from the two files. Statistical matching is not a viable alternative to exact matching for evaluating census coverage.

Another application of exact matching uses a post-enumeration survey designed to evaluate a survey's content rather than its coverage. In this case, both ownership and content of the A-records must be linked to file B, for example, to estimate income reporting bias in the census. Similarly, file B can be used to validate or correct information contained in file A. Statistical matching could be used as a surrogate for exact matching in estimating census income reporting bias. But in this case, only exact matching would suffice for data validation at the individual level.

A third use of exact matching is merging two files. Merging may be done to enrich file A with information contained in file B, for example, to supplement aggregate income data with income components from tax records. Indeed, irrespective of cost savings, exact matching is preferred over direct data collection if the quality of the file B data is considered superior to that obtained from interviewing. For financial data such as tax data, this is often the case. An important special case of file enrichment is the creation of a longitudinal file by exact matching.

In file enrichment, roughly the same set of owners are represented in both files. On the other hand, if the information in the two files is similar but ownership is significantly different, then merging produces a conflated file representing a larger sample. An important special case of merging is when the two files represent substantially the same set of owners and the objective of the linkage is to create an unduplicated file, viz., a file with one record per owner.

File merging and creation of an unduplicated file are essential components of modern automated statistical data processing. The frame for the mail-questionnaires of the U.S. census of agriculture is developed by merging and then eliminating duplication among several lists of businesses broadly classified as "farms." Similarly, the U.S. decennial census is based upon geographic files containing address lists compiled through conflation. Address matching is another example where statistical matching simply will not suffice as a surrogate for exact matching. Like matching person-name, address matching presents difficult problems in methodology and implementation. In the strictest sense, file enrichment requires exact matching, although statistical matching is often used as a surrogate. As described in the next section, the price for this expediency can be as high as a merged file that consistently produces misleading results. Despite this fact, the consequences of using statistical matching in lieu of exact matching often go unexplored.

2.2. *Statistical applications of statistical matching*

Many applications require matching, but often exact matching is neither feasible nor necessary and statistical matching could serve as a surrogate. For other applications, such as our example of male and female wage earners, statistical matching is appropriate on its own merits.

An important application of statistical matching is imputation. Questionnaire data, as collected or transcribed, may contain missing or faulty data which are detected during editing. Records containing missing values or inconsistent data must be provided with suitable values. The process of determining and providing these new values without a reinterview is called imputation (Sande (1982)). In general, imputed values may be obtained either by using statistical matching to locate a donor record or employing a statistical model to compute the imputed values.

In the donor method, the fields to impute in the target are identified during the editing phase. The donor record is then selected from a set of complete and consistent records by a statistical match between the "good" data in the target and this "clean" data set. Conceptually, and often typically, this amounts to associating the target record with an imputation cell, i.e., a subset of the clean data which matches statistically the good data on the target. The final imputed values for the target are obtained (donated) by a suitably chosen record in the imputation cell. Sometimes the match is based upon other similarities, as in hot-deck imputation where a serial correlation is assumed. For example, it may be assumed that people with similar characteristics who live near one another are more likely to have similar incomes than those who live farther apart. In this case, the file is sorted geographically and the donor is the first previously processed record with identical match characteristics.

The second method is model-based imputation, in which a clean data file is used to model relationships between the data items. When a record shows missing data or other inconsistencies, its good data are used in a statistical model that generates the imputed values. Typically, the model is a collection of models each specific to certain patterns of good and faulty data. These models are constructed from subsets of the data file – imputation cells – produced by statistical clustering and matching methods. The model has the effect of associating the target record with this cell by a statistical match based on the valid data on the target. The corresponding statistical model is applied to the valid data to produce the imputed values.

Similar techniques can be used to merge or conflate two files using a statistical match as a surrogate for an exact match. For example, we might create a surrogate longitudinal file from data obtained from an ongoing cross-sectional survey.

