

The Use of Names for Linking Personal Records

Howard B. Newcombe, Consultant
Martha E. Fair and Pierre Lalonde, Statistics Canada

The skill of a human who searches large files of personal records depends much on prior knowledge of how the names vary in successive documents pertaining to the same individuals (e.g., as with ANTHONY-TONY, JOSEPH-JOE, WILLIAM-BILL). Now, an essentially exact procedure enables computers to make similar use of an accumulated memory of their own past experiences when searching for, and linking, records that relate to particular persons. This knowledge is further applied to quantify the benefits from various refinements of the rules by which the discriminating powers of names are calculated when they do not precisely agree or are substantially dissimilar. Of the six refinements tested, by far the most important is the recently developed exact approach for calculating the ODDS associated with comparisons of names that are possible synonyms.

KEY WORDS: Data base maintenance; File searching; Probabilistic linkage; Quantitative judgment; Record linkage.

Personal documentation in machine-readable form has become so extensive in any advanced society as to constitute, collectively, a detailed but highly fragmented life history for virtually all its members. The files exist to serve the needs of people and of society as a whole, and frequent access is involved. Much of the searching is necessarily based on names and personal particulars that are apt to be reported differently on successive documents for the same individuals. The problems are familiar to clerks, but now access by computer is becoming the norm.

With automated searching, many choices are possible between *refinements* and *simplifications* in the way that names get compared. Rarely, however, have the merits of alternative approaches been quantified in terms of gains or losses of discriminating power, so as to reduce the guesswork when designing a system. The potential for sophistication in automated comparisons of names is substantial. Humans develop special skills in recognizing nicknames, ethnic variants, diminutives, and corrupted forms due to truncations, misspellings, and typographical errors. This is known to be based on a relatively simple rationale, supported by remembered data. If a machine is to acquire similar ability, it too must rely on past experience (Newcombe, Fair, and Lalonde 1989; Newcombe, Kennedy, Axford, and James 1959). Although there is now an essentially exact way of measuring the discriminating powers of comparison pairs like CARL-KARL, GEORGE-GYORGY, JACOB-JAKE, JOHN-JACK, and WILLIAM-BILL, much clerical labor and large amounts of data are needed to set it up (Fair, Lalonde, and Newcombe 1990, 1991; Newcombe et al. 1989). Simpler comparisons are, therefore, likely to remain popular in many procedures that use names to access files.

Whether or not this exact approach becomes widely applied, its existence now provides a convenient standard against which to judge the performance of other treatments of names. So we have used the approach in this article to quantify the gains and losses of discriminating power due to various *refinements* and *shortcuts* commonly used in automated searching and linkage.

* Howard B. Newcombe is a consultant, P.O. Box 135, Deep River, Ontario K0J 1P0, Canada. Martha E. Fair is Chief and Pierre Lalonde is Project Manager, Occupational and Environmental Health Research Section, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada. The authors thank John Armstrong, Michael Eagen, and William E. Winkler for helpful critical comments on an early version of this article, and also the associate editors and referees who substantially influenced its final form.

The test is special to names as identifiers; is suitable for fine-tuning this component of a record linkage system; and is uninfluenced by the adequacy of the rest of the identifiers. It differs from, but is complementary to, more direct tests of overall performance.

1. COMPUTER LINKAGE

Where a computer is used to search large files of personal records and bring together the records for particular individuals, it may emulate with varying degrees of success the strategies of a human clerk who does the same job. To determine whether a pair of records is correctly matched, the names are compared along with other identifiers (e.g., year, month, and day of birth; sex and marital status; and various geographic particulars such as place of birth, residence, work, or death). Sometimes, however, these comparisons point in different directions.

The problem then is to determine, as in a court of law, where the preponderance of the evidence lies. The comparisons must be considered not only separately but also in combination. A particular comparison outcome (e.g., JOHN-JOHN or JOHN-JACK) will argue *for* linkage when it is more common among correctly matched pairs than among random false matches. Conversely (as with JOHN-JOE), an outcome will argue *against* linkage when the opposite is the case. These *likelihood ratios* (or individual ODDS in favor of linkage) may be combined to assess the collective evidence from the full set. But this is not the whole of the relevant information.

In addition, a human clerk may recognize two further factors: the *size* of the file being searched and the *likelihood* that the individual is represented in it. Thus, when looking for a particular JOHN BROWN in the telephone directory for a small town where he is thought to reside, finding the name suggests that it may well belong to the right person. This would definitely *not* be so when searching a large national death register, especially if this JOHN BROWN were unlikely to have died.

Automated searches have from the outset used much the same reasoning as does a human clerk; this provides numerous options when calculating the ODDS for particular

theory alone (Newcombe and Kennedy 1962; Newcombe et al. 1987). (The need for practice and theory to complement each other is discussed elsewhere; see Scheuren, Alvey, and Kilss 1986; Winkler 1989b.)

When Statistics Canada first actually used probabilistic linkage in the early 1980s, based on the Fellegi–Sunter theory, it was to search the newly established Canadian mortality data base, which extended back to 1950. Their linkage system, known as CANLINK or GIRLS (for Generalized Iterative Record Linkage System) included innovations described by Howe and Lindsay (1981), Hill (1981), Hill and Pring-Mill (1985). In particular, a *preliminary linkage* step was introduced that temporarily ignored specific values of names, thereby eliminating in a simple fashion many unpromising record pairs, and an *iterative update* of the outcome frequencies from LINKED pairs of records was used. The preliminary step was needed because the death files were now blocked by just a single surname (as a NYSIS phonetic code; see Appendix H of Newcombe 1988), and the blocks were larger than those based on pairs of surnames for family linkage. The iterative updates were required because to get new linkage jobs started, outcome frequencies from earlier linkages were often used initially and replaced later with increasingly appropriate data as the new files of LINKS were progressively improved. (The effect of omitting this update is considered in Sec. 3.4.) A further intended refinement, recognition of *partial agreements* of names (like THOMAS–TOM), was less successful; as a result, modified procedures had to be devised (Eagen and Hill 1987; Fair et al. 1990, 1991; Newcombe 1988; Newcombe et al. 1987, 1989; Winkler 1985, 1989a.) The matter is referred to again in Section 3.5.

Howe and Lindsay (1981) also recognized explicitly, for the first time, the concept of the *prior odds* or *prior likelihood* but failed to apply it to create a scale of *absolute* ODDS that might be used for setting thresholds. Earlier, two thresholds had been proposed as part of the Fellegi–Sunter theory to distinguish *positive links* and *positive nonlinks*, plus an intermediate category of ambiguous matches called *possible links*. The thresholds were to be calculated in advance as “error bounds” that would limit the numbers of false-positive and false-negative links and would identify pairs in need of special assessment. But when the ODDS from the full sets of identifiers were combined, it was found that the resulting overall ODDS served only to array the record pairs, *relative to one another*, in descending order of the likelihood of a correct match. Thus, in practice, the two thresholds got assigned subjectively. On the scale of *relative* ODDS available at the time, they fell high above the crossover or 50/50 odds point (e.g., in the case of the death searches by a factor of well over 1 million, and greater than the size of the file being searched).

An empirical conversion to a scale of presumed *absolute* ODDS indicated why. When allowance was made for the size of the death file, $1/N(\text{File B})$, and for the proportion of search records that find a matching death record in it, $N(A | \text{LINK})/N(\text{File A})$, the new scale brought the subjective thresholds close to the crossover or 50/50 odds point. Together, these two factors were taken to represent the prior

likelihood of a correct match on a single random pairing (i.e., before examining any identifier or blocking information).

The new scale of absolute ODDS was controversial at first, although the results were consistently believable over many empirical tests, whereas those from the alternative were not. Later, it was shown to use just a variant of the prior odds, $P(\text{LINK})/P(\text{NONLINK})$, already recognized by Howe and Lindsay (1981). The implications are substantial but were not explored by those authors (see Secs. 2.3 and 3.1 and Fig. 1). In practice, however, it was soon found that the concept of the prior likelihood could be applied with great flexibility in many ways. For example, as a refinement it was calculated separately for subsets with differing prior likelihoods (see Newcombe 1988, chap. 28 and apps. B and D.3).

What refining the practice achieved, as distinct from formal theory, was enhanced flexibility in the access to discriminating power. *Individual identifiers* were compared freely, just as a human might do when seeking clues to the true linkage status of a record pair; and the prior likelihood of a correct match, in the case of a death search, was exploited to take into account the age of the individual in a given year, and the actuarial likelihood that he or she might have died in that year. For linkages of cancer records with death files, the approach even used survival curves appropriate to particular diagnoses. The practices are fully described, but in nontechnical language for those working close to the files, who design, implement, and test the detailed procedures (see, for example, Newcombe 1988, sec. 28.2 and apps. D.2 and D.3).

This is the technological setting within which the current study has been carried out.

1.2 General Method

Any formal statement of the comparison procedure for individual identifiers should allow for the flexibility that exists in practice. This is especially true of names when they do not precisely agree (e.g., as allowing recognition of the comparison DANIEL–DANNY). Moreover, because some kind of grouping of possible synonyms is inevitable, this too must be exceedingly flexible if discriminating power is not to be wasted (Scheuren 1985). We will deal first with formal expressions that permit flexibility when estimating likelihood ratios (or ODDS in favor of linkage as indicated by particular comparisons), and second with grouping under conditions of minimum constraints. (Other accounts use logarithms of the likelihood ratios and refer to them as “weights.” The ratios may also be viewed as factors by which comparisons of particular identifiers raise or lower the overall “betting odds” in favor of linkage.)

