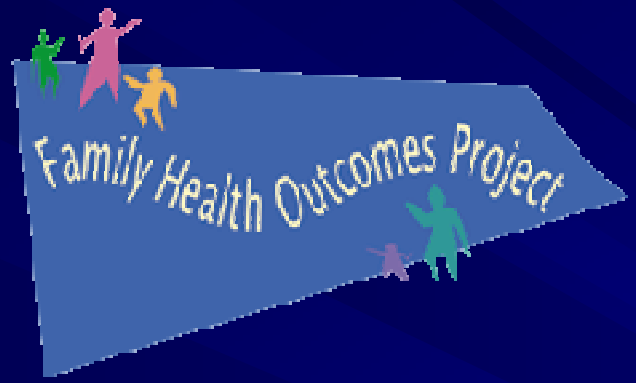


Political and Public Policy Implications of Missing Data and Unlikely Values: The Example of California's Birth Certificate Data

APHA 2007

Washington DC

Linda Remy, MSW PhD
Gerry Oliva, MD MPH
Jennifer Rienks, PhD
Ted Clay, MS

Family Health Outcomes Project
Department of Family and Community Medicine
University of California, San Francisco

Background

- Gould J and Chavez G AJPH 92(1) 79-81 2002 found that:
  - In California, birth certificates are more likely to be incomplete for infants who subsequently die.
  - the higher a sub-population's risk of poor outcomes, the greater the likelihood that birth records will be incomplete.
  - Gestational age is much less likely to be calculated if the mother's race is other than non-Hispanic White
- Excluding records with missing and unlikely values when calculating health indicator rates likely underestimates cases at high risk of poor outcomes and incorrectly estimates progress toward Healthy People 2010 objectives.

Background (cont.)

- The National Center for Health Statistics (NCHS) addresses data quality in several ways
  - NCHS calculates % missing for birth certificate variables by State. If a state falls below 1.5 times the 1998 median and above 1% remedial action is required
  - NCHS edits BC data to correct for missing and unlikely values before calculating indicators
- Kotelchuck M (1994) APNCU index uses a SAS algorithm to impute gestational ages based on baby weight and gender. A PDF copy of his code is available: www.mchlibrary.info/databases/HSNRCPDFs/APNCU99420SAS.pdf

California's Experience

- In 2004, California did not meet NCHS standards on mother's Hispanic origin, education, and month and year of last menstrual period.
- Under California law, the Birth Stat Master File, used for all indicator reports at the state level, must reflect exactly what is on the BC. Therefore no edits are done and missing values are excluded.
- The CADPH Center for Health Statistics (CHS) and the Maternal and Child Health Branch (MCAH) collaborated to develop a strategy to address the issue of data quality

California's Experience (cont.)

- CHS began a targeted training program for birth clerks in hospitals to improve the completeness and quality of data for the fields specified by NCHS
- FHOP was asked to study the quality of BC data fields at the county and hospital level before and after the training program
- In the process FHOP assessed the impact of poor quality on key perinatal health indicators that used the deficient fields, to explore the implications for planning and policy development at the state level

