

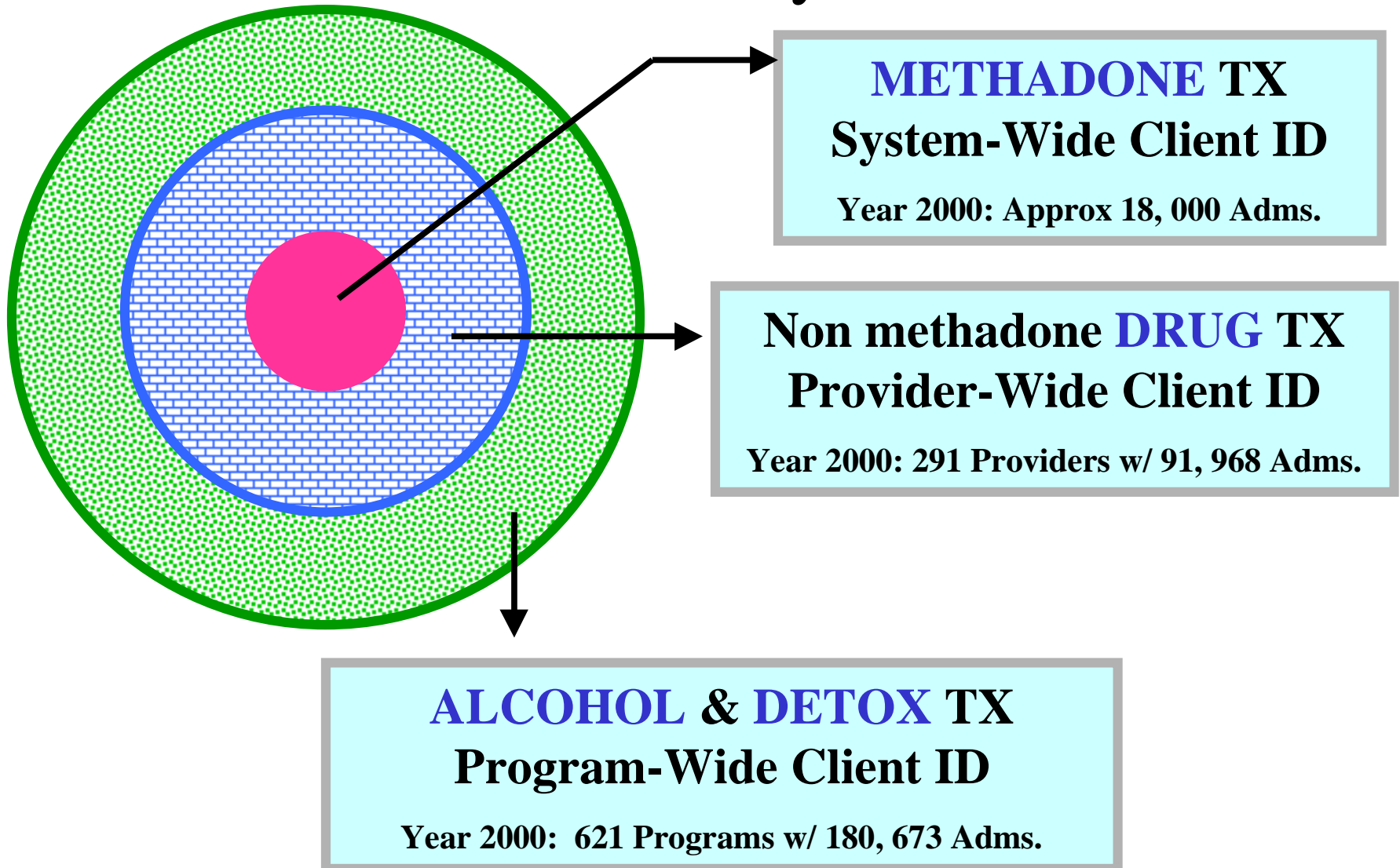
**Approaches to Large Scale Probabilistic
Record Linking for New York State
Alcohol and Other Drug Treatment Data**

Nelson Toth

**New York State Office of Alcoholism
and Substance Abuse Services**

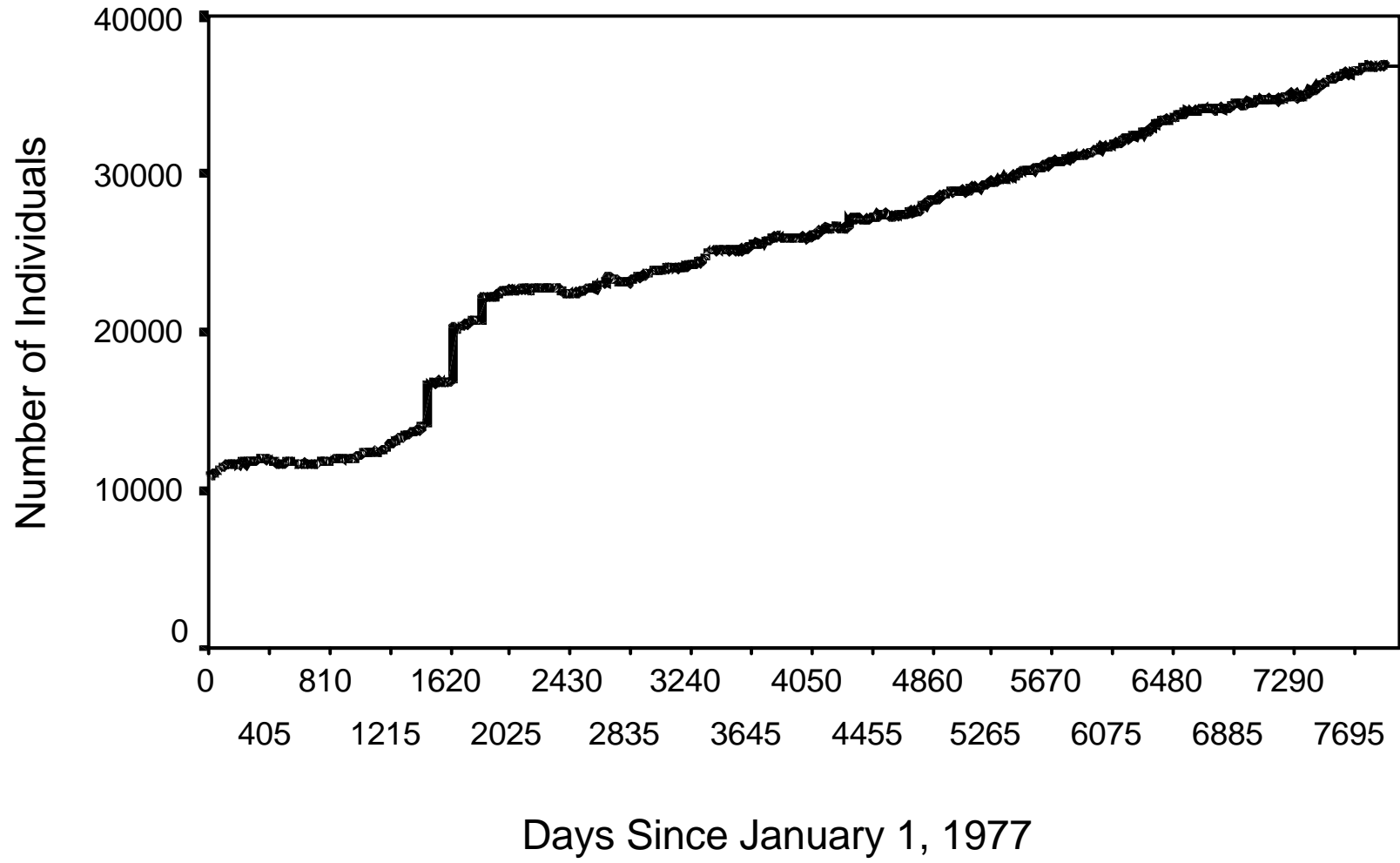
Research Supported by Grant 5-H79-T1112237 from the Center for Substance Abuse Services