Conceptually, each first given name on one file is compared with every first given name on the other file, and second given names are likewise compared. Generally, LINKED pairs (of names or records) are vastly outnumbered by *possible* NONLINKED pairs, i.e., actual plus potential. (This concept is fundamental and is not altered by “blocking” that reduces the *actual* numbers of comparison pairs; see Fellegi 1985.) Although LINKS and NONLINKS are thought of as uncon-

taminated with pairs of the opposite kind, modest admixtures have only slight effects on the ODDS.

When comparing value A_x from a Record A (which is used to initiate a search) with value B_y from a Record B (which is in the file being searched), the ODDS in favor of a correct LINK associated with outcome $A_x \cdot B_y$ (i.e., the comparison pair of values) may be written in terms of the relative probability of occurrence of the particular outcome in LINKS as compared with NONLINKS; that is,

$$\text{ODDS} = P(A_x \cdot B_y \mid \text{LINK}) / P(A_x \cdot B_y \mid \text{NONLINK}). \quad (1.1)$$

But except where files A and B are both very small, the denominator in this expression will be closely approximated by $P(A_x) \cdot P(B_y)$, because any fortuitous LINKS in the random pairs will be vastly outnumbered by the NONLINKS. Thus the expression may be converted to

$$\text{ODDS} = P(A_x \cdot B_y \mid \text{LINK}) / P(A_x) \cdot P(B_y). \quad (1.2)$$

This implies that we need to know in advance the number of LINKS with values A_x and B_y . In practice crude approximations are estimated initially from sample linkages carried out manually or from previous linkage studies and are revised iteratively as the current LINKS are progressively refined.

An expanded form of this procedure is sometimes used to support an existing practice in the case of death searches. This involves ignoring the frequency of value A_x , both in File A and in the LINKS, on the grounds that names are unlikely to be strongly correlated with the probability of death and with whether a Record A is LINKED to a Record B. Justification depends on the magnitude of the error introduced by the assumption. The expanded version has two parts:

$$\text{ODDS} = \frac{P(B_y \mid A_x \cdot \text{LINK})}{P(B_y)} \cdot \frac{P(A_x \mid \text{LINK})}{P(A_x)}. \quad (1.3)$$

SIMPLIFIED FORMULA CORRECTION FACTOR

Current practice views the second part (the "correction factor") as approximating unity, so it can be ignored, except where the assumption is thought to be seriously misleading (as it might be if ethnicity and ethnic names were correlated with mortality).

What the relative probabilities fail to do is indicate explicitly how the ODDS should be calculated using data that are in short supply. Examples include outcome values $A_x \cdot B_y$ that are represented only once or twice in an available real file of LINKS and, especially, numerous other outcome values representing pairs of possible synonyms that have not actually occurred in the available LINKS but probably would occur if that file were larger. Because crucial steps in the reasoning have to do with numbers of outcome values, as distinct from their likelihoods, it is helpful to convert the last two expressions to a form actually used to obtain estimated relative probabilities, as

$$\text{ODDS} = \frac{N(A_x \cdot B_y \mid \text{LINK}) / N(\text{LINKS})}{N(A_x \cdot B_y \mid \text{NONLINK}) / N(\text{NONLINKS})} \quad (1.4)$$

and

$$\text{ODDS} = \frac{N(A_x \cdot B_y \mid \text{LINK}) / N(A_x \mid \text{LINK})}{N(B_y) / N(B)}$$

SIMPLIFIED FORMULA

$$\times \frac{N(A_x \mid \text{LINK}) / N(\text{LINKS})}{N(A_x) / N(A)}, \quad (1.5)$$

CORRECTION FACTOR

where the general term $N(* \mid \text{LINK})$ represents the number of records among LINKED pairs that have attribute (*), $N(\text{LINKS})$ = number of linked pairs, $N(A)$ = number of records in File A, $N(B)$ = number of records in File B, $N(A_x)$ = number of records in File A with value x , and $N(B_y)$ = number of records in File B with value y . (For the origins of this version, see Newcombe et al. 1989.)

It is convenient to retain the distinction between a search file (File A) and a file being searched (File B), even though conceptually the roles could be reversed. For one thing, the search file usually is smaller than the file being searched. Also, the distinction has special significance for the death searches, because informal versions of a given name (e.g., nicknames) are more commonly used by employers and others while one is alive rather than by undertakers after one has died.

Here we need to introduce two concepts related to the ways in which the range of possible outcomes may be handled:

1. Grouping or "pooling" of similar values of $A_x \cdot B_y$, which individually are represented poorly or not at all in the available LINKS (the "quantity" problem)
2. Increasing sacrifice of discrimination as the within-group heterogeneity grows when its definition is broadened to ensure representation in the LINKS (the "quality" problem).

A tradeoff between "quantity" and "quality" is unavoidable. The definition of an outcome group needs to be broad enough so that $N(A_x \cdot B_y \mid \text{LINK})$ is represented by at least one comparison pair. Otherwise, no ODDS can be calculated. But because the definition is widened to increase the representation, it will also let more heterogeneity into the group. (Thus as the error due to statistical fluctuation diminishes, so the error due to lessened specificity increases.)

The earliest linkage operations simplified matters by recognizing just two categories of outcome—*agreements* and *disagreements*—and by attributing specificity for value only to the former category. But major errors arose from an unsuccessful attempt to adapt the earlier procedures, to recognize "partial agreements" such as JOSEPH-JOE (Newcombe et al. 1987). (The term "partial agreement" is commonly applied, for reasons of convenience, to any possible synonyms regardless of similarity, as with ELIZABETH-BETTY.)

The problem posed by the value-specific partial agreements of names may be handled in various ways, but only one of

these appears to be precise. A compromise solution, now in routine use, is based on the numbers of early characters that agree. ODDS are first calculated for different levels of agreement (i.e., one, two, three, four or more agree); actual values are ignored at this stage. Such "global ODDS" are later adjusted upward or downward, depending on whether the particular values of the *agreement portions* are rare or common (Eagen and Hill 1987; Newcombe 1988; Newcombe et al. 1987), but this neglects the values of the *disagreement portions* (e.g., it wrongly treats diverse name pairs like JOHN-JONATHAN and JOHN-JOSEPH as equally likely to be synonyms). An alternative approach that recognizes phonetic components common to the two names has also been developed (Winkler 1985, 1989a).

A precise treatment of partial agreements of names recognizes both values in a comparison pair and avoids resorting to globally defined (i.e., value-nonspecific) levels of agreement. This permits it to deal with outwardly dissimilar comparison pairs (e.g., EDWARD-TED, MARGARET-PEGGY). Any necessary groupings must be defined in value-specific ways. The frequency with which the two values are related by actual usage then determines the magnitude of the precise ODDS. A modest manual test showed that the approach worked where sufficient data from LINKED pairs of records could be made available (Newcombe et al. 1989). That was followed by an expanded application based on an accumulated composite file of LINKS from many past searches of the Canadian mortality data base (Fair et al. 1990, 1991). This refinement will be considered further in Section 3.5.

(The current emphasis on flexibility also extends to other identifiers that are apt to be reported differently on separate occasions or that may change over time, as with MARITAL STATUS, OCCUPATION, INDUSTRY, and PLACES OF RESIDENCE, WORK, and DEATH. For these, there likewise is no need to prejudge in which direction the comparisons will argue. "Agreement" and "disagreement" are often poor indicators, but the ODDS—when they have been calculated—will decide.)

1.3 Combining the ODDS

When the likelihood ratios or ODDS for particular identifiers are combined over the full set in a record pair, it is usual to assume as a tolerable approximation that the identifiers are independent of one another. The overall *absolute ODDS* (in the sense of "betting odds" in favor of linkage) may then be represented by

$$\text{Absolute ODDS} = R_1 \cdot R_2 \cdot \dots \cdot R_n \cdot P(\text{LINK}), \quad (1.6)$$

where R_1 to R_n are the likelihood ratios (ODDS) for identifiers 1 to n (including any used for blocking) and are independent of each other, and $P(\text{LINK})$ is the prior likelihood of a correct match on a singly random pairing. The latter term is similar to the *prior odds*, $P(\text{LINK})/P(\text{NONLINK})$, recognized but not used by Howe and Lindsay (1981). Confusion remains concerning the implications, and is not explicitly addressed by existing formal theory (see Sec. 2.1).