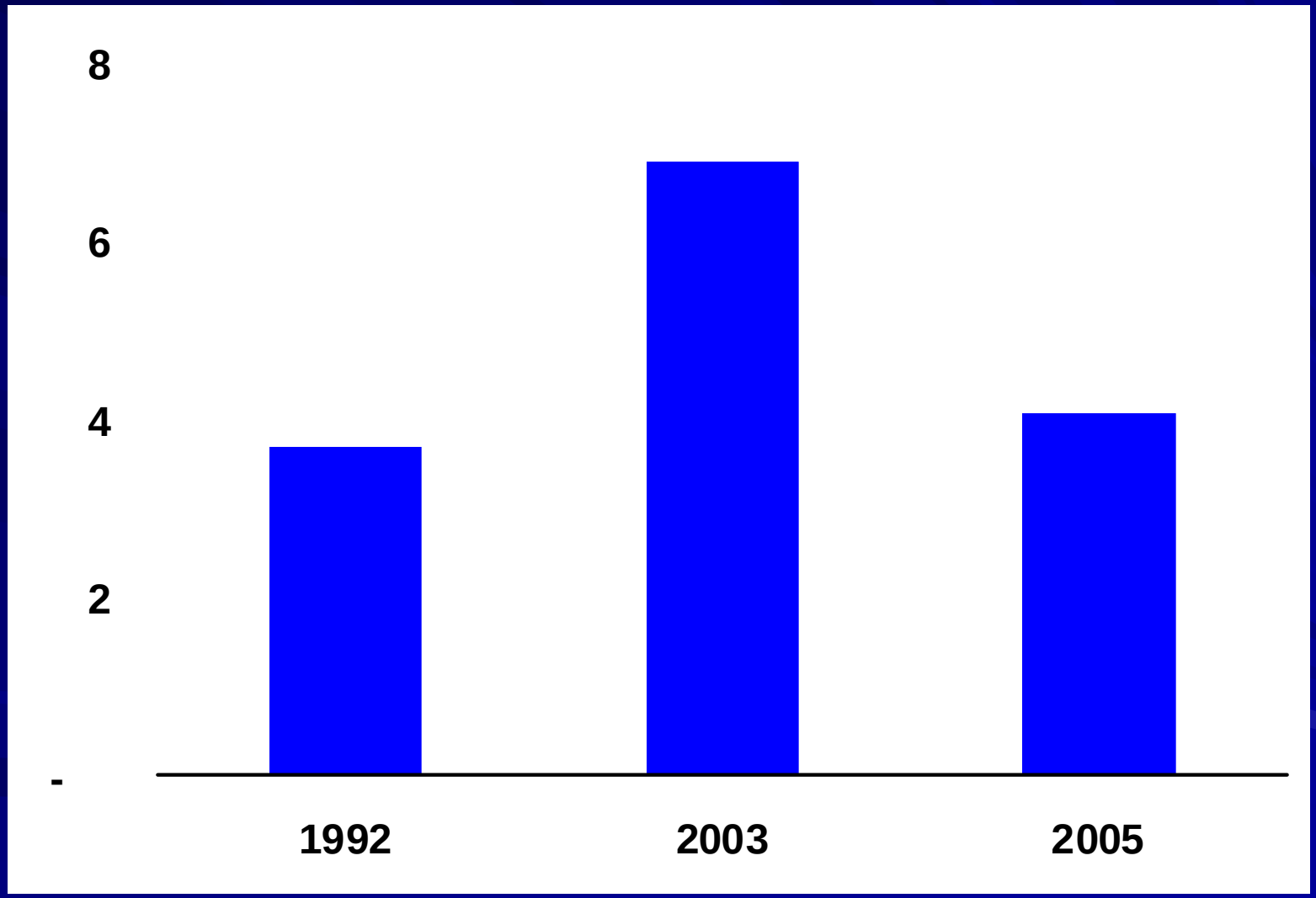
Study Objectives

- To quantify and describe rates of missing and unlikely values for gestational age between 1989 and 2005
- To quantify the impact of using the Kotelchuck algorithm to impute preterm birth rates by comparing these rates with those calculated with unedited data
- To assess the impact of edited and unedited data on trends in preterm birth rates
- To assess the impact of data quality on the assessment of race/ethnic disparities

Findings: State Rates for Missing and Improbable Values of Gestational Age (GA)

	Improbable Gestational Age (Weeks)				
Year	Missing	Lt 18	Gt 47	Total	Pct
1989	17,484	306	6,397	24,187	4.2
1990	17,009	273	5,718	23,000	3.8
1991	16,768	292	5,387	22,447	3.7
1992	17,018	150	4,950	22,118	3.7
...	...	...	...	...	...
2002	30,124	286	3,378	33,788	6.4
2003	34,093	204	2,853	37,150	6.9
2004	33,482	262	2,756	36,500	6.7
2005	18,537	324	3,408	22,269	4.1