Client Tracking in the OASAS Client Data System



Daily Treatment Load

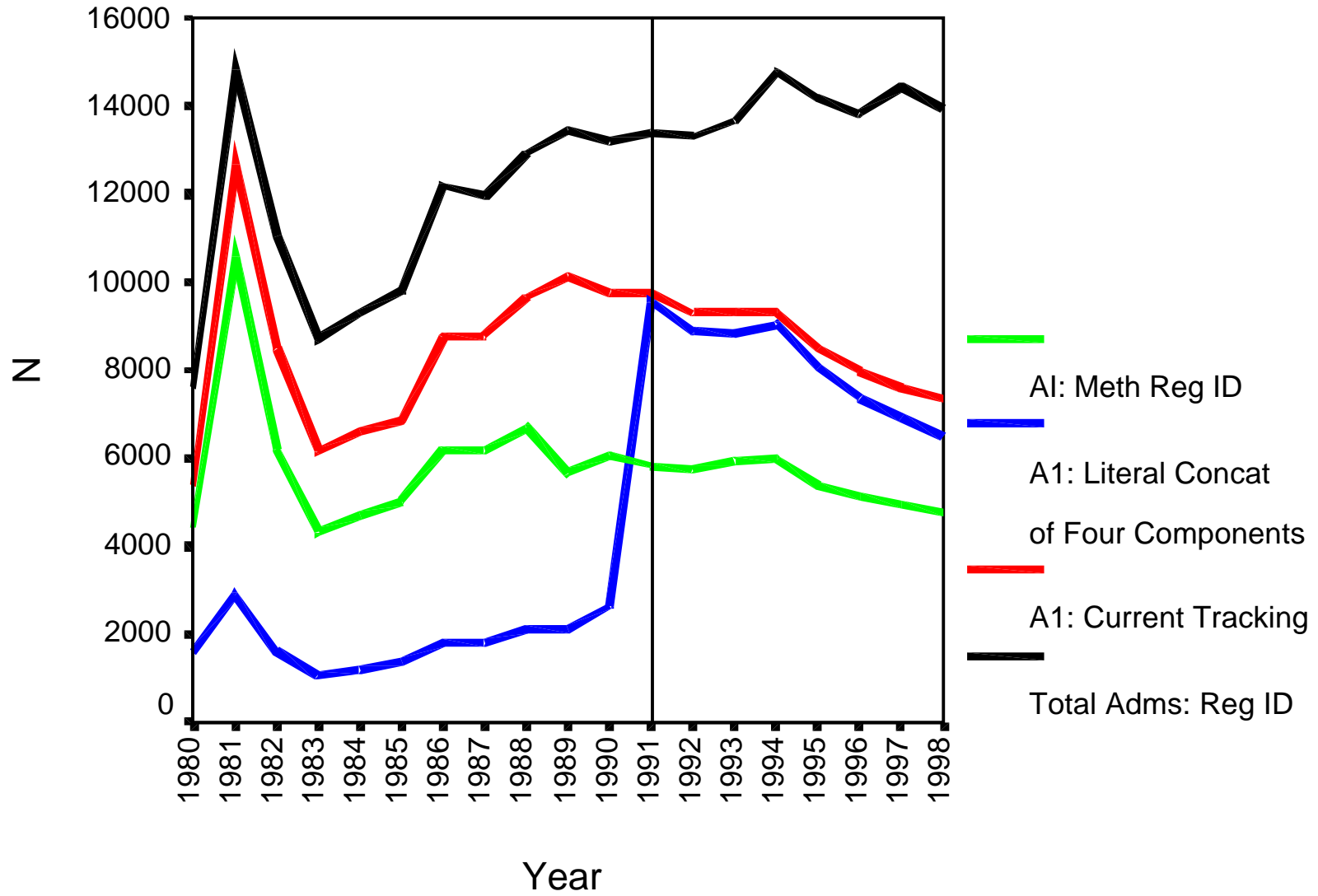
Methadone Maintenance: 1977 - 1997



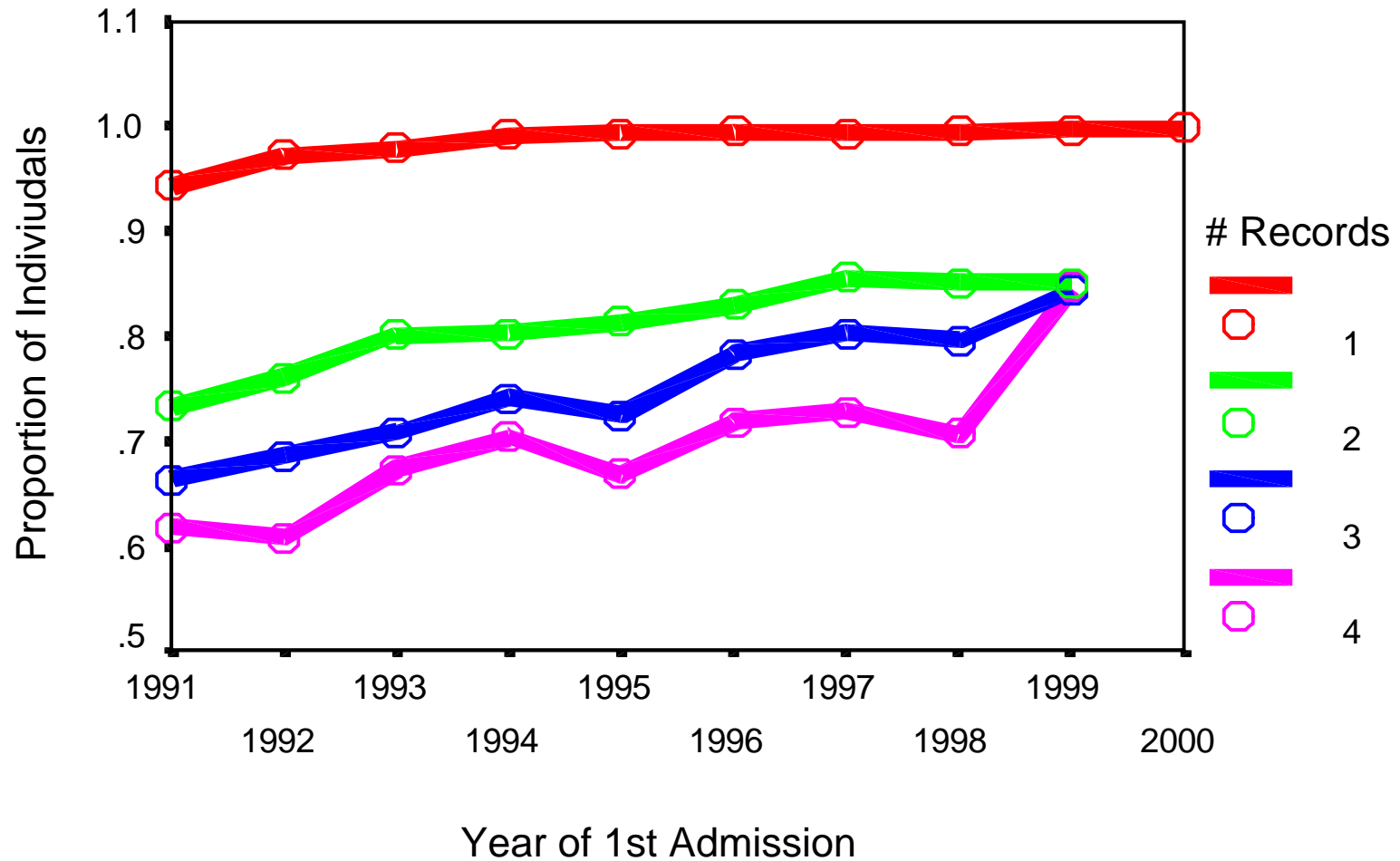
Components of Client Tracking

- 1. Gender**
- 2. Year of Birth**
- 3. Month of Birth**
- 4. Day of Birth**
- 5. Last Four Digits of SSN**
- 6. First Two Characters of Last Name**