The version of this expression used to calculate *estimated*

absolute ODDS from actual counts is unfamiliar to many, so it is necessary to be explicit: R_1 to R_n become frequency ratios, and $P(\text{LINK})$ becomes $N(\text{LINKS})/N(\text{LINKS} + \text{NONLINKS})$. Because each linked pair contains one record from File A and one from File B, $N(\text{LINKS}) = N(A|\text{LINK}) = N(B|\text{LINK})$. Also, where each record on File A is compared in succession with every record on File B, the total number of comparison pairs, regardless of their linkage status, will together equal the product of the two file sizes; that is, $N(\text{LINKS} + \text{NONLINKS}) = N(\text{File A}) \cdot N(\text{File B})$. The concept is valid even where, in practice, only the pairings that occur within blocks are actually seen; but this implies that likelihood ratios for blocking identifiers will be taken into account. Thus by substitution we may obtain

Absolute ODDS

$$= R_1 \cdot R_2 \cdot \dots \cdot R_n \cdot \frac{N(A|\text{LINK})}{N(\text{File A})} \cdot \frac{1}{N(\text{File B})}. \quad (1.7)$$

Howe and Lindsay (1981) had felt that their prior odds, $P(\text{LINK})/P(\text{NONLINK})$, could not be readily estimated. The solution came to us by observing *human stratagems* and through reasoning based on *counts* rather than on *probabilities*. At first, it was hard to persuade others that this practice is valid, perhaps because our way of thinking was unconventional (David Binder and Geoffrey Howe, personal communication, November 10 to December 11, 1982). A further possible reason might be the common custom of *not* calculating frequency ratios for blocking identifiers; but then NA and NB would represent the sizes of Files A and B *within* the particular block, and the prior likelihoods would differ from block to block.

Calculation (1.7) has been used over the past decade for searches of Canadian death files. The application is exceedingly flexible and allows refinement through redefinition of Files A and B to represent, separately, a multiplicity of subsets (based on age, death year, selected diagnoses, and so on) of populations that are internally heterogeneous. (For details, see Newcombe 1988 chap. 28 and apps. B and D.2.)

2. EMPIRICAL DISTRIBUTIONS OF LINKS AND NONLINKS

A feedback of empirical data from the LINKS and NONLINKS is the most basic requirement of a linkage system. For example, the expressions by which the ODDS for the individual identifiers are calculated require these data as input. Also, such data are needed when assessing errors due to assumptions that are not strictly correct.

Above all, direct observation of individual record pairs often yields clues to more suitable comparison steps. These clues are most likely to become apparent to humans when resolving difficult matches manually. An experienced person can be less bound by artificial constraints than the automated system, and he or she is still, given existing linkage systems, in a better position to be guided by memories of past encounters with similar problems.

Theoretical papers on linkage make strong assumptions to get results, and linkage practice does the same to simplify

procedures. Examples include the use of artificially simplified ways of comparing names, which may not adequately exploit their true discriminating power, and the practice of simply multiplying the ODDS for individual identifiers to combine them for a whole set, which would be strictly proper only if they were independent of each other (Fellegi and Sunter 1969; Howe and Lindsay 1981).

Only with better data from LINKS and NONLINKS can many of the uncertainties be resolved. Recognition of this has led, in part, to the idea of accumulating large files of LINKS and creating even larger files of NONLINKS (see, for example, Fair et al. 1990, 1991; Lalonde 1989; Newcombe et al. 1989). It has also emphasized the use of additional evidence on the *true linkage status* of record pairs assigned borderline absolute ODDS in an automated operation (Fair, Newcombe, and Lalonde 1988a; Fair, Newcombe, Lalonde, and Poliquin 1988b).

We will deal first with the latter point.

2.1 The Assumption of Independence

Calculated overall "absolute ODDS" usually assume that the components in the identifier sets are independent of each other. Rarely is this assumption strictly correct. It can be seen to be misleading when scanning visually for record pairs

that were wrongly classed as positive LINKS and positive NONLINKS. Our unpublished observations include examples of *multiple agreements* (e.g., of rare ethnic names and related places of birth) that have spuriously raised the ODDS to create false positives. Conversely, there are examples of *multiple disagreements* (especially on year, month, and day of birth—perhaps due to multiple wrong guesses by an informant at the time of a death), which have spuriously lowered the ODDS to create false negatives.

The effects of these and other such biases are best visualized in the overlap between the numbers of verified LINKS and NONLINKS, when distributed along a scale of absolute ODDS that assumes independence, as in Figure 1 (data of Fair et al. 1988a, 1988b; and Lalonde 1986). We will refer to points on this scale as "theoretical" ODDS to distinguish them from the "empirical" ODDS, which are the ratios of observed counts of LINKS/NONLINKS at various points on the same scale. (Total LINKS and NONLINKS are not shown in the Figure; but conceptually the latter vastly outnumber the former.)

In practice there is no need to actually create the bulk of the possible NONLINKS, because most would fall so very low on the scale. Major misunderstanding arises, however, when the enormous preponderance of actual plus potential NONLINKS over LINKS is not kept in mind. Thus the distributions and their crossover points serve little purpose if

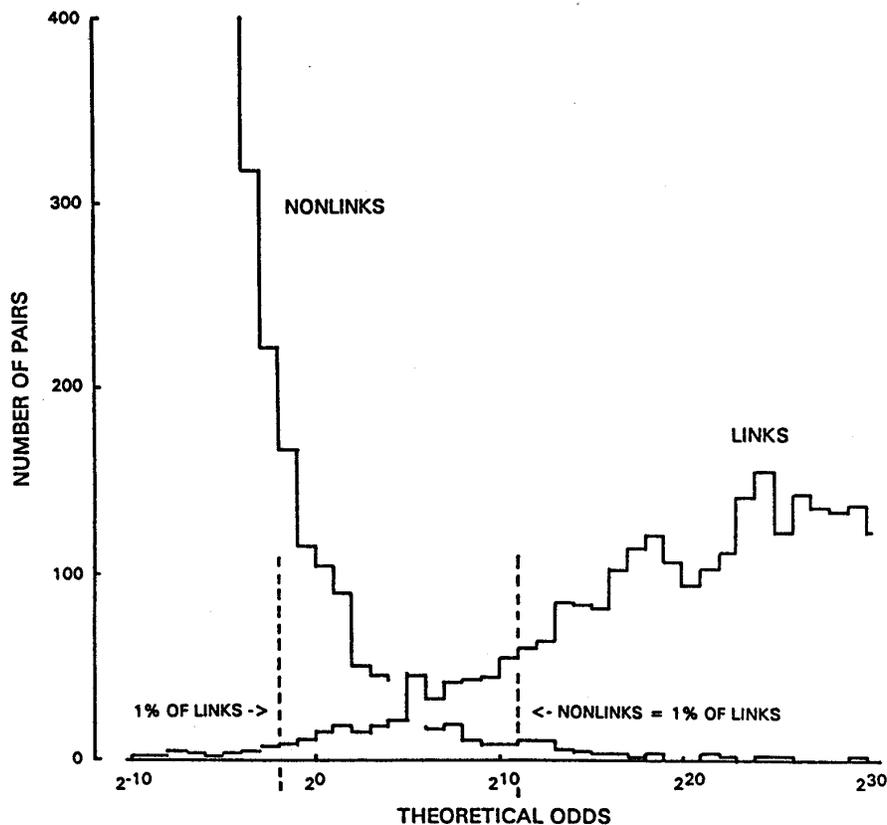


Figure 1. Overlapping Parts of the Distributions of LINKS and NONLINKS, on a Scale of Theoretical ODDS (Lalonde 1986). Note that empirical error bounds (broken lines), set at the 1% levels, are displaced upward on the theoretical scale.

plotted as proportions of LINKS compared with proportions of NONLINKS. Likewise, upper and lower "error bounds," when expressed in such terms, make nonsense of the concept. (The data in Figure 1 are from searches of 1,300,000 death records, initiated by 30,000 work records, yielding 2254 LINKS; the vital status of doubtful pairs was confirmed using taxation files. Because the number of possible pairings, i.e., actual plus potential, is the product of the two file sizes, NONLINKS outnumber LINKS by 17,000,000 to 1.)

Marked discrepancies are revealed in Figure 1 between the theoretical ODDS scale, based on the assumption of independence, and the corresponding observed ratios of LINKS versus NONLINKS. For example, where the theory indicates that the ODDS in favor of linkage are 1/1, in reality they are only 1/6; and where the observed ODDS are 1/1, the theory says that they should be 16/1. Moreover, if one wants to set lower and upper thresholds to limit the number of LINKS wrongly classed as "positive nonlinks" to 1% of all LINKS, and to likewise limit the NONLINKS wrongly classed as "positive links" to a similar number (i.e., 1% of the LINKS), the correct thresholds would be represented by theoretical ODDS of approximately 1/4 and 2,000/1. Thus the true error bounds are displaced upward on a scale of ODDS that assumes independence.

There has been confusion in the past, which is best avoided by thinking in terms of numbers (i.e., *counts*) as distinct from proportions. One does *not* limit false positives to 1% of NONLINKS because, in our example, that would create 17,000,000 times as many false positives as false negatives. Indeed, the Fellegi-Sunter theory emphasizes that NONLINKS typically will greatly outnumber LINKS; for example, see slides #9 and #10 of Fellegi 1985. More explicitly, where this is the case "no one could possibly conclude" that the two error bounds would be properly set at equal proportions (i.e., 1%) of the LINKS and of the NONLINKS (I. P. Fellegi, personal communication July 8, 1987).