The reader may associate these uses of statistical matching with taxonomic or clustering methods. Statistical matching and clustering are related concepts. Matching seeks to link records which are close together; clustering seeks to separate classes of similar records into distinct groups. Both methods require the identification and use of variables known as classification or cohort variables. Matching and clustering techniques used in tandem can stratify the population and form stratum profiles from which nested samples may be selected. Analogously, important units of analysis, such as households, can be formed. Clustering methods are also used in matching; for example, U.S. Department of the Treasury (1985) discusses a method of limiting the search space by blocking the files.

3. Problems Inherent in Record Linkage

3.1. Statistical policy issues

Suppose we want to use record linkage to create an enriched data file, then we have to answer three important questions. (1) How reliable is the linkage expected to be? (2) How does the linked file compare with a similar (perhaps hypothetical) file obtained via direct data collection? (3) What limitations does the linkage process impose on statistical uses of the linked file?

Record linkage is not easy to do. The Office of Population Censuses and Surveys of the United Kingdom reports that for an exact match into its population census file, the first four letters of surname, sex, exact date of birth and place of birth were needed to achieve 98 % accuracy. But the costs were so high that the method is no longer employed (EUROSTAT (1986)).

The reliability of an exact match refers to the accuracy of the matching method and its robustness to small changes in the matching variables. The principal factors affecting match reliability are the choice of matching variables and the quality of the input files. In a pilot linkage using exact matching, match reliability can be assessed beforehand by checking match results on a representative sample of matched record pairs. This is called a pilot linkage. It is less easy to define and assess the reliability of a statistical match, since usually there is no single correct answer. One method is to statistically match sets of records which are known to match exactly and evaluate the results. Another approach is to compare statistics and estimates derived from the linked file with reliable sources. Quantitative measures of match reliability are needed. These measures would, in turn, lead to statistical standards for creating, documenting, and assessing the usefulness of linked files.

Are data on the linked file a valid representation of data obtained via direct data collection? Or does the matching process alter meaningful relationships between the variables? For exact matching, validity depends upon match reliability and factors such as the choice of match variables and the extent to which the data in the two files are comparable. For statistical matching, validity is a more subtle and troublesome concept.

When a statistical match is used to create an enriched data file, a crucial assumption about the data in the two files is being made, i.e., the conditional independence assumption (CI). Let X denote the matching variables common to the two files and let Y and Z denote the remaining variables in these files. By creating a linked file (X, Y, Z) to study Y–Z interactions, one assumes implicitly that Y–Z relationships can be inferred entirely from X–Y relationships and X–Z relationships. It is assumed that, conditional on X, Y and Z are independent. The statistical match, therefore, provides no information about the true conditional relationship between Y and Z. However, inferences from the linked file depend upon the CI assumption, regardless of how well this assumption holds (Kadane (1978)). Rodgers (1984) cites match evaluation studies, Paass and Wauschkuhn (1980), Barr, Stewart, and Turner (1982), and Rodgers and DeVol (1981) all discuss problems related to the CI assumption.

The use of statistical matching for file enrichment raises several issues. Has the conditional independence assumption been examined beforehand? How robust to CI are estimates derived from the matched file? How can additional information such as accurate information on the conditional relationship between Y–Z pairs be used to improve match validity? Is complete information on CI and other assumptions and restrictions on the linkage properly documented for future users of the linked file? Paass (1985b) suggests solutions, particularly the use of additional information, to these types of problems.

The data analyst must carefully account for the limitations of a statistically matched file. Is the statistical match constrained or unconstrained? That is, can each B-record be linked only to one A-record or to more than one? This choice will affect the true degrees of freedom of the matched file. Are all A-records linked to at least one B-record? What are the relationships between the distributions of the X-variables on files A and B and X-variables in the linked file? What biases does the matching introduce? Rodgers (1984) considers these problems in detail.

Statistical matching distorts and dilutes the information of the linked data. As a surrogate for exact matching, statistical matching distorts information because it is an imperfect process and yields less than 100 % match reliability. The user is typically unaware or unable to correct for data imperfections.