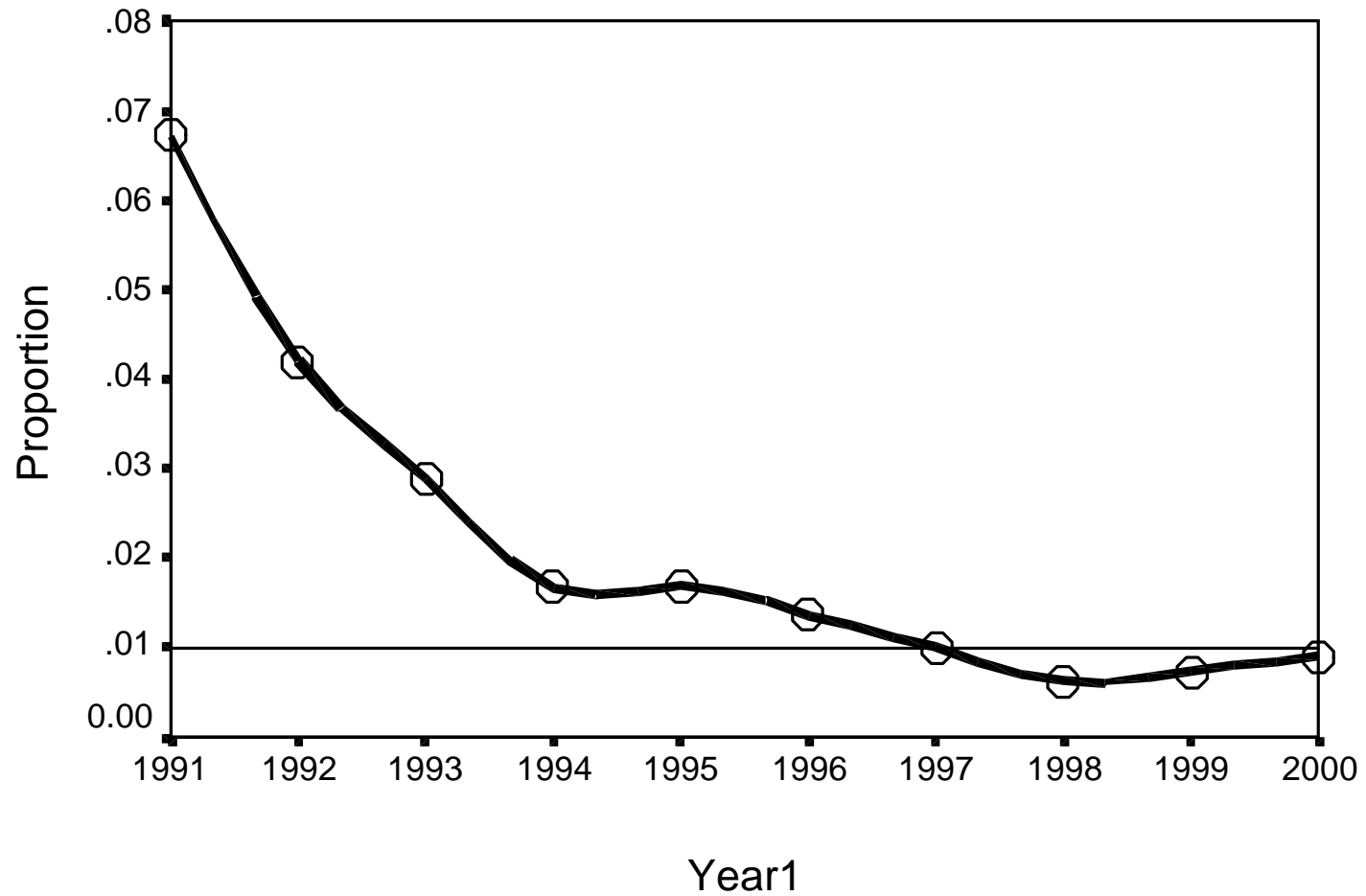
Admissions To Methadone Maintenance



Proportion of Individuals w/ Valid & Consistent Tracking Components



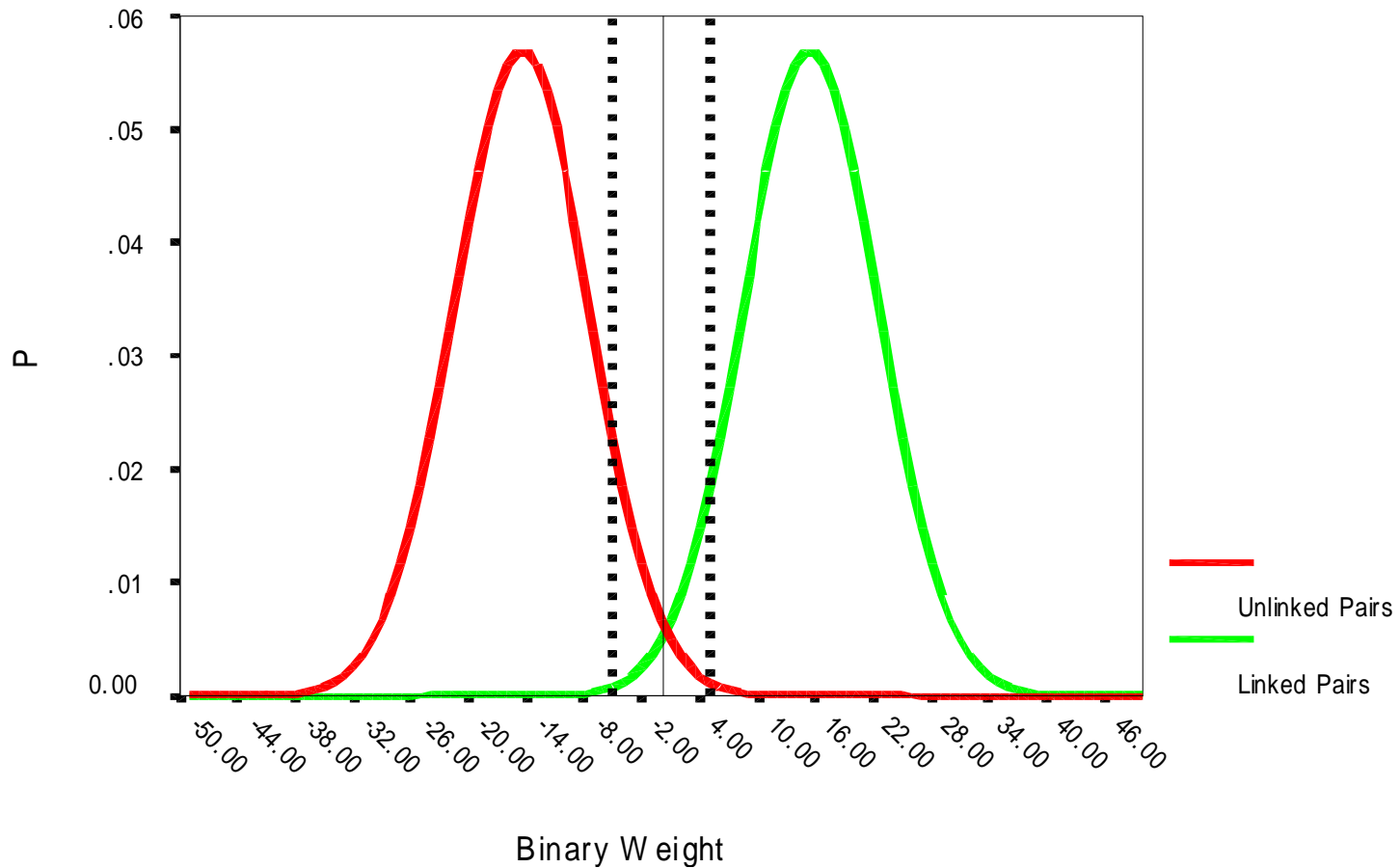
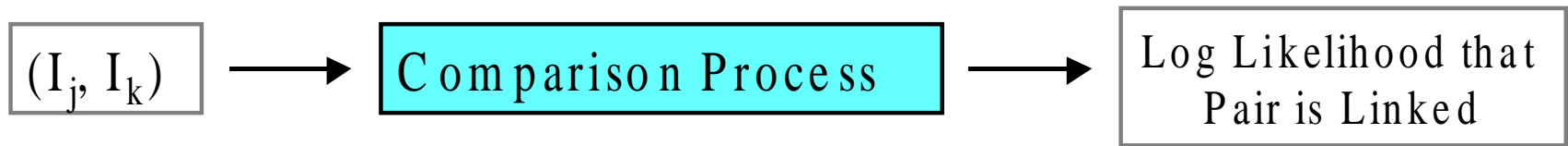
Mean Proportion of Linked Pairs with Invalid L4SS or LNAME values



Inconsistency of Tracking Data

REG. ID	Year	Sex	DOB	L4SS	LNAME
129119	1995	1	19700911	3191	LO
129119	1995	1	19700911	3192	LO
129119	1992	1	19700911	9999	LO
129119	1991	1	19700911	3119	LO
129119	1991	1	19700911	3292	LO
159811	1995	2	19020106	8601	ME
159811	1994	2	19620105	8601	ME
159811	1994	2	19620105	8610	ME
159811	1993	2	19620105	8601	ME
159811	1993	2	19620105	9999	XX

Overlapping Distributions and Threshold Setting



Pairwise Comparison of Tracking Components



Tracking Components

1. Gender
2. Birth Year
3. Birth Month
4. Birth Day
5. Last 4 Digits of SSN
6. First 2 Characters of Last Name

N_1 : # Positioned Correctly
 N_2 : # Positioned Incorrectly
 N_3 : # In I_2 and Not In I_1

Gender: $(1, 1) \longrightarrow (1,0,0)$

L4SS: $(8762, 8675) \longrightarrow (1,2,1)$

Birth Month: $(12, 09) \rightarrow (0,0,2)$

Total of 145, 800 Categories

Standard Model

Multinomial Distribution in Six Variables Conditioned on Year1-Year2 Pair

$$\mathbf{f}_L([g_1 g_2 g_3], [y_1 y_2 y_3], [m_1 m_2 m_3], [d_1 d_2 d_3], [l4_1 l4_2 l4_3], [ln_1 ln_2 ln_3] \mid \mathbf{yp})$$

$$\mathbf{f}_U([g_1 g_2 g_3], [y_1 y_2 y_3], [m_1 m_2 m_3], [d_1 d_2 d_3], [l4_1 l4_2 l4_3], [ln_1 ln_2 ln_3] \mid \mathbf{yp})$$