Change in Improbable GA



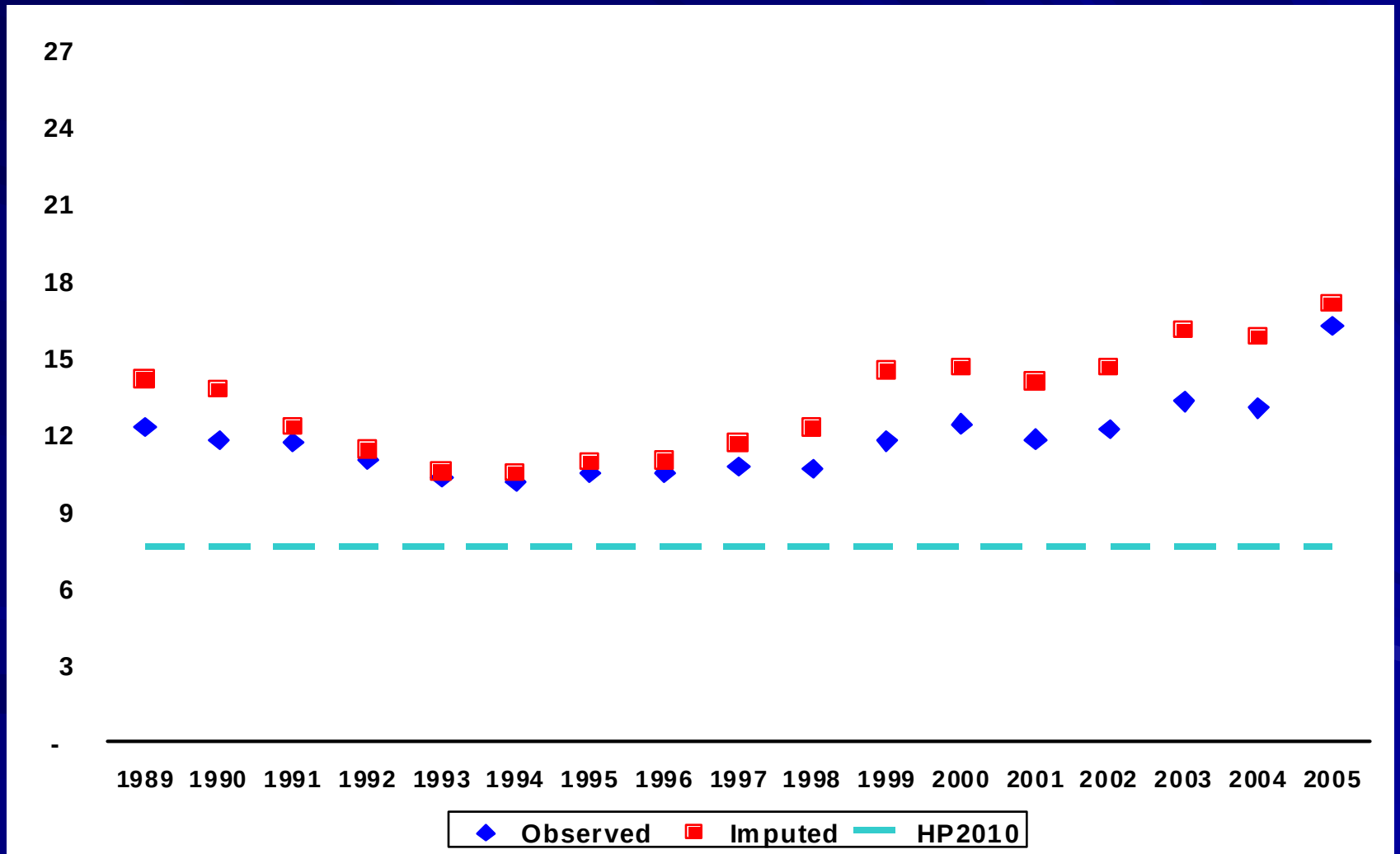
Between 1989 and 2005

- Improbable GA (missing, less than 18 weeks, more than 47 weeks) ranged from a low of 3.7% (1992) to 6.9% (2003).
- Of records with improbable GA, 72% were due to missing data in 1992, compared with 92% in 2004.
- After CHS began training in 2005 for selected hospitals, the statewide number of records with improbable GA dropped 64% compared with 2004, and number of cases fully missing gestational age dropped 80%.