2.2 Data on Name Comparisons Involving Synonyms

Value-specific information to do with $N(A_x \cdot B_y | \text{LINK})$, heretofore lacking in quantity, is contained in a composite file of 64,937 LINKED pairs of male given names derived from 26 linkage projects. All of the projects involved searches of the Canadian Mortality Data Base (File B, containing 3,397,860 male given names), initiated by records of various study cohorts, including employment records, survey responses, cancer registrations, birth records, and entries in a national radiation dose register (composite File A, containing

Table 1. Common Male Given Names From the Canadian Death File, 1950-1977

Rank	Name*	Total observed		Rank	Name*	Total observed	
		Number	Percent			Number	Percent
<i>Formal Names</i>							
1	JOHN	187,486	5.30	26	JEAN (male)	22,661	.64
2	WILLIAM	170,669	4.83	27	FRANCIS	21,596	.61
3	JAMES	111,513	3.16	28	HAROLD	21,588	.61
4	JOSEPH	104,767	2.96	29	GORDON	19,158	.54
5	GEORGE	95,188	2.69	30	HERBERT	19,133	.54
6	CHARLES	70,040	1.98	31	SAMUEL	18,927	.54
7	ROBERT	66,575	1.88	32	ANDREW	18,440	.52
8	THOMAS	64,182	1.82	33	DONALD	17,416	.49
9	HENRY	55,718	1.61	34	DANIEL	16,076	.46
10	EDWARD	55,837	1.58	35	STANLEY	14,575	.41
11	ARTHUR	52,221	1.48	36	PATRICK	13,402	.38
12	ALBERT	47,660	1.35	37	NORMAN	13,270	.38
13	ALEXAND (ER)	38,343	1.09	38	ROY	12,943	.37
14	FREDERI (CK)	36,864	1.04	39	RAYMOND	12,338	.35
15	DAVID	33,530	.95	40	EMILE	12,261	.35
16	ERNEST	32,041	.91	41	HENRI	12,107	.34
17	ALFRED	30,902	.87	42	KENNETH	12,076	.34
18	FRANK	29,376	.83	43	DOUGLAS	11,843	.34
19	PAUL	26,919	.76	44	LEONARD	10,978	.31
20	PETER	26,889	.76	45	EUGENE	10,968	.31
21	WALTER	26,718	.76	46	VICTOR	10,797	.31
22	HARRY	24,830	.70	47	GEORGES	10,446	.30
23	MICHAEL	24,645	.70	48	ALLAN	10,384	.29
24	RICHARD	24,070	.68	49	LEO	10,200	.30
25	LOUIS	23,860	.68	50	EDWIN	10,156	.29
				51	CLARENC (E)	9,974	.28
<i>Informal Variants</i>							
1	FRED	7,947	.23	8	JOE	866	.025
2	JACK	5,575	.16	9	DAN	781	.023
3	ALEX	3,550	.10	10	BILL	314	.009
4	MIKE	3,267	.10	11	PETE	265	.008
5	SAM	2,014	.06	12	DON	240	.007
6	RAY	1,911	.056	13	ANDY	220	.006
7	TOM	990	.029	14	DAVE	179	.005
				15	ED	43	.001

* Truncated at seven characters in the records of the Canadian mortality data base.

Table 2. Pooling of Synonyms in Value-Specific Groups: Example Based on CHARLES Compared with KARL and Related Variants

Value of name	Numbers in File B*	Value of name	Numbers in File B*	Value of name	Numbers in File B*
KARL	3,002	KARLIOU	1	KARLS	2
KARLA	1	KARLIS	82	KARLSEN	2
KARLDON	1	KARLMER	1	KARLSON	2
KARLE	6	KARLO	36	KARLSSO	1
KARLEY	1	KARLOFF	1	KARLTON	1
KARLHEI	2	KARLOL	1	KARLY	2
KARLIE	1	KARLOS	1		

* Based on an alphabetic listing from the death file. Of these names, only KARL was actually interchanged with CHARLES in the linked pairs of records. However, the other potential combinations with CHARLES cannot be classed as full disagreements.

1,574,661 male given names). (For details, see Fair et al. 1990, 1991.)

The data used in the current study are from the LINKED pairs of names containing any of the 51 most common given names in the death file or any of the 15 most common informal variants. These names are listed in Table 1, together with their counts and percentage frequencies in the death file.

The 51 common names account for more than half (1,842,327/3,397,860) of all given names in the death records of males. Among 64,937 LINKED pairs of male given names, they were present 33,183 times on the Records A (25,673 as first names and 7,510 as second names) and 33,988 times on the Records B (26,536 as first names and 7,452 as second names), for a total of 67,171 times. A name pair that partially agrees may occur in either of two configurations, e.g., as FRANK-FRANCIS or as FRANCIS-FRANK, depending on which value comes from File A and which value comes from File B. Where two or more of the 51 names get interchanged with each other (as happens with HARRY, HENRI, and HENRY), some of the same information may be duplicated in a slightly different form within the tables.

The 15 common informal variants represent less than 1% (28,164/3,397,860) of all given names in the death records of males. Among the 64,937 LINKED pairs of male given names, these were present 1,554 times on the Records A

Table 3. Examples of Partial Agreements That Are Well Represented

Rank	Values*		Numbers observed		
	x	y	Total	$N(A_x \cdot B_y \mid LINK)$	$N(B_x \cdot A_y \mid LINK)$
1.	MICHAEL-MIKE		173	12	161
2.	FREDERI-FRED		169	12	157
3.	ALEXAND-ALEX		152	11	141
4.	JOHN -JACK		90	23	67
5.	FRANCIS-FRANK		73	19	54
6.	JOSEPH -JOE		62	2	60
7.	FREDERI-FREDRIC		52	28	24
8.	ALLAN -ALLEN		47	28	19
9.	HENRY -HENRI		44	40	4
10.	SAMUEL -SAM		37	3	34
11.	PETER -PETE		33	3	30
12.	THOMAS -TOM		33	7	26
13.	WILLIAM-WILLI		20	18	2

* Truncated at seven characters in the LINKS of Fair et al. (1991).

Table 4. Examples of Partial Agreements That Are Not Well Represented

Values*		Total observed	Values*		Total observed
x	y		x	y	
ALBERT	-ALBERTO	1	ALBERT	-ALBERTS	0
ARTHUR	-ARTIMUS	1	ARTHUR	-ARTIMON	0
DOUGLAS	-DOUGLES	1	DOUGLAS	-DOUGLIS	0
ERNEST	-ERNES	1	ERNEST	-ERNE	0
HAROLD	-HARLOD	1	HAROLD	-HARLOE	0
LEO	-LEODA	1	LEO	-LEODAS	0
PETER	-PEDER	1	PETER	-PEDAR	0
VICTOR	-VIATEUR	1	VICTOR	-VIATIAI	0

* Truncated at seven characters in the files of Fair et al. (Fair, Lalonde, and Newcombe (1991)). The synonyms are all represented in the parent files A and B.

(1,426 as first names and 128 as second names) and 701 times on the Records B (633 as first names and 68 as second names), for a total of 2,255 times.

Application of the linkage rationale to outcomes defined in wholly value-specific ways depends on more than just the ODDS formula for its success. The chief obstacle is created by the many value pairs that are rare in the available LINKS, plus the even more numerous possible ones that have not been observed at all. Grouping is necessary, but must be based on wholly value-specific group definitions. The roles played in the process by Files A and B and the LINKS are illustrated in Tables 2-5. Group definitions are based on selected blocks of names in alphabetic listings, chosen to bring rare synonyms into the same groups with common forms (Table 2). Comparison pairs that are common in the LINKS present no special problem (Table 3). However, possible pairs that are rare or absent in the available LINKS need to be grouped with others that are more common (Table 4). ODDS are calculated for specific name pairs and for specific groups as a whole, using expression 1.4 (Table 5). (For details see Fair et al. 1990, 1991.)

There are no rules explicitly stating how the boundaries of the groups should be determined, except that variants known to yield widely different ODDS on their own should not be put into the same group. Apart from this, the process is unavoidably subjective—but it is far from entirely arbitrary. In particular, it is greatly aided by strong impressions gained while perusing alphabetical listings of names from Files A and B.

3. APPLICATION: REFINEMENTS AND SHORTCUTS

Many choices have had to be made in the past between shortcuts in the way the ODDS are calculated versus corresponding refinements in which the shortcuts are not used. Such choices are inescapable, but only rarely have their effects on the calculated ODDS been quantified. Indeed, where data to support the more refined alternative were lacking, the comparison often was not possible. But now the extensive data from large files of LINKS accumulated at Statistics Canada make it attractive to assess the effects on discriminating power when people's names are compared in alternative ways.

Table 5. Comparison Outcomes for the Given Name *GEORGE*, With Examples of Possible Groupings

Values*		Total outcomes	ODDS
x	y		
<i>Full Agreement</i>			
GEORGE-GEORGE		3,130	89.7/1
<i>Partial Agreement</i>			
GEORGE-GEO		6	87.9/1
GEORGE-GEOR to GEORGZ (including GEORDIE)		11	14.9/1
GEORGE-GEORGES		28	12.1/1
GEORGE-GEORGET to GEORGZ (including GEORGIO)		3	21.6/1
<i>Other (including disagreements)</i>			
GEORGE-G* (* = other; few synonyms)		16	1/5.6
GEORGE-non-G (full disagreements)		175	1/13.2

* Data for $A_x \cdot B_y$ and $B_x \cdot A_y$ are pooled.