It also creates additional distortions by introducing biases attributable to the matching. It dilutes data because it introduces subtle, limiting assumptions such as CI and, in the absence of proper documentation, can mislead the user about the true number of independent observations in the file. In these ways, a linked file can appear to contain more information than it actually does. This is compounded if the linked file is later used to create another linked file. The problem of dilution through record linkage has an analog in survey sampling. Samples get old and must be replenished or redesigned through a benchmark like a census. To counteract the dilution of linked files, it would be beneficial to benchmark linked files and discard or enrich them as needed.

3.2. *Benefit-cost issues*

Few attempts to conduct formal benefit-cost analyses for statistical data linkage have been undertaken. Conclusions from these analyses are then tentative.

We can say, however, that formal benefit-cost analysis cannot and should not be done prior to all linkage efforts. The most interesting of such analyses involve actual pilot linkage and empirical comparisons between the linkage results and those achieved through normal (nonlinkage) procedures. In some administrative settings, this can involve running small randomized experiments that produce estimates of benefits and costs under linkage and conventional conditions. There can be negative effects from requiring a benefit-cost analysis. A "chilling effect" might occur in that research administrators become reluctant to do a linkage because a benefit-cost analysis is required. Benefit-cost analysis of ongoing, routine linkages are often not so informative partly because the results can be ambiguous.

We can also say that the empirical data available to sustain a pre-linkage benefit-cost analysis is generally sparse. Nontheless, empirical data of acceptable quality must be obtained if a benefit-cost analysis is to be done. The most persuasive data is produced by a pilot linkage. A pilot linkage provides empirical estimates of some costs and benefits and may provide information that reduces privacy concerns. However, the benefits of the pilot may be temporal. One cannot obtain information on benefits of repeated linkages, on the decay of the benefit-cost ratio with time, etc. Moreover, the costs of a pilot are said to be close to the cost of a full-scale record linkage, although conclusive evidence is lacking. It is also important to consider the expert opinion of program administrators, audit reports and auditors' judgments, and the results of a manual search and linkage. A proper pilot linkage can be used to estimate (by regression analysis or other means) costs in measurable quantities such as expected number of records or expected number of computations per linkage. Data quality, for example, in expected numbers of false positive and false negative linkages, might be modelled in terms of cost. The pilot linkage should be used to model these relationships whenever possible.

It is also true that the depth and sophistication, i.e., the level of a benefit-cost analysis, can vary. The Department of Health and Human Services uses an illustrative taxonomy of the different levels of benefit-cost analysis, when evaluating potential administrative linkages. According to their taxonomy, level I (simple analysis) compares direct monetary costs with direct monetary savings. Level II (complex analysis), includes all comparisons made at level I, and adds the costs associated with planning, privacy and legal issues, imposition of bureaucracy, and the cost of false positive matches. These costs are then compared with benefits from deterrence, prosecuted cases, monetary savings, and public confidence. At level III (comparative approach), costs associated with matching versus direct interviewing are added to the costs evaluated at level I and II.

How extensive the benefit-cost analysis should be depends on the reason the information is needed and the way it will be used. Comparative analysis is demanding, akin to running a formal experiment. The first two levels seem appropriate when the proposed linkage is not competing with other proposals. The first level is less expensive, but does not take into account qualifying concerns, e.g., privacy and confidentiality.

And finally we can say that pre-linkage benefit-cost analysis differs from post-linkage analysis in purpose and difficulty. Prelinkage analysis can be useful in deciding whether the linkage is warranted. It is a forecast. Post-linkage analysis can be used to determine whether expectations have been met.