- Distributions empirically derived from Methadone data.
- No assumption of independence.
- Weight of evidence in favor of H_L given observed categories is

$$\log_2(f_L/f_U) \text{ bits}$$

Examples of Standard Model

$$f_L([1,0,0], [4,0,0], [2,0,0], [2,0,0], [4,0,0], [2,0,0] \mid 91-93) = .740800$$
$$f_U([1,0,0], [4,0,0], [2,0,0], [2,0,0], [4,0,0], [2,0,0] \mid 91-93) = .000003$$

$$(f_L/f_U) = 251,520.4$$

$$\log_2(f_L/f_U) = +17.94$$

$$f_L([1,0,0], [2,1,1], [1,0,1], [0,0,1], [4,0,0], [2,0,0] \mid 94-97) = .000299$$
$$f_U([1,0,0], [2,1,1], [1,0,1], [0,0,1], [4,0,0], [2,0,0] \mid 94-97) = .000002$$

$$(f_L/f_U) = 1,788.2$$

$$\log_2(f_L/f_U) = +10.80$$

$$f_L([1,0,0], [2,1,1], [1,0,1], [0,0,2], [0,1,3], [0,1,1] \mid 94-99) = .000365$$
$$f_U([1,0,0], [2,1,1], [1,0,1], [0,0,2], [0,1,3], [0,1,1] \mid 94-99) = .002796$$

$$(f_L/f_U) = .1304$$

$$\log_2(f_L/f_U) = -2.94$$

Enhanced Model

Distribution for Value of Tracking Components

Conditioned on Comparison Outcome Category and Year1-Year2 Pair

$$\lambda_g = \log_2 (f_L(\text{gender pair} \mid [g_1, g_2, g_3], \text{yp}) / f_U(\text{gender pair} \mid [g_1, g_2, g_3], \text{yp}))$$

$$\lambda_y = \log_2 (f_L(\text{doby pair} \mid [y_1, y_2, y_3], \text{yp}) / f_U(\text{doby pair} \mid [y_1, y_2, y_3], \text{yp}))$$

$$\lambda_m = \log_2 (f_L(\text{dobmm pair} \mid [m_1, m_2, m_3], \text{yp}) / f_U(\text{dobmm pair} \mid [m_1, m_2, m_3], \text{yp}))$$

$$\lambda_d = \log_2 (f_L(\text{dobdd pair} \mid [d_1, d_2, d_3], \text{yp}) / f_U(\text{dobdd pair} \mid [d_1, d_2, d_3], \text{yp}))$$

$$\lambda_{ln} = \log_2 (f_L(\text{lname pair} \mid [ln_1, ln_2, ln_3], \text{yp}) / f_U(\text{lname pair} \mid [ln_1, ln_2, ln_3], \text{yp}))$$

Log Likelihoods for the Values of Component Pairs are added to the Standard Model

$$\log_2(f_L/f_U) + \lambda_g + \lambda_y + \lambda_m + \lambda_d + \lambda_{ln}$$

Formal Measures of Information

$$I(L : U) = \sum f(c)_L \log_2 \left(\frac{f(c)_L}{f(c)_U} \right)$$

Mean Information per Observation from {L} in Favor of H_L over H_U

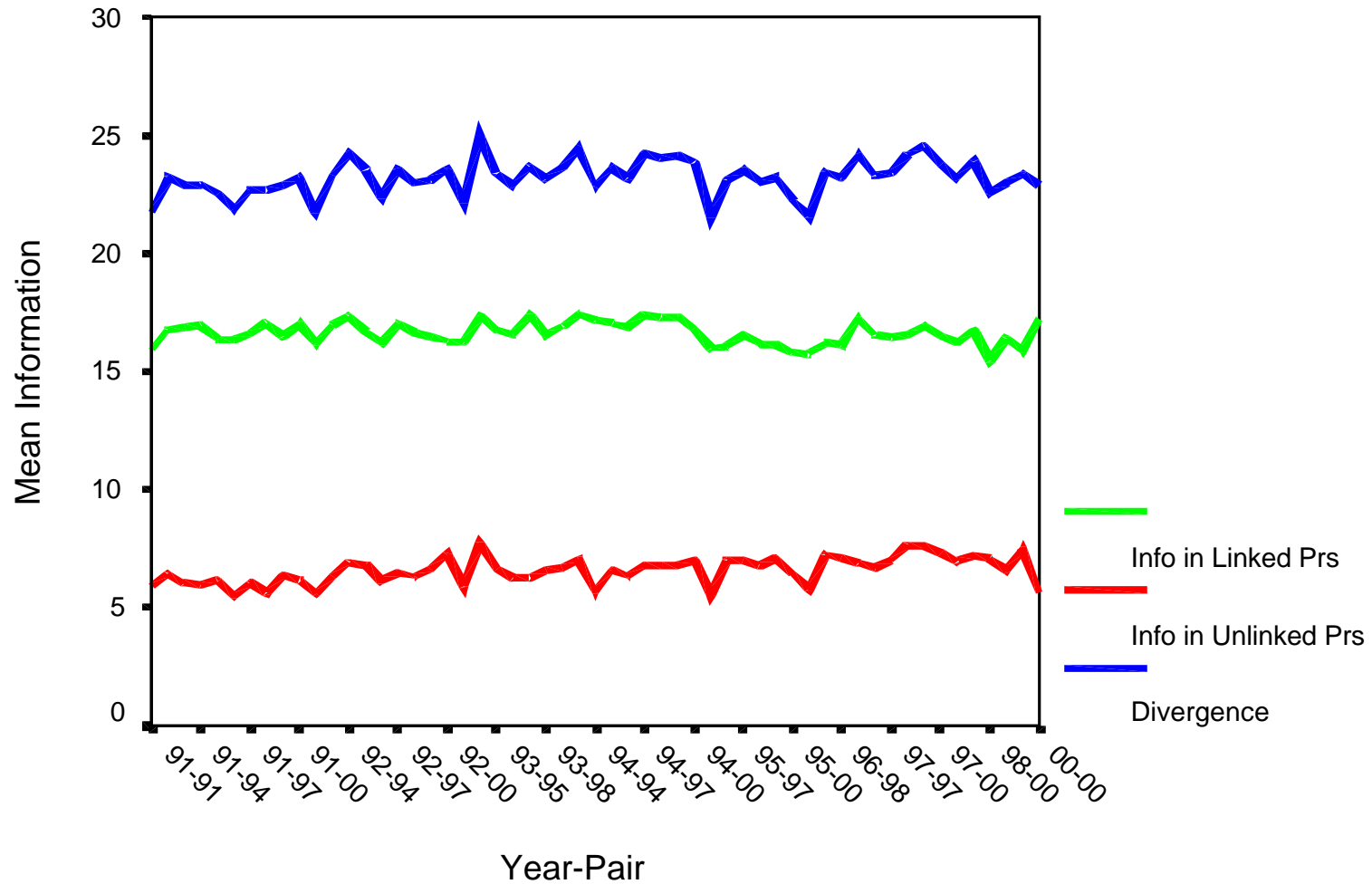
$$I(U : L) = \sum f(c)_U \log_2 \left(\frac{f(c)_U}{f(c)_L} \right)$$