We consider here six shortcuts (and their corresponding refinements):

1. Use of the simplified formula (see expression 1.5)
2. Pooling of first and second given names, to reduce the number of look-up tables of the value-specific frequencies, $N(B_x)/N(B)$, when using the simplified formula
3. Use of a wholly versus a partially global term in the numerator of the simplified formula when calculating ODDS for the various levels of outcome (i.e., both A_x and B_y being nonspecific in the LINKS, versus A_x being specified as equal, successively, to each of the 51 common names)
4. Not updating the global ODDS
5. Recognizing the specificities of just the agreement portions of names that only partially agree
6. Pooling complementary partial agreements (e.g., $A_x \cdot B_y = \text{MICHAEL-MIKE}$, plus $A_y \cdot B_x = \text{MIKE-MICHAEL}$).

Past and current practices with regard to these shortcuts are reviewed elsewhere (Hill 1981; Howe and Lindsay 1981; Newcombe 1988).

The importance of a given refinement as compared with its corresponding shortcut is assessed by comparing the ODDS when calculated in the two ways. The ratios of the two ODDS will be termed "error factors" or "correction factors." These factors vary for different names as represented in File A (e.g., the given name JOHN) and for different comparison outcomes (e.g., JOHN-JACK). One such type of "correction factor" is defined in the second part of expression 1.5. Its use as part of the full expression constitutes a refinement, its omission constitutes a shortcut, and its use on its own reveals the factor difference between the ODDS as obtained in the two ways.

Comparisons between different refinement/shortcut choices may be based either on the frequency distributions of the error levels, as defined earlier, or on the median and maximum error factors. Sometimes a combination of the two may be appropriate. Data from the six types of comparisons are presented in Figure 2 (parts a to f) and Table 6 (lines 1 to 6). The histograms in Figure 2 are appropriately weighted throughout; for example, in part a of Figure 2 by the frequencies of the names in File A.

The magnitudes of such error factors may vary with the particular name or linkage project; that is, forming a distribution of error factors as shown in Figure 2. The log error factor approach, with base 2, is used in this Figure. (Log error factor = 1 indicates a difference by a factor of 2, log error factor = 2 indicates a difference by a factor of 4, and so on.) Because we are dealing with a spectrum of error factors and need to divide it into discrete levels, we have recognized central values of 1, 2, 4, 8, 16, and so on (equivalent to logs to the base 2 = 0, 1, 2, 3, 4, and so on). Standard rounding of the logs is used to assign the appropriate central values.

3.1 Ranking the Choices

The effect of choosing a shortcut, or its corresponding refinement, is best seen in a listing of the associated error factors in descending order. These create in the mind a compelling picture. What they teach us is that the feedback of actual data does away with the need for guesswork. For our current purposes it is sufficient that the results of the tests be summarized (Fig. 2, Table 6) and that examples be given.

Use of the simplified formula, for example, results in error factors as high as 6.4, with 13% of the 34,737 comparisons associated with the four-fold level of error. Nine of the 51 common names and 5 of the 15 informal names are involved (i.e., DOUGLAS, ERNEST, EMILE, FRANK, HAROLD, CLARENCE, ALFRED, HERBERT, HARRY, FRED, PETE, MIKE, SAM, ALEX). Similarly modest error factors result from pooling of first plus second names, use of a wholly global numerator, and pooling complementary partial agreements. In these examples the magnitudes of the error factors vary with the values of the given names.

The effects of not updating the ODDS differ in that the error factors vary with the quality of the files used to initiate the death searches and, therefore, with the particular linkage study. Error factors are greater for the partial agreements than for the full agreements and disagreements, independent of the actual values of the names; for this reason, only the partial agreements are considered here. Again, the effects of the shortcut are modest. The largest are associated with search files (Files A) in which the quality of the identifiers differed most widely from the average; that is, were either much better

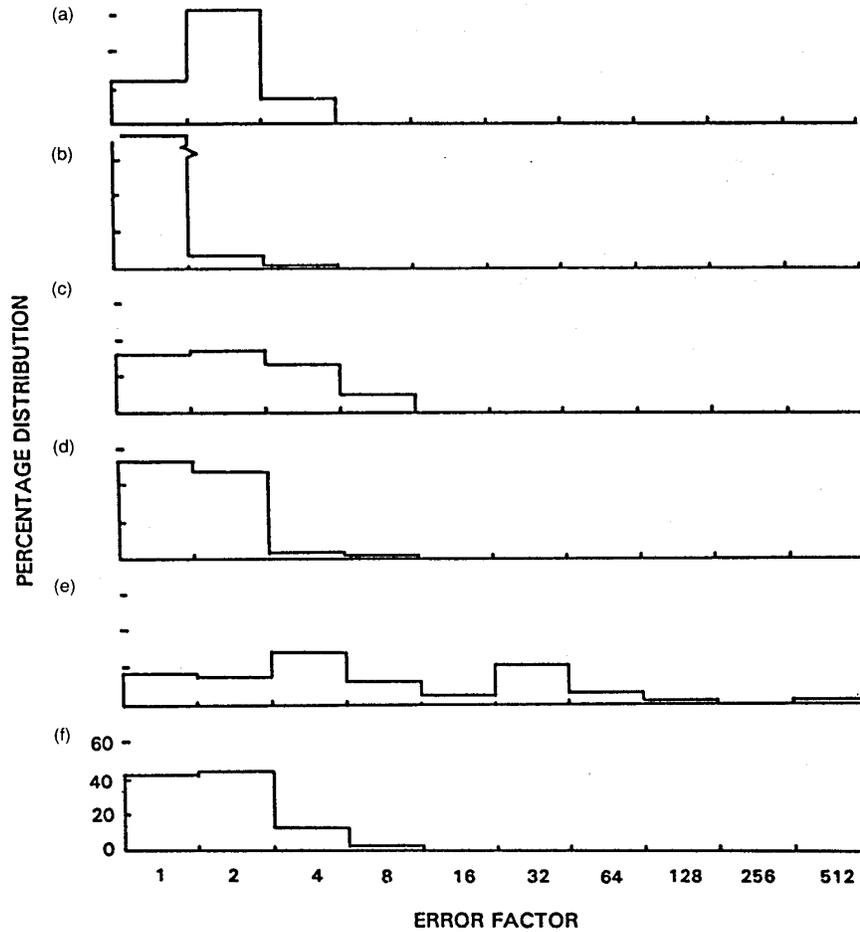


Figure 2. Frequency Distributions of Error Factors Resulting From Shortcuts in the Comparison Procedures for Male Given Names. (a) Simplified formula; (b) Pooled first plus second names; (c) Wholly global numerator; (d) Not updating the odds; (e) Recognizing the specificities of just the agreement portions; (f) Pooling complementary partial agreements.

(as with infant death-to-birth linkages) or much worse (as with certain employment records). This is because in such cases the composite ODDS most poorly represent the ODDS appropriate for the particular project.

Only for one kind of choice are the error factors truly large. This has to do with the practice of recognizing the specificities of only the agreement portions of names that do not fully agree (versus recognizing the full specificities of both members). The most extreme examples

(with their error factors) include WALTER-WLADYSL (686.7), ERNEST-EARNEST (412.7), PETER-PIO (190.2), WILLIAM-BILLY (160.6), ROY-LEROY (155.8), JOHN-JUHO (82.6), LEONARD-LENARD (82.4), RAYMOND-RAIMOND (77.6), LOUIS-LOIS (72.7), and JOHN-JAN (57.7). Only when the full specificities are taken into account does the discriminating power get efficiently exploited.

4. CONCLUSIONS AND RECOMMENDATIONS

Current tests assess the degree to which inherent discriminating power is exploited where names are used to bring together records of the same persons, especially when alternative forms of a name are compared. The emphasis differs from that of procedures based on degrees of phonetic similarity plus lists of exceptions, in that both values get recognized and necessary data are drawn from large accumulations of linked pairs of records.

Motivation to achieve maximum refinement in record linkage comes from the social trend towards larger and more

Table 6. Ranking the Choices Between Refinements Versus Shortcuts: Partial Agreements Only

Shortcut	Median error	Maximum error	Rank*
1. Simple Formula	1.7	6.4	(3)
2. Pooling First and Second	1.1	14.2	(6)
3. Global Numerator	2.1	12.2	(2)
4. Update Omitted	1.4	6.4	(5)
5. Partial Specificity	4.9	686.7	(1)
6. Complementary Partials	1.4	11.2	(4)

* Rank based on median error factor, followed by maximum.

numerous personal data banks. Complex influences govern the trend. Records proliferate because people rely on governments and the commercial sector for increased security and benefits of many sorts, plus conveniences and luxuries where possible. The process is slowed by fears that the right to privacy might suffer, but it is also accelerated by public insistence on a right to know whether perceived threats to health and well-being are real, because the best answers often come only through increased access to personal data banks (in Canada, see Bouchard, Roy, and Casgrain 1985; Fair 1989; Jordan-Simpson, Fair, and Poliquin 1988; Leyes 1990; Medical Research Council of Canada 1968; Newcombe et al. 1983; Roos, Wajda, and Nichol 1986; Smith and Newcombe 1980, 1982; elsewhere, see Arellano, Petersen, Pettiti, and Smith 1984; Baldwin, Acheson, and Graham 1987; Copas and Hilton 1990; Jaro 1989; Kilss and Alvey 1985; Patterson 1980; Rogot, Sorlie, Johnson, Glover, and Treasure 1988; Winkler 1989a,b,c,d; also see early reviews by Acheson 1967 and Farr 1875). A logical step in this evolution is the automation of registers embracing whole populations (Dunn 1946; Leyes 1990; Marshall 1947; Redfern 1990; Scheuren 1990).