3.3. *Privacy and confidentiality issues*

Most file linkages are not planned at the time the original data are being collected, so informed consent for the linkage is not given. Do statements made to the respondent regarding use and confidentiality of the data collected make subsequent record linkage ethical and legal? If so, are such statements sufficient? Gastwirth (1986) raised these issues with regard to data collection by the U.S. Federal statistical system. On one extreme, government could require that all permissible linkages be prespecified and made known to the respondent before data collection; or that the respondent be advised of unforeseen linkages as they arise and have the right to refuse participation. This may be appropriate for administrative and regulatory data, but to do so for statistical data would mean the curtailment of important statistical analyses. Recent findings by the Constitutional Court of the Federal Republic of Germany leading to the development of the Census Law for the Federal Republic's 1987 Census have been construed to mean: "Statistical data collection does not require a close and concrete connection to the purposes data are collected for" (Herberger (1986)). This implies to us that collected census data may be used by the Federal Statistical Office for legitimate statistical purposes without the respondent's pre- or post-approval.

Statistical data are used for different purposes than administrative and regulatory data, and laws governing statistical data collection should respect this distinction. Also, the assiduous and successful manner in which statistical offices have lived up to their confidentiality protection responsibilities over long periods also deserve consideration. Nevertheless, real and current issues remain. For administrative data, there exist large, present privacy issues such as those associated with transborder data flow and credit reporting. However, despite numerous attempts at defining and distinguishing between administrative and statistical data (American Statistical Association (1977)), this distinction is still not clear. The recent postponement of the 1981 census of the Federal Republic of Germany stemmed from confusion over administrative and statistical data and their proper uses. The premiere issue facing the statistical community in record linkage is that of distinguishing statistical data and its use from other data and making that distinction equally clear to the public, to the legislative and legal systems, and to the news media. Privacy laws are being written, laws which potentially can affect statistical programs.

Statisticians have an obligation to explain the case for statistical data to the public and lawmakers. Experience from the West German census crisis could prove valuable here.

The tension between confidentiality protection and record linkage is profound. Exact matching poses a potential threat to the respondent if either file contains confidential data. To protect confidentiality, the linked file should be handled in the same manner as for either of the original (confidential) files, perhaps with even greater intensity because, as confidential and potentially identifying data increases, so does disclosure risk.

The increased disclosure risk and the need to reduce this risk often can have the effect of undermining the purpose of the linked file. The data enrichment created by the linked file may have to be sacrificed to reduce the disclosure risk. In such cases, if tools for assessing disclosure risk are available before linkage or before concerns about release of the linked file are raised, a prudent decision not to release a linked file may be reached. Alternatively, confidentiality protection techniques which distort the data (Cox, Mc Donald, and Nelson (1986); Kim (1986)) could be applied. In either case, something less than an exactly matched file is then available. It is important to have indicators of the usefulness of this file prior to its release and use.

On the other hand, statistical matching does not increase disclosure risk because it does not link record owners. Indeed, in cases where the user has incomplete information on the linkage, the risk of disclosure is reduced by statistical matching. The statistician interested in making data available is faced with a difficult choice. We can create an exactly matched file, and accept the disclosure risks. Or we can simulate an exact match through statistical matching, thus reducing the disclosure risk but accepting what could be questionable data quality. These are areas that need much more research. The interplay between record linkage and confidentiality protection methodologies remains largely unexplored. A notable exception is the work of Paass (1985a,b).

3.4. *Public perception issues*

Perhaps the most widely held false perception regarding record linkage is that of a monolithic Federal bureaucracy that collects and compiles data indiscriminately and shares the data freely among the constituent agencies. The image is that of George Orwell's "Big Brother" state which seeks to control the individual through compiling information about him and shows callousness to his need for privacy. Terminology such as "data banks" tends to reinforce the notion. Discussions of this subject, such as Burnham (1983), emphasize the potential for such government abuse. For the case of statistical data and statistical researchers, investigatory bodies such as the Privacy Protection Study Commission (1977) find no evidence to support this concern. Despite the many opportunities for abuse, there still remains a long-standing record free from abuse. We suggest that this speaks for a successful balance of checks and good will that ultimately deters possible abuses.