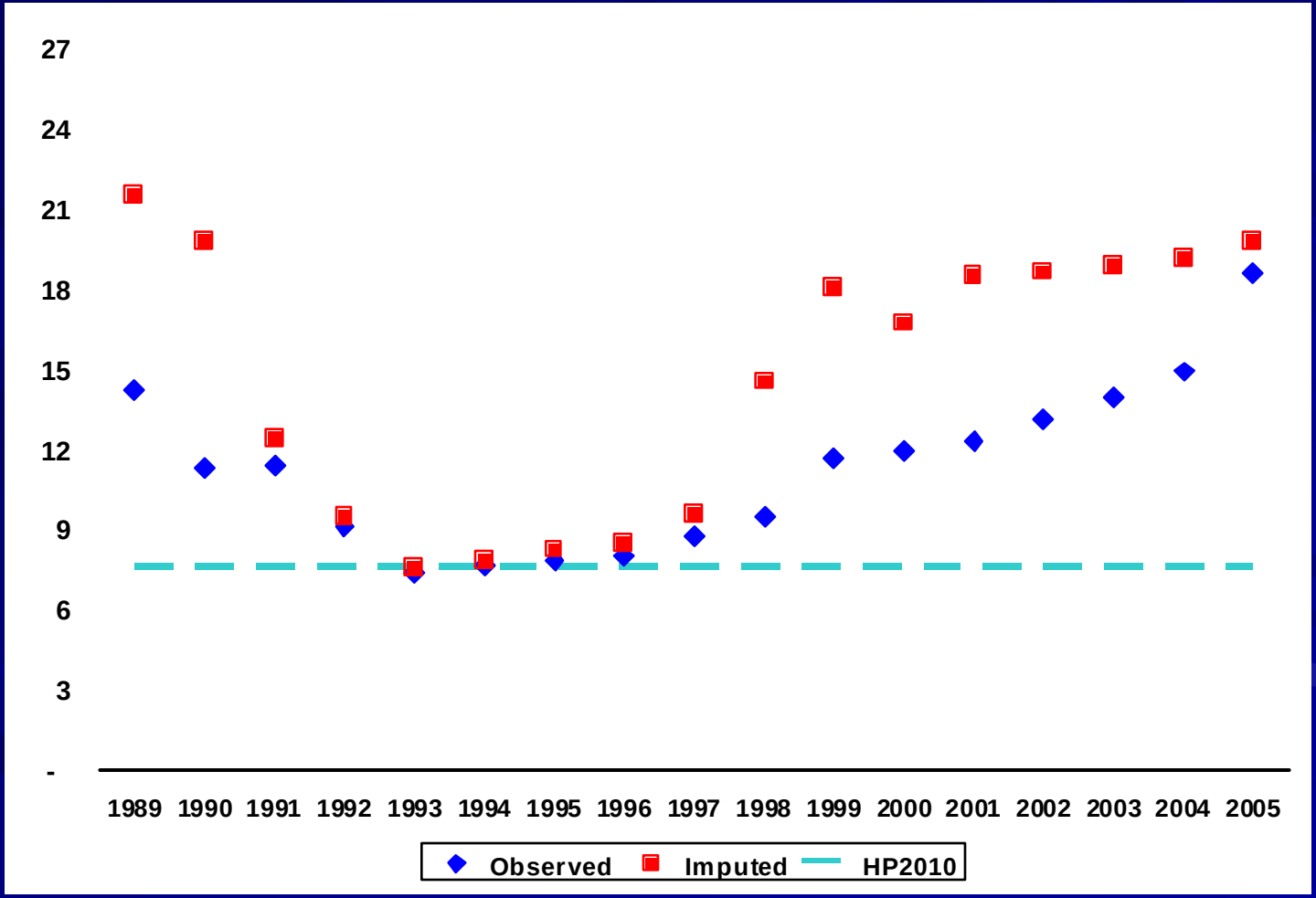
County Variations

- The 1992 county range for improbable GA was 0.0%-13%, median 3.2%. The 2003 range before training was 0.0%-17.7%, median 5.3%.
- In 1992, most counties with data quality problems were rural. In 2003, more counties had data quality problems and most were larger.
- In 1992, only 5 counties with 5,000 or more births had more than 5% of records with improbable data. Of 20 counties with 5,000 or more births in 2003, 13 had improbable GA above 5%.