Mean Information per Observation from {U} in Favor of H_U over H_L

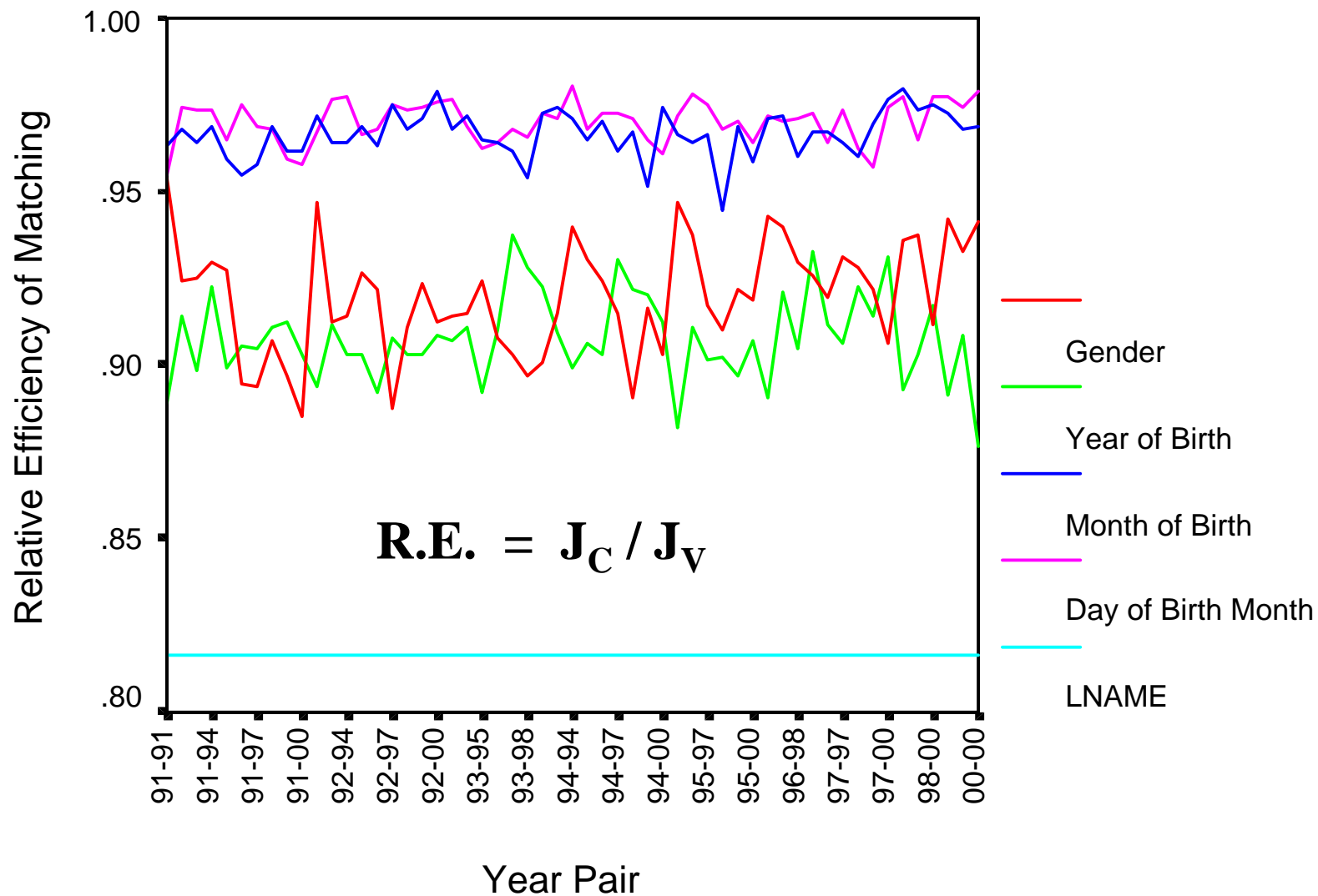
$$J(L : U) = \sum (f(c)_L - f(c)_U) \log_2 \left(\frac{f(c)_L}{f(c)_U} \right)$$