Statistical agencies operate under legislative mandates, regulations, and policies separate and different from those of regulatory and enforcement agencies. This is particularly true in countries with decentralized statistical systems such as the United States. In the United States, even when statistical and nonstatistical agencies are parts of one agency or department, as is the case in the Internal Revenue Service, operative strictures, policies, and safeguards are established to insulate data collected for statistical purposes from administrative units. However, the extent to which this insulation varies is unknown.

Within the decentralized U.S. statistical system, there is a separation of rights and responsibilites regarding sharing and linking individually identifiable data. A one-way street model, in which data held by one agency can be shared with another agency but not vice versa, is strictly observed. The U.S. Bureau of the Census occupies the terminal end of this one-way street in that it receives shared data but does not share data with others. This arrangement is complex administratively but is maintained vigorously to guarantee the confidentiality of individual respondent data and the integrity of the statistical system as a whole. This model is followed in other countries, both with centralized and decentralized statistical systems (EUROSTAT 1986)).

Despite the intentions of statisticians towards privacy and their good record of confidentiality protection, the public and, to a lesser extent, the media tend to vastly overrate the ability of the statistician to link files. Accurate linkage on a large "Big Brother" scale may be technically infeasible and certainly would be extremely costly. The U.S. Congress is currently considering regulations that place tighter controls on the linking of government data. These regulations also set standards of reliability for data coming from linked files (Fredell (1986)). These regulations exempt statistical data from such controls, because of its already high security status and the inherent need for flexibility in combining and using statistical data.

To assure that the privileged legal status[2] of statistical data continues, statisticians and statistical policy makers should continue to be scrupulous and cautious regarding record linkage. Extreme care and sound judgment should be exercised when considering linkage. Reviews and checks at various levels, if not currently in place or active, should be made operative.

In the statistical community, we are not aware of any file linkage that was motivated by other than valid statistical reasons. Statistical agencies are well-respected for their considerable and constant disclosure protection. These efforts should be continued and extended to meet the increased disclosure risk associated with individual linked files and, as more linkages are made, sets of overlapping linked files. Moreover, these disclosure protection efforts, not always understood or appreciated by the media and the public, should be publicized in a forth-right and understandable way to reduce false or exaggerated concerns regarding respondent privacy and confidentiality. It is the statistical community's responsibility to inform the public about the importance put on privacy and confidentiality issues.

4. Policy Guidelines on Record Linkage for Research Purposes

The following guidelines for linking information for research are tentative. They are based on advice from a variety of professional groups. The groups include the American Statistical Association, the Evaluation Research Society, government agencies such as the U.S. General Accounting Office and the National Institute of Justice, committees of the U.S. National Academy of Sciences and the Social Science Research Council (Committee on Comparative Evaluation of Longitudinal Surveys (1985)), and especially the international Bellagio Conference.

Four premises were important in the Bellagio Principles (Flaherty et al. (1978)) and we reiterate them here. (1) There are valid and socially significant fields of research for which linked microdata are indispensable. (2) There are legitimate research purposes requiring the use of individual linked records for which

[2] The privileged status is not sacrosanct – the U.S. Office of Management and Budget is currently considering regulations which would limit data release in the U.S. statistical system under the informed consent doctrine to those users (including linkages) made known to the respondent explicitly, presumably at the time of data collection.

public-use samples are inadequate. (3) There are legitimate research purposes requiring the use of linked identifiable data within the framework of concern for confidentiality. (4) The distinction between a research file and other microdata files is fundamental in discussions of privacy and dissemination of microdata.

All of the citations below are taken from Flaherty et al. (1978).

Guideline on Privacy and Confidentiality: "The privacy of individuals and institutions providing information that is shared must be preserved where appropriate and to the extent possible."

There are several justifications for this guideline. First, professional codes of ethics frequently demand that the privacy interests of respondents be recognized. Second, much information is provided under assurances that the information will remain confidential and there is evidence that such assurances are necessary for public cooperation in surveys, see, for example, National Research Council (1978).

Guideline on Procedures for Linkage: "Some research and statistical activities require the linking of individual data for research and statistical purposes. The methods that have been developed to permit record linkage without violating law or social custom regarding privacy should be used whenever possible and appropriate."