The current approach follows a general trend in statistics, which is to develop empirical reference distributions using computers, rather than to rely mainly on theoretical distributions. Here, we use large composite files of LINKS (Fair et al. 1990, 1991) and even larger files of random pairs to serve as NONLINKS (Lalonde 1989). Examples as applied to other statistical problems include uses of the "bootstrap" method (Efron and Tibshirani 1986, 1992). Moreover, those involved with linkage technology stress the need to archive empirical data from past linkage studies, and use it to compare the performances of different systems (see, for example, Howe 1986; Howe and Spasoff 1986a,b; Jabine and Scheuren 1986; Scheuren et al. 1986; Science Council of Canada 1986; Smith 1986).

In a sense, we emphasize here a role for semiautomated "learning," from past experience. Complexity need not be a serious barrier, because complex procedures, once developed, may be used repeatedly and can evolve through successive refinements.

[Received October 1989. Revised May 1991.]

REFERENCES

- Acheson, E. D. (1967), *Medical Record Linkage*. Oxford, U.K.: Oxford University Press.
- Arellano, M. G., Petersen, G. R., Pettiti, D. B., and Smith, R. E. (1984), "The California Automated Mortality Linkage System," *American Journal of Public Health*, 74, 1324-1330.
- Baldwin, J. A., Acheson, E. D., and Graham, W. J. (eds.) (1987), *Textbook of Medical Record Linkage*. Oxford, U.K.: Oxford University Press.
- Bouchard, G., Roy, R., and Casgrain, B. (1985), *Reconstitution Automatique des Familles, le Système SOREP* (Vols. I and II), Chicoutimi, Quebec: Centre Interuniversitaire de Recherches sur les Populations (SOREP).
- Copas, J. B., and Hilton, F. J. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society, Ser. A*, 153 (Part 3), 287-320.
- Dunn, H. L. (1946), "Record Linkage," *American Journal of Public Health*, 36, 1412-1416.
- Eagen, M., and Hill, T. (1987), "Record Linkage Methodology and its Application," in *Statistical Uses of Administrative Data, Proceedings of an International Symposium*, eds. J. W. Coombs and M. P. Singh, Ottawa: Statistics Canada, pp. 139-150.
- Efron, B., and Tibshirani, R. (1986), "The Bootstrap Method for Assessing Statistical Accuracy" (with discussion), *Statistical Science*, 1, 54-77.
- (1992), "Statistical Data Analysis in the Computer Age," *Science*, in press.
- Fair, M. E. (1989), *Studies and References Relating to Uses of the Canadian Mortality Data Base*, Ottawa: Statistics Canada, August 1989.
- Fair, M. E., Lalonde, P., and Newcombe, H. B. (1990), *Tables of ODDS For Partial Agreements of Male Given Names in Linking Records*, Report OEHRs No. 9, Ottawa: Statistics Canada.
- (1991), "Application of Exact ODDS for Partial Agreements of Names in Record Linkage," *Computers and Biomedical Research*, 24, 58-71.
- Fair, M. E., Newcombe, H. B., and Lalonde, P. (1988a), *Improved Mortality Searches for Ontario Miners Using Social Insurance Index Identifiers*, Report No. INFO-0264, Ottawa: Atomic Energy Control Board.
- Fair, M. E., Newcombe, H. B., Lalonde, P., and Poliquin, C. (1988b), "Alive" Searches as Complementing Death Searches in the Epidemiological Follow-Up of Ontario Miners, Report No. INFO-0266, Ottawa: Atomic Energy Control Board.
- Farr, W. (1875), in *Supplement to the 35th Annual Report of the Registrar General*, London: Her Majesty's Stationery Office, p. 110.
- Fellegi, I. P. (1985), "Tutorial on the Fellegi-Sunter Model for Record Linkage," in *Record Linkage Techniques—1985 (Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985)*, eds. B. Kilss and W. Alvey, Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 127-138.
- Fellegi, I. P., and Sunter, A. B. (1969), "A Theory of Record Linkage," *Journal of the American Statistical Association*, 40, 1183-1210.
- Hill, T. (1981), *Generalized Iterative Record Linkage System: GIRLS*. Ottawa: Statistics Canada.
- Hill, T., and Pring-Mill, F. (1985), "Generalized Iterative Record Linkage System," in *Record Linkage Techniques—1985 (Proceedings of the Workshop on Exact Matching Methodologies Arlington, Virginia, May 9-10, 1985)*, eds. B. Kilss and W. Alvey, Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 327-333.
- Howe, G. R. (1986), "Possible Future Directions in Record Linkage," in *Proceedings of the Workshop on Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 231-233.
- Howe, G. R., and Lindsay, J. (1981), "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," *Computers and Biomedical Research*, 14, 327-340.
- Howe, G. R., and Spasoff, R. A. (eds.) (1986a), *Proceedings of the Workshop on Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*. Toronto: University of Toronto Press.
- (1986b), "Recommendations of the Workshop on Computerized Linkage in Health Research," in *Proceedings of the Workshop on Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 18-23.
- Jabine, T. B., and Scheuren, F. (1986), "Record Linkages for Statistical Purposes: Methodological Issues," *Journal of Official Statistics*, 2, 255-277.
- 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.
- Jordan-Simpson, D., Fair, M. E., and Poliquin, C. (1988), "Canadian Farm Operator Study: Methodology," *Health Reports (Statistics Canada)*, 2, 141-155.
- Kilss, B., and Alvey, W. (eds.) (1985), *Record Linkage Techniques—1985 (Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985)*, Washington, DC: Department of the Treasury, Internal Revenue Service.
- Lalonde, P. (1989), "Deriving Accurate Weights Using Non-Links," in *Proceedings of the Record Linkage Sessions and Workshop, Canadian Epidemiology Research Conference—1989*, eds. M. Carpenter and M. E. Fair, Ottawa: Statistics Canada, pp. 149-157.
- Leyes, J. (1990), "Release of a Pilot Longitudinal Administrative Database," *The Daily (Statistics Canada)*, Monday, October 22, 1990, p. 6.
- Marshall, J. T. (1947), "Canada's National Vital Statistics Index," *Population Studies*, 1, 204-211.
- Medical Research Council of Canada (1968), *Health Research Uses of Record Linkage in Canada*. Report No. 3, Ottawa: Author.
- Newcombe, H. B. (1967), "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," *American Journal of Human Genetics*, 19, 335-359.
- (1988), *Handbook of Record Linkage: Methods for Health and Sta-*

- tistical Studies, Administration and Business*, Oxford, U.K.: Oxford University Press.
- Newcombe, H. B., Fair, M. E., and Lalonde, P. (1987), "Concepts and Practices that Improve Probabilistic Record Linkage," in *Statistical Uses of Administrative Data. Proceedings of an International Symposium (Ottawa, Ontario, November 23-25, 1987)*, eds. J. W. Coombs and M. P. Singh, Ottawa: Statistics Canada, pp. 127-138.
- (1989), "Discriminating Powers of Partial Agreements of Names for Linking Personal Records, Part I: The Logical Basis, and Part II: The Empirical Test," *Methods of Information in Medicine*, 28, 86-91, 92-96.
- Newcombe, H. B., and Kennedy, J. M. (1962), "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Communications of the Association for Computing Machinery*, 5, 563-566.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.
- Newcombe, H. B., Smith, M. E., Howe, G. R., Mingay, J., Strugnell, A., and Abbott, J. D. (1983), "Reliability of Computer versus Manual Death Searches in a Study of Eldorado Uranium Workers," *Computers in Biology and Medicine*, 13, 157-169.
- Patterson, J. E. (1980), "The Establishment of a National Death Index in the United States," in *Cancer Incidence in Defined Populations (Banbury Report No. 4)*, eds. J. Cairns, J. L. Lyon, and M. Skolnick, Cold Spring Harbor, Long Island, New York, Cold Spring Harbor Laboratory, pp. 443-451.
- Redfern, P. (1990), "Sources of Population Statistics: An International Perspective," in *Population Projections: Trends, Methods and Uses*, OPCS Occasional Paper 38, London: Office of Population Censuses and Surveys, Her Majesty's Stationery Office.
- Rogot, E., Sorlie, P. D., Johnson, N. J., Glover, C. S., and Treasure, D. W. (1988), *A Mortality Study of One Million Persons: First Data Book*, NIH Publication No. 88-2896, Bethesda, MD: Public Health Service, National Institutes of Health.
- Roos, L. L., Wajda, A., and Nicol, J. P. (1986), "The Art and Science of Record Linkage: Methods that Work with Few Identifiers," *Computers in Biology and Medicine*, 16, 45-57.
- Scheuren, F. (1985), "Methodological Issues in Linkage of Multiple Data Bases," *Record Linkage Techniques—1985*, Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 155-178.
- (1990), Discussion of "Rolling Samples and Censuses," by L. Kish, *Survey Methodology*, 16, 72-79.
- Scheuren, F., Alvey, W., and Kilss, B. (1986), "Record Linkage for Statistical Purposes in the United States," in *Proceedings of the Workshop in Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 198-210.
- Science Council of Canada (1986), *Proceedings: A National Workshop on the Role of Epidemiology in the Risk Assessment Process in Canada*, Catalogue No. SS24-23/1985, Ottawa: Author.
- Smith, M. E. (1986), "Future Needs and Directions for Computerized Record Linkage in Health Research in Canada: Future Study Plans," in *Proceedings of the Workshop in Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 211-230.
- Smith, M. E., and Newcombe, H. B. (1975), "Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories," *Methods of Information in Medicine*, 14, 118-125.
- (1979), "Accuracies of Computer Versus Manual Linkages of Routine Health Records," *Methods of Information in Medicine*, 18, 89-97.
- (1980), "Automated Follow-up Facilities in Canada for Monitoring Delayed Health Effects," *American Journal of Public Health*, 73, 39-46.
- (1982), "Use of the Canadian Mortality Data Base for Epidemiological Follow-up," *Canadian Journal of Public Health*, 73, 39-46.
- Sunter, A. B. (1968), "A Statistical Approach to Record Linkage," in *Record Linkage in Medicine (Proceedings of the International Symposium, Oxford, July 1967)*, ed. E. D. Acheson, London: E & S Livingstone, pp. 89-109.
- Winkler, W. E. (1985), "Preprocessing of Lists and String Comparison," in *Record Linkage Techniques—1985*, eds. W. Alvey and B. Kilss, Washington, DC: Department of the Treasury, U.S. Internal Revenue Service, pp. 181-187.
- (1989a), *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage (Technical Report)*, (paper presented at the Annual ASA Meeting in Anaheim, CA) Washington, DC: Statistical Research Division, U.S. Bureau of the Census.
- (1989b), "The Interaction of Record Linkage Practice and Theory," in *Proceedings of the Record Linkage Sessions and Workshop, Canadian Epidemiology Research Conference—1989*, eds. M. Carpenter and M. E. Fair, Ottawa: Statistics Canada, pp. 139-148.
- (1989c), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Fifth Census Bureau Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 145-155.
- (1989d), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, 15, 101-117.