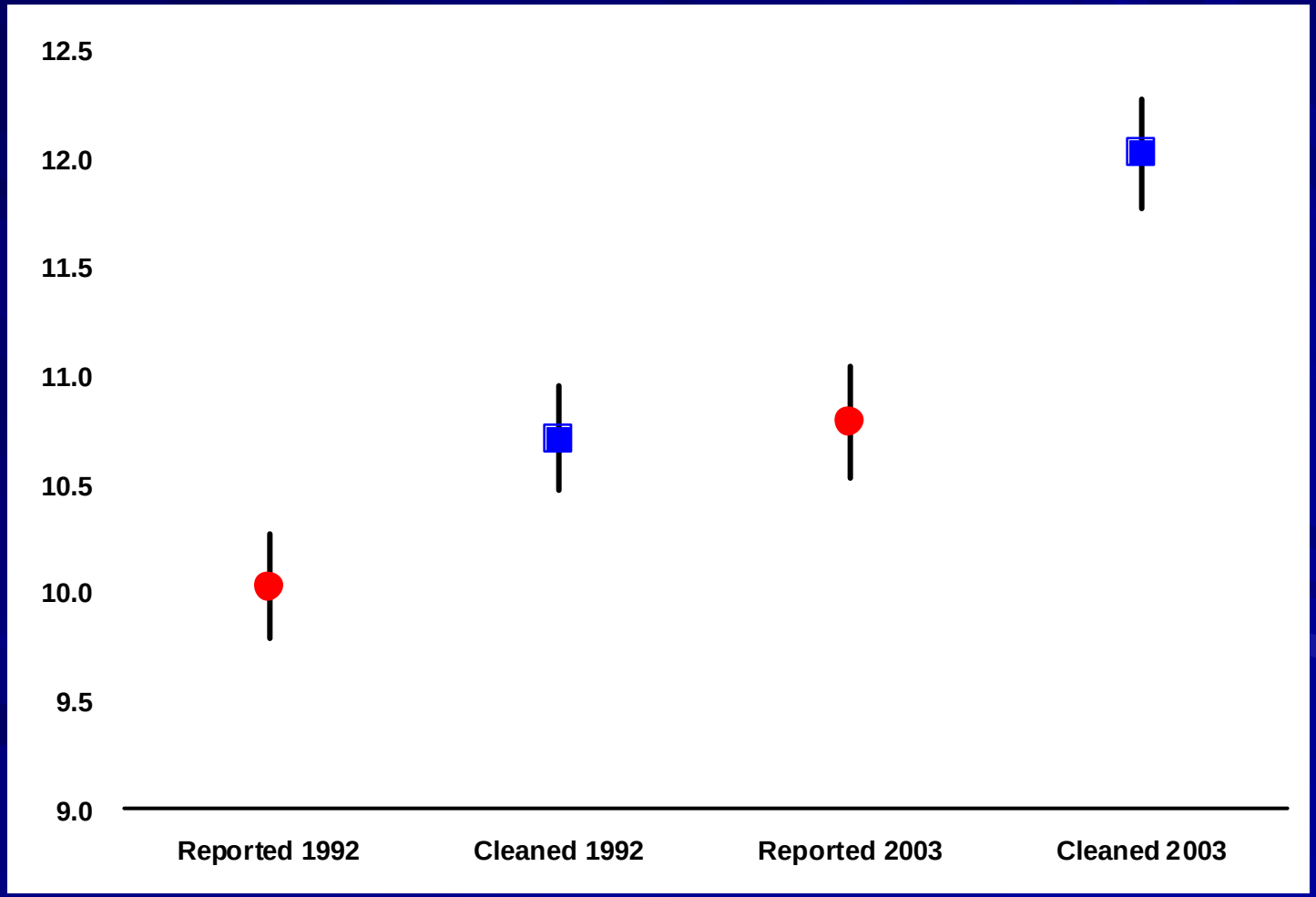
Comparison of Observed and Imputed County Preterm Birth Rates 1989-2005