The procedures available to accomplish linkage while not undermining privacy concerns can be cumbersome but are effective at times. Some, like insulated linkage systems, are nominally effective but permit deductive disclosures. Others, like statistical matching, may jeopardize the thoughtful analysis of the data but do protect privacy.

Guidelines on Special Linkages: "There is considerable potential for development of more economical and responsive customized user services, such as (1) record linkage under the protection of the statistical office, (2) special tabulations, (3) public-use samples for special purposes. Such services must often involve some form of cost recovery."

The guidelines emphasize the capacity of statistical agencies to implement services economically and responsively. For example, researchers who independently design special purpose studies, especially randomized field tests to complement an ongoing survey, would be linked to the system if (a) the study is compatible with the agency's general work and research responsibilities, (b) the risks of disruption to the system can be managed, (c) the agency is responsible for oversight of the process, and (d) the project engenders no appreciable cost to the agency. These points go well beyond simple notions of record linkage and "data sharing." Special linkages are feasible, however, only for a few projects, perhaps only one every year or two. This is because of the difficulty of adjoining evaluation studies or experiments to an already complex statistical system.

Guideline on Appeals: "In making linked data available to researchers, national statistical offices should provide some means to ensure that decisions on selective access are subject to independent review and appeals."

The Bellagio Conference emphasized that access to government information for research should be handled equitably. Decisions on access to data should not rest solely in the hands of the bureaucracy and formal means of appeal should exist.

Guideline on Access–Policy Research and Evaluation: "To the extent possible, information based on linked records and analyzed for major public policy research and applied work, such as for evaluation and used as the basis for a published report, should be made available for reanalysis."

The rationale here is that any statistical data that affect public policy should be made available for independent inspection, reanalysis, and criticism. In the long run, the data may be combined with other information for better policy research purposes. Precedents include reanalysis of the Equality of Educational Opportunity surveys by Mosteller and Moynihan (1972), of the New Jersey Negative Income Tax Experiments by Boeckmann (1981), and of the Kansas City Police Patrol Experiments by Fienberg, Larntz and Reiss (1976). The Joint Committee on Standards for Educational Evaluation (1981) specified that data should be released for "responsibly planned reviews, sharing being particularly important for facilitating simultaneous analysis by independent analysts." This places the public's interest in open policy research above proprietary interests of the investigator. And other guidelines emphasize that this must be accomplished while recognizing privacy concerns.

Guidelines on Evaluating Rules and Procedures for Record Linkage: "To the extent possible, the usefulness of data linkage strategies should be evaluated periodically." Evaluation here means systematically addressing questions such as: (a) Who needs and uses the linked data? (b) How efficiently are data made available and used? (c) What are the products of use? What is the product's value?

The main justification for these guidelines is that more research is needed on the conditions under which linkage is warranted and worthwhile, and on the effective legal, procedural or other mechanisms for linkage.

Anecdotal and ex post facto cases are available, but more formal systematic studies are sparse. Not all such questions are easily answered, especially ones regarding the "value" of ultimate uses. The state of the art in measuring efficacy and efficiency of data linkage arrangements, and in measuring the utility of the linked information needs to be improved.

Guidelines on Inventory of Linked Files and Benefit-Cost Analyses: "Building knowledge about the propriety, costs and benefits of linkage is essential to good management of government statistical research programs. Each agency should maintain a catalog of linkages and benefit-cost analyses of linkage. Moreover, these ought to be open to public scrutiny."

Studies on the consequences of linkages are an essential source of experience. Reports on benefit-cost analysis of linkage are not always readily accessible in the United States even when the analyses have been produced for administrative purposes. They should be. Whenever possible, such analyses should include models of linkage costs and data quality.

Guideline on Documentation: "The linked information made available for research should be accomodated by documentation that conforms to reasonable standards of quality."