Divergence: A measure of the discriminability between H_L and H_U

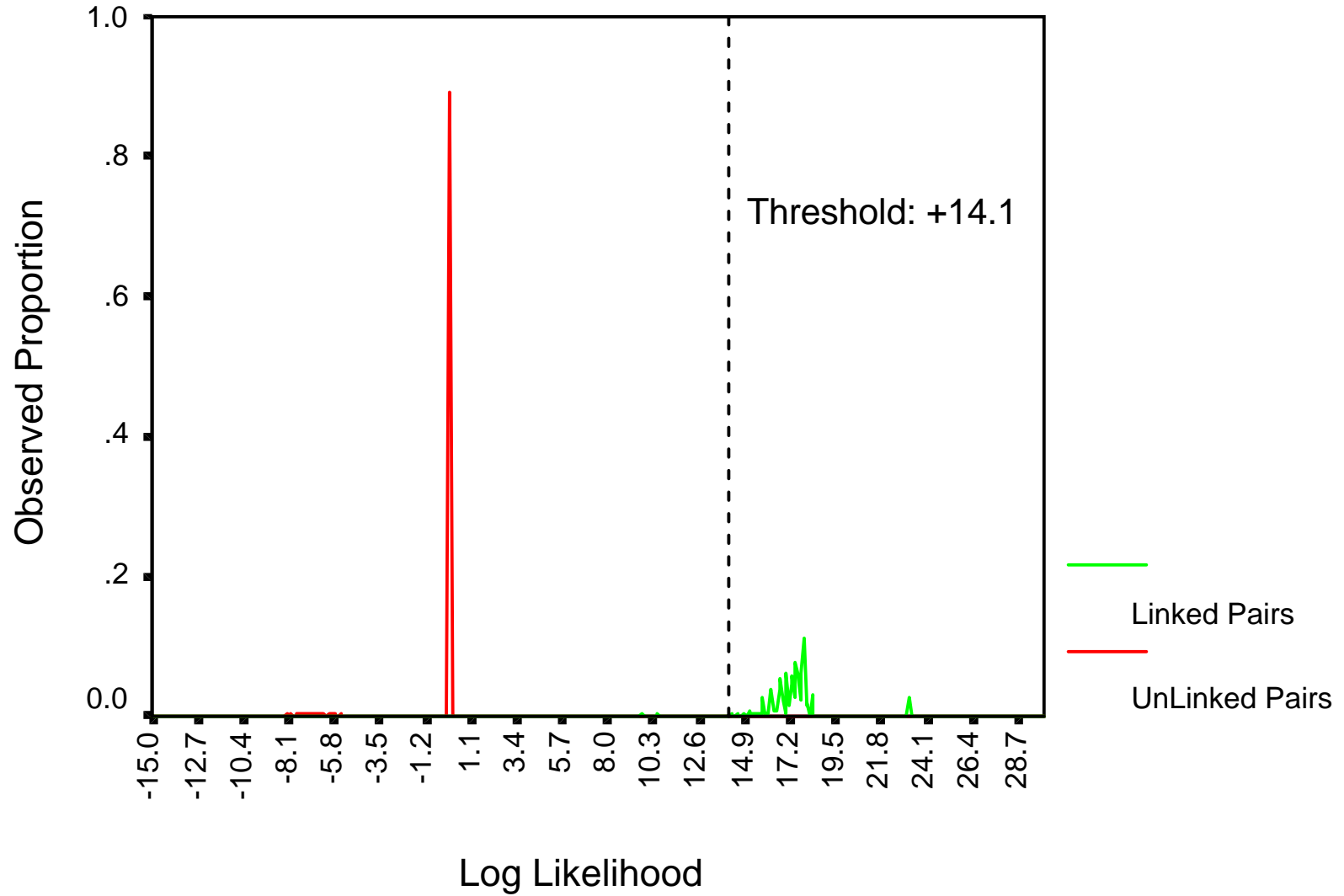
Information Measures for Standard Model



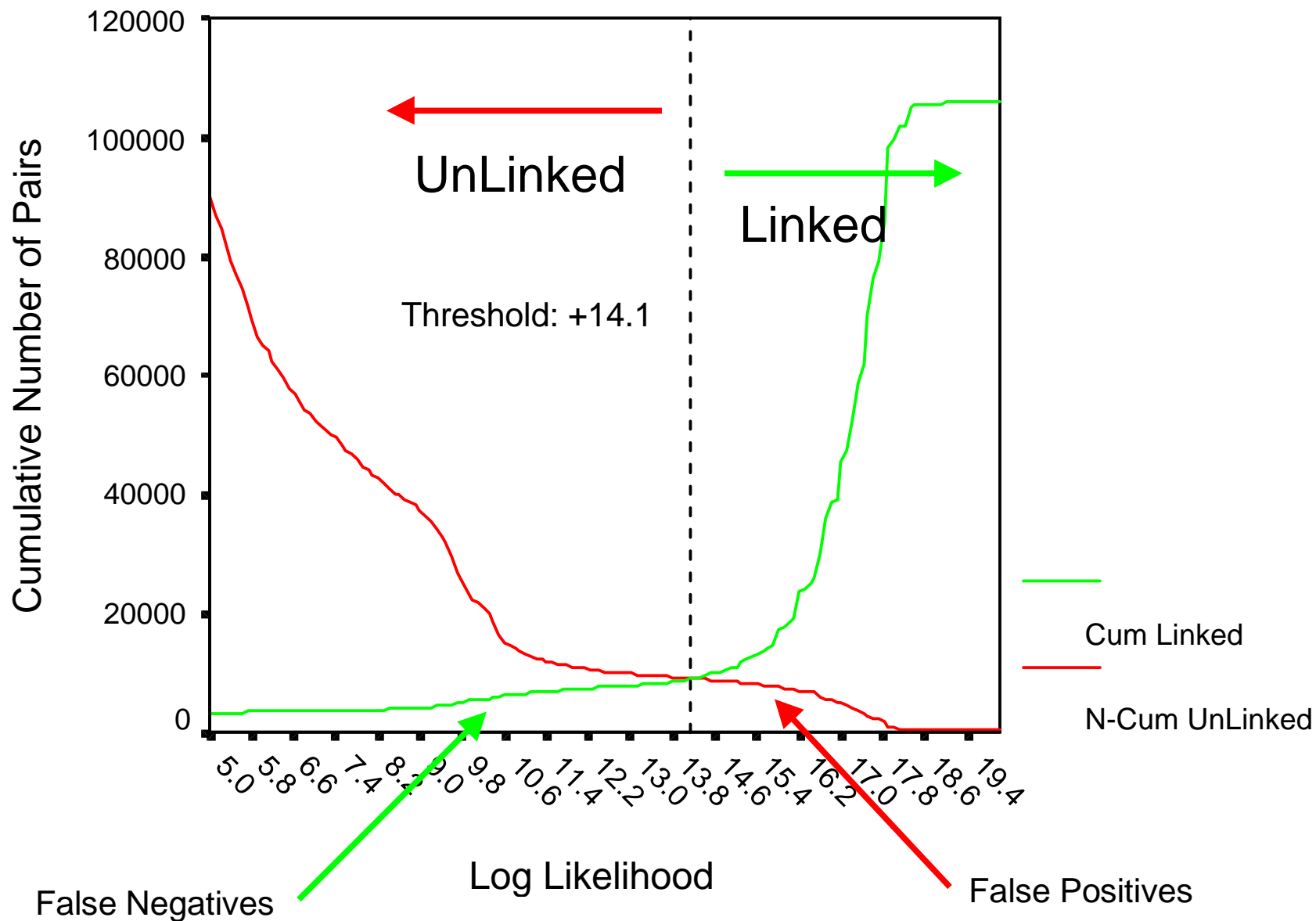
Relative Efficiency of Matching



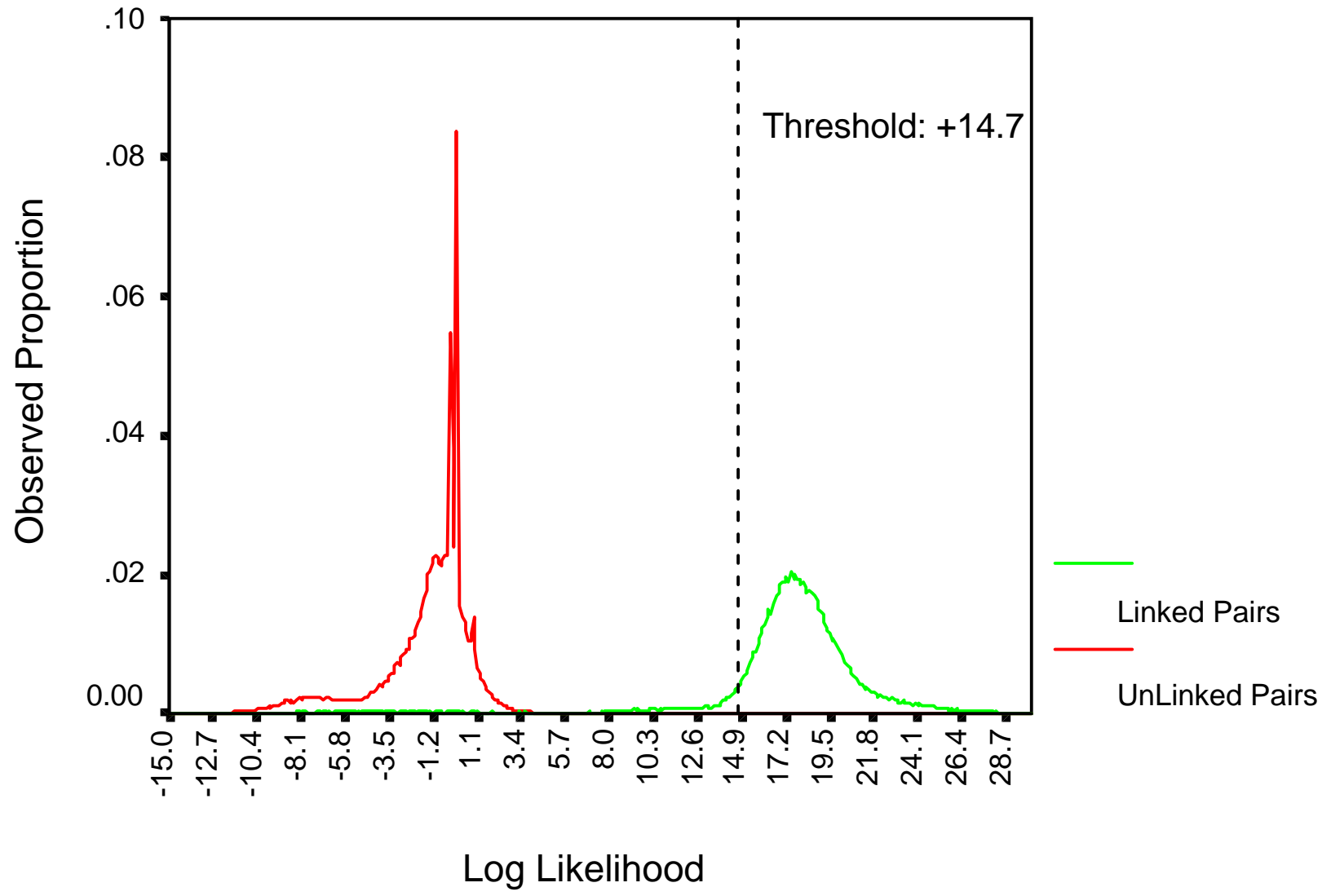