Comment

MAX G. ARELLANO*

Because the discussion is focused primarily on first name variants, the title perhaps should more appropriately be "The Use of Given Names for Linking Personal Records." While not stated, the implication is that the "special skill" developed by humans is of considerable value in the decision making process. The fact is that the "special skills" vary considerably from person to person and that the biases that they bring to the evaluation may hinder rather than assist in the record linkage process.

The "past experience" argument is spurious. There is no reason to believe that the lessons learned from a Canadian

mortality study will be of any benefit to an evaluation of a Cuban expatriate population or that experience gained in a study of mortality among Chicago nurses will be of any benefit to a study of child abuse in Seattle.

It does not follow at all that "if a machine is to acquire similar ability, it too must rely on past experience." For instance, an analysis of the decisions made by the operators may well reveal that their judgments are based primarily on their perceptions of probability of occurrence and the reliability of the data. These factors are quantifiable and not dependent on past experience.

* Max G. Arellano is Chief Scientist, Advanced Linkage Technologies of America, Inc., Berkeley, CA 94707.

1. COMPUTER LINKAGE

The presentation is much too informal. It is difficult enough to figure out how the authors derive their likelihood ratio without trying to deduce how they arrived at the “two additional factors.” What are the consequences of failure to recognize these two factors? The authors’ arguments would be much easier to follow if they were presented in mathematical terms.

I fail to see the relevance of “describing the insights on which their success depends” in “plain language . . . such as they would understand” to the clerical searchers?

Probabilistic linkage procedures must be based on a probability model with definable probability distribution or density functions. The Fellegi-Sunter model is a probability model; I see no evidence of a probability model in this article. This is not to say that there is no merit in the approach presented in this article; however, it should properly be presented as a subjective probability or expert system.

The concept of falsely matched random pairs is a fascinating topic. But if, as the authors state in the first paragraph of page 7, “it was not difficult to determine the corresponding likelihoods for a control group of falsely matched random pairs (NONLINKS),” then why didn’t they present the procedures that they used to obtain this value? I believe that this would have contributed immensely to their presentation.

I understand the context within which the historical development is being presented, and I am in complete sympathy with the authors’ objectives. However, the point must be made that the validity of the linkage rationale that they describe is a function of the correctness of the linkage decisions that were made. A statement is badly needed regarding whether it was possible to confirm their decisions or how they were able to establish a level of confidence in them. After all, this is the central issue in record linkage.

The authors seem to feel that refinements in linkage decision criteria only proceeded “independent of the formal theory” (p. 1194). There is no reason, however, to believe that these or similar developments could not have or did not proceed within the context of the formal theory, perhaps without the knowledge of the authors.

Routine cross-comparisons can also be extremely wasteful of available resources if they are not called for by the nature of the data. In most linkage evaluations, 85–90% of the correct linkages can readily be detected as exact matches on name and birthdate.

The authors state that “frequent close scrutiny of difficult matches provided insights that would have been missed had refinement been sought through theory alone” (pp. 1194–1195). In view of the fact that, as the authors readily admit, their procedures are not based on the formal theory, the validity of this statement is doubtful. How can they be sure of the correct direction of these “difficult matches” without reference to the subjects whose records are being linked?

In the development of their decision criteria, Fellegi and Sunter stated very clearly that the effect of their weight computation is “to array the record pairs, relative to one another, in descending order of the likelihood of a correct match.”

If the authors had observed the strict requirements of the

Fellegi-Sunter model, they would have realized that the restriction of the comparison-space to linkages with identical surname phonetic codes requires an adjustment to the computation of the surname weights. This adjustment would have compensated for the distortion that they observed in the “crossover” point.

The discussion of prior likelihood is unnecessarily vague. What are prior likelihoods? How are they estimated? It is not sufficient to simply show these as $P(\text{LINK})/P(\text{NONLINK})$.

The authors would do better to present their derivation in terms of the Fellegi-Sunter model. Within the context of the Fellegi-Sunter model, there is no need for concern about “fortuitous LINKS in the random pairs.”

“ODDS” should be expanded on. ODDS of what?

One cannot have conditional probabilities without either a probability distribution or density function. I don’t see any evidence of either.

The derivation leads to the conclusion that we need to know the number of links with value A_x and B_y (p. 1196). But this is exactly what we are trying to accomplish with the linkage; that is, this information is not known. The authors gloss over this point without explaining how they intend to fill in the blanks.

The “tradeoff” argument (p. 1196) is completely spurious. The categories are determined by the characteristics of the data. It is not reasonable to assume that the operators of linkage software can be expected to ensure that every outcome group is broad enough so that “ $N(A_x \cdot B_y | \text{LINK})$ is represented by at least one comparison pair. Otherwise no ODDS can be calculated” (p. 1196). This sounds as though the procedure is controlling the application. Linkage software can readily be designed so that empty categories are either assigned zero values or some predetermined default value.

Partial given name agreements can be easily handled by phonetically encoding the name and constructing an exception list. This procedure has been in use by most organizations with which I am familiar for at least the past 16 years.

The authors state that “confusion still remains concerning the implications, and is not explicitly addressed by existing formal theory” (p. 1197). The authors are obviously privy to some controversy to which I am not.

We keep coming back to the fact that $N(A | \text{LINK})$ is unknown. The authors should have expanded on how they obtain this value.

2. EMPIRICAL DISTRIBUTIONS OF LINKS AND NONLINKS

The authors apparently believe that the results of particular linkage evaluations can be extrapolated to other linkage evaluations. Although this may be true in general, it cannot be relied on as a matter of policy. For instance, the reporting of demographic information by psychiatric patients may be much less reliable than information gathered for epidemiologic research purposes, the point being that “memories of past encounters with similar problems” may well lead you astray.

Although the authors criticize the practice of simply multiplying the ODDS for individual identifiers to combine them

for a whole set “which would only be proper if they were independent of each other,” (p. 1198) this appears to be exactly what they do—or do they believe that the $P(\text{LINK})$ term corrects for the dependence among the identifiers?

The authors appear obsessed by the presence of false-positive links and false-negative links. The purpose of a record linkage, however, is not to eliminate these links, but rather to minimize them. There is a point beyond which the cost of refining the rules outweighs the advantages of applying them, particularly if the refinement requires an extensive amount of manual review.

The authors state that the number of possible pairings is the product of the two file sizes. This is true, however, only if all possible pairwise comparisons are actually formed between the two files—a practice that would be prohibitively expensive. The actual number of pairings is a function of the blocking strategy that was used. The difference is not at all trivial.

The problem to which the authors allude beginning on page 1199, of establishing upper and lower threshold values is not related to the independence problem. It is a function of the far greater size of NONLINKS relative to the LINKS —a fact, by the way, that is well known to persons involved in probability linkage, despite the concerns expressed by the authors. The threshold problem would exist even if a correction for the dependence of the identifiers could be incorporated into the computation of the total odds.

The “strong impressions gained while perusing alphabetic listings of names from Files A and B” (p. 1200) are of value only if their validity can be established by reference to the truly valid linkages. Under any circumstances, however, unless these “impressions” can be translated into formal rules, these procedures are obviously not suitable for mass production purposes.

3. CONCLUSIONS AND RECOMMENDATIONS

Rarely, if ever, does an experienced human clerk obtain feedback regarding the validity of a difficult linkage decision. Without this information, the clerk cannot possibly know whether his intuition was correct or not. If the clerk is not routinely receiving this feedback, the rules he has been developing may well lead to the systematic introduction of error into the decision criteria he is applying to the linkages.