The main justification for this guideline is that linked files are often poorly documented. This results in errors, delays, and difficulty in using linked data, suspicion of data that are not generated in-house, and other problems.

Guidelines on Government Practices, Regulation and Law Pertinent to Record Linkage: "With respect to law, special efforts have to be made to assure that data linkage is (a) not restricted unnecessarily by privacy law and, (b) appropriately fostered by access statutes. The law can also be a vehicle assuring that proprietary interests are balanced against public interests in this arena. Those efforts are the responsibility of the research community and government professionals as well as lawmakers."

These guidelines are an articulation of the concern generated by privacy laws that accidentally cut off linkage useful for research despite the absence of a real threat to individual respondents, e.g., the Tax Reform Act's curtailment of use of addresses for the NAS Medical Follow-up Study (Boruch and Cecil (1979)). The need to make special provisions in privacy laws for access and linkage has been recognized by the Privacy Protection Study Commission (1977). The Commission's recommendations have been incorporated into recent legislative initiatives designed to preserve confidentiality, e.g., Privacy of Research Records Act and Confidentiality of Statistical Records Act.

Guideline on Encouraging Use of Linked Data: Archives, Professional Societies, and Research Sponsors: "The potential consumers of information can and should be vigorously encouraged to use linked data for scholarly research. A variety of ways to encourage use should be exploited by archives, professional societies, and the agencies that sponsor studies."

Merely assuring access to data will not foster use. Information must be marketed and prospective users need to understand how it can be used. Existing infrastructures can be exploited for storage and distribution. These include centralized facilities such as the National Archives, university-based and state archives, and the agencies sponsoring the original studies that can and do store and distribute information. The network of professional organizations, such as the International Federation of Data Organizations, the Association of Public Data Users, and IASSIST should identify repositories, advertise the availability of data, and keep track of the way the information is used. The public supplies the data and deserves to be assured that the data are used repeatedly and well.

Guideline on Public Information: "Considerable efforts should be made to explain to the general public the procedures in force for the protection of the confidentiality of records linked for research and statistical purposes."

The Bellagio Conference concluded that there is considerable public misunderstanding of the special nature of data use involved in research and statistical activities. Fear of the abuse of administrative data by government agencies generates suspicions that carry over to legitimate research. It may be difficult to convince the general public of the long-standing concern for confidentiality shown by statistical agencies. The conference stressed the importance of improving and continuing the information that the public receives on data-access issues.

5. Concluding Comments

Decisions to use record linkage must be carefully thought through, from the standpoint of the agency providing the data, the user, and the public's interests. Administrative guidelines for increased use of record linkage to create files for research purposes have been discussed. Record linkage has been used for many valid statistical purposes. Its use (exact matching) raises legitimate public and professional privacy concerns, as well as legitimate statistical questions on the effects of linkage upon the quality of the data and the ways the data can then be used. In the context of large statistical data files held by government agencies, these questions become serious policy issues affecting the data provider, the data user, and the respondent. Policy guidelines for record linkage must be based upon both statistical evidence and practical experience. The information, principles, and proposals presented in this paper should be considered when forming policy. The broad spectrum of policies on data access for research purposes should also be considered. Among important data-access policy issues are: increased use of restricted-use agreements regarding individual data files, privileged access to identifiable agency data by trusted researchers, and making confidential data available in a secure manner beyond the physical boundaries of agency headquarters.

7. References

American Statistical Association (1977): Report of the ASA Ad Hoc Committee on

Privacy and Confidentiality. The American Statistician, 31, 2, pp. 59–78.

Barr, R.S., Stewart, W.H., and Turner, J.S. (1982): An Empirical Evaluation of Statistical Matching Methodologies. Southern Methodist University, Dallas, unpublished.

Boeckmann, M.E. (1981): Rethinking the Results of a Negative Income Tax Experiment. In Boruch, R.F., Wortman, P.M. and Cordray, D.S. (Eds.): Reanalyzing Program Evaluation. Jossey-Bass, San Francisco, pp. 341–355.