Standard Model: Aggregated Data



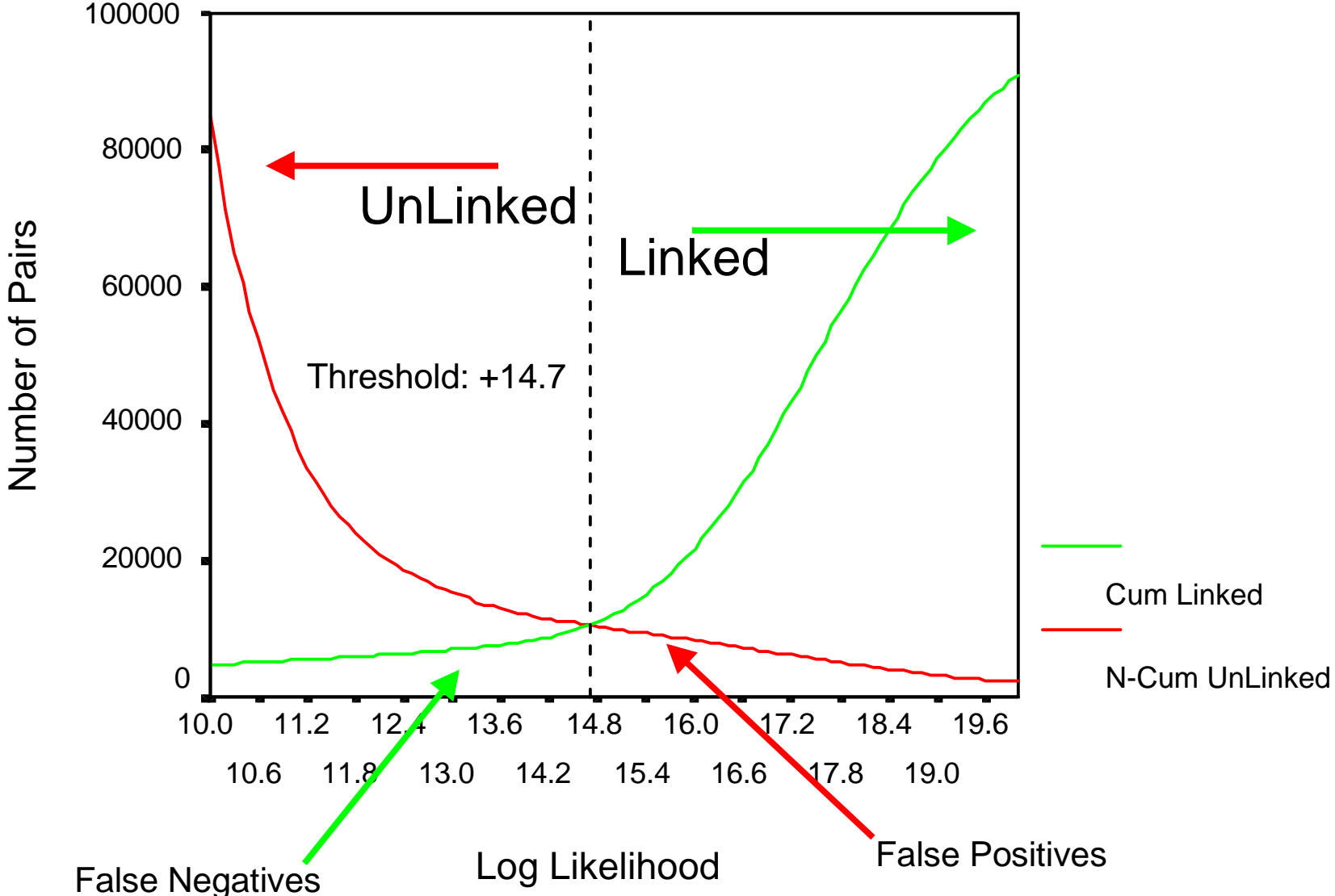
Standard Model: Aggregated Data



Enhanced Model: Aggregated Data



Enhanced Model: Aggregated Data



OVERVIEW

Probabilistic Record Linkage in the CDS