The authors contend that the thought patterns (of the “experienced human clerk”) clearly differ from those of a skilled mathematician. However, the consensus among most persons involved in probability linkage with whom I am familiar is that subjective judgment is based on perceptions of prob-

abilities of occurrence, a feel for the reliability of the data, and a familiarity with the various ways in which the same item of information can be recorded. There is no mystery; all of these factors are readily quantifiable.

Before one can “learn” from past experiences (p. 1203), two elements are necessary:

1. One must rigorously define how to measure a “success.” The authors have failed to do so.

2. One must demonstrate that the lessons learned from a particular linkage evaluation have relevance to the new linkage evaluations that are under active consideration. Personally, I would hesitate to apply the lessons which the authors have learned from their Canadian experience to our ongoing linkage evaluations in California.

4. REVIEWER'S SUMMARY

The authors' bias toward an informal approach to the development of linkage decision criteria is obvious, as is their sentiment that no real value can come from pursuing formal probability linkage models such as the Fellegi-Sunter model. One must ask, however, if the authors are aware of any objective basis for their assertion that an informal approach is superior to an approach based on a formal mathematical model.

Organizations with which I have been affiliated have used various versions of the Fellegi-Sunter probability linkage model for the past 17 years, with a great deal of success. Our linkage evaluations have included files with over one million records. Although manual review of the borderline linkages is an essential element of our linkage processing, because of the very large number of linkages identified it would be impractical for us to become overly involved in resolution of the difficult matches. Although we routinely observe the instances in which there is a substantial amount of conflict among the identifiers, I would question the wisdom of applying the lessons learned from the outcome of one difficult match to another difficult match.

Newcombe would do well to explore the operation of systems that use a formal probability linkage model; perhaps he would then gain a greater appreciation of them. We welcome his call for greater mutual cooperation. If there is sufficient interest, we would be glad to participate in a comparative linkage methodology evaluation study.

REFERENCE

Fellegi, I., and Sunter, A. (1969), “A Theory for Record Linkage,” *Journal of the American Statistical Association*, 64, 1183–1210.

HOWARD B. NEWCOMBE, MARTHA E. FAIR, and PIERRE LALONDE

Arellano has provided a detailed critique of our article, much of which does not actually contradict what we have said or conflict with our own understanding, even though the language may differ. Any rejoinder, therefore, should confine itself to major points of difference on matters of emphasis or fact.

We do *not*, for example, believe that added refinement is always cost-effective in all situations. But by exploring the ways in which the comparison space may be more finely partitioned, we hope to expand both the present and the future *potential* for improved linkage performance at acceptable cost. It was the crudity of the popular agreement-disagreement distinction that provided the initial major motivating force. What impressed us as a source of innovation was the wealth of alternative comparison procedures and of multiple alternative outcome definitions, that got applied freely by a person's mind. Many of these proved highly effective in the case of difficult links, once the true status of the record pairs was confirmed later by independent means.

The emphasis we have placed on multiple partitioning of the comparison space has applications that are not confined to any particular identifier field. For example, colleagues at one time were concerned that our recognition of multiple outcomes from comparisons of place of work with place of death, when doing death searches, might be contrary to linkage theory. The observation was that workers at an Ontario uranium refinery who migrated before dying tended to die more often, either in the home province or in western Canada, but less often in eastern Canada and only rarely in the most easterly province, as compared with the random expectation. A somewhat different pattern (i.e., empirical distribution) was observed for workers in the uranium mines of Saskatchewan and the Northwest Territories; so there was no question of extrapolating from one subset of the cohort to the other. Here, final verification of the linkage status of the record pairs was not in doubt. Even before that verification, however, approximate likelihood ratios contributed to the linkage process and to the updating and iterative refinement, both of the linked files and of the likelihood ratios together. To establish useful comparison rules, we first needed to "learn" what only the linked files could "teach" concerning the empirical distributions and the outcome definitions most likely to exploit their discriminating power to good advantage. Earlier objections to the approach were later withdrawn. But if this broad emphasis on added partitioning of the comparison space to reveal a greater diversity of usable differences in observed versus random distributions is indeed fundamentally flawed, as Arellano seems to believe, we would welcome from him a concrete example to that effect.

We also appreciate Arellano's stated interest in "comparative linkage methodology evaluation studies," especially if this interest encompasses the current focus on given names. Thus he could readily compare his own practice of recog-

nizing phonetic similarity plus an exception list with our wholly value-specific approach, using Canadian data that have been published in great detail for just such a purpose (Fair, Lalonde, and Newcombe 1990). Moreover, Figure 2 of our article indicates a convenient way to display the results. Indeed, the two approaches need not be mutually exclusive, since ours provides what might be viewed as just a very long "exception list" based on the most appropriate data for searches of the particular File B.

We are aware that in principle any use of data from old linkages when starting a new linkage operation must involve some degree of extrapolation, at least initially. But this is not necessarily so for the later stages, after there has been opportunity for iterative adjustments based on the new links.

Arellano has alluded to a number of exceedingly simple concepts which appear to him to give rise to logical difficulties. For example:

- "We keep coming back to the fact that $N(A|LINK)$ is unknown."
- "The concept of falsely matched random pairs is a fascinating topic. But, . . . why didn't they present the procedures that they used to obtain this value?"
- "The problem . . . of establishing an upper and lower threshold value is not related to the independence problem."

At the risk of repeating what is in the article, we will consider these together here:

- $N(A|LINK)$: The simple answer is that one may do a small preliminary linkage, perhaps manually, to arrive at the approximate proportion of records in File A that will find a correct match in File B. There is no serious obstacle to this because, as Arellano points out, often 85 to 90% of the linkages are easy anyway. What is curious about the question itself is that this first step is the same as is routinely employed to obtain preliminary estimates of the likelihood ratios. The process thereafter, of iteratively refining early crude estimates, has been repeatedly emphasized in the literature (e.g., see Howe and Lindsay 1981).
- Random Pairs: Again, only modest ingenuity is needed to solve the problem. Where the outcomes of interest are defined in complicated ways, there is no need to resort to theory to determine their frequencies of occurrence in random pairs. Instead, one uses the computer to put together large numbers of random pairs, among which the proportions of the outcomes of special interest may be determined by tabulation (Lalonde

1989). Alternatively, for simple value specific outcomes such as ROBERT compared with BOB, the random expectation is just the product of the proportions of these two values in Files A and B (or Files B and A) prior to linking.

- **Thresholds and Independence:** The statement that lack of independence has no effect on the placing of the upper and lower thresholds is too sweeping to be correct. Where *correlated disagreements* (e.g., due to multiple wrong guesses on the part of an informant) have spuriously moved true links downward below the lower threshold, or where *correlated agreements* of rare specific values (e.g., of ethnic surnames and forenames, plus places of birth) have spuriously moved false matches upward above the upper threshold, preset thresholds will no longer accurately perform their intended function. Such effects are often too large to be ignored when setting the thresholds.

Initially it had *not* been our intention to raise in this article the contentious matter of the “prior likelihood” of a correct match on a single random pairing. Indeed, we did not invent the concept—but we did devise the procedure for estimating the magnitude. For all practical purposes, prior likelihoods are essentially similar to the “prior odds” that appear explicitly in the weight formula of Howe and Lindsay. The idea is also implicit in the Fellegi-Sunter theory, where two conditional probabilities (i.e., of a link and of a nonlink) are described (Fellegi and Sunter 1969, exs. 6 and 7, pp. 1185–1186). Each contains a term for a prior probability (of a match and of a non-match, respectively) before the comparison of any identifiers. These terms are $P[(a, b) | M]$ and $P[(a, b) | U]$, and their ratio represents the prior odds contained in the Howe-Lindsay weight formula. In an early version of our article, Figure 1 drew criticism from reviewers as being unsupported and incorrect. This is why details of

our use and derivation of an estimated “prior likelihood” are included here together with the related idea that *blocking* be treated as *not* altering, either the total number of *possible* record pairings (actual plus potential), or the use of likelihood ratios derived from the blocking identifiers. Indeed, unless valid links are known to be lost due to blocking and their numbers can be estimated, there is no special reason why blocking need make any difference at all to the calculation of total weights or absolute odds in favor of a correct match.

Alternatively, of course, one may legitimately view each block as containing its own Files A and B; then, likelihood ratios for blocking identifiers are ignored, but a separate prior likelihood is required for every block, which may be cumbersome. Falling in between these two legitimate alternatives is a common practice that recognizes blocks and ignores likelihood ratios based on blocking identifiers, but *omits* the prior likelihood. Test results from this might seem satisfactory where the blocks happen to be small and most search records find a correct match, but it is hardly justified on logical grounds. As well, for searches of an accumulated national death file, with large blocks based only on a single surname code and with most cohort members still alive, the scale of odds that this incomplete treatment yields does not even remotely approximate the absolute scale needed for predefined error bounds.

Finally, although we are mindful of major differences of emphasis in various workers, we are unaware of any fundamental conflict between our approach and existing theory. If Arellano believes that there is such a conflict, we hope that its nature will get spelled out clearly in the future. Because much of record linkage development and application is of necessity in the hands of people trained in disciplines other than mathematics, any such clarifications ought to be in a form understandable by all who are engaged in implementing the linkage rationale.