Boruch, R.F. and Cecil, T.S. (1979): Assuring the Confidentiality of Data in Social Research. University of Pennsylvania Press, Philadelphia.

Burnham, D. (1983): The Rise of the Computer State. Random House, New York.

Committee on Comparative Evaluation of Longitudinal Surveys (1985): Report of the Committee. Social Science Research Council, New York.

Cox, L.H. and Boruch, R.F. (1985): Emerging Policy Issues in Record Linkage and Privacy. International Statistical Institute, Proceedings of the 45th Session, Amsterdam, 1, pp. 17–32.

Cox, L.H., McDonald, S., and Nelson, D. (1986): Confidentiality Issues at the United States Bureau of the Census. Journal of Official Statistics, 2, 2, pp. 135–160.

EUROSTAT (1986): Protection of Privacy, Automatic Data Processing and Progress in Statistical Documentation. Eurostat News, Luxembourg.

Fienberg, S.E., Larntz, K., and Reiss, A.J. (1976): Redesigning the Kansas City Patrol Experiment. Evaluation, 3, pp. 124–131.

Fienberg, S.E., Martin, M.E., and Straf, M.L. (Eds.) (1985): Sharing Research Data. National Research Council, Washington, D.C.

Flaherty, D.L. et al. (1978): The Bellagio Conference on Privacy, Confidentiality and the Use of Government Microdata. New Directions in Program Evaluation, 4, pp. 19–30.

Fredell, E. (1986): Strengthening ADP Match Regs Seen. Government Computer News, Washington, D.C., 5, 19.

Gastwirth, J.L. (1986): Comment. Journal of the American Statistical Association, 81, 393, pp. 23–25.

Herberger, L. (1986): Planning of West Germany's Next Population and Housing Census. Unpublished.

Joint Committee on Standards for Educational Evaluation (1981): Standards for Evaluation of Education Programs, Projects, and Materials. McGraw Hill, New York.

Kadane, J.B. (1978): Some Statistical Problems in Merging Data Files. 1978 Compendium of Tax Research, Office of Tax Analysis, Washington, D.C., pp. 159–171.

Kim, J.J. (1986): A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. American Statistical Association, Proceedings of the Survey Research Methods Section, to appear.

Mosteller, F. and Moynihan, D.P. (Eds.) (1972): On Equality of Educational Opportunity. Vintage, New York.

National Research Council (1978): Privacy and Confidentiality as Factors in Survey Response. National Academy of Sciences, Washington, D.C.

Paass, G. (1985a): Disclosure Risk and Disclosure Avoidance for Microdata. IASSIST Conference on Public Access to Public Data, Amsterdam.

Paass, G. (1985b): Statistical Record Linkage Methodology: State of the Art and Future Prospects. International Statistical Institute, Proceedings of the 45th Session, Amsterdam, 1, pp. 33–48.

Paass, G. and Wauschkuhn, U. (1980): Experimentelle Erprobung und Vergleichende Bewertung Statistischer Matchverfahren.

Gesellschaft für Mathematik und Datenverarbeitung, Bonn.

Pearson, R.W. (1986): Research Access to Publicly Collected Data. Social Science Research Council, New York.

Privacy Protection Study Commission (1977): Personal Privacy in an Information Society. U.S. Government Printing Office, Washington, D.C.

Rodgers, W.L. (1984): An Evaluation of Statistical Matching. Journal of Business & Economic Statistics, 2, 1, pp. 91–102.

Rodgers, W.L. and DeVol, E. (1981): An Evaluation of Statistical Matching. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 128–132.

Sande, I.G. (1982): Imputation in Surveys: Coping with Reality. The American Statistician, 36, 3, pp. 145–152.

U.S. Department of Commerce (1980): Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper 5. Government Printing Office, Washington, D.C.

U.S. Department of the Treasury (1985): Record Linkage Techniques – 1985. Internal Revenue Service, Statistics of Income Division, Washington, D.C.

Received January 1987
Revised February 1988