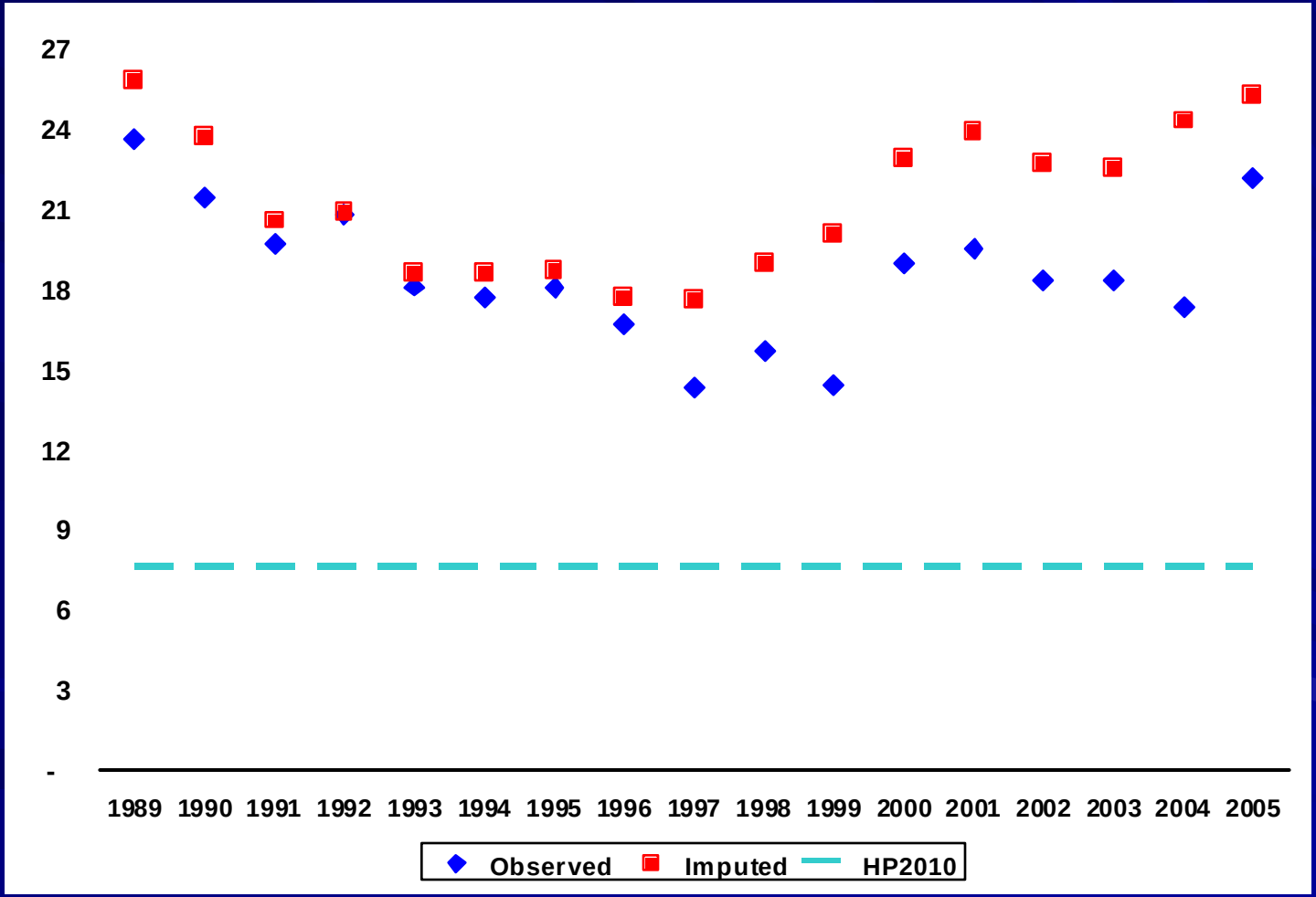
Trend Analysis of Asian Preterm Birth Rates Observed and Imputed 1989-2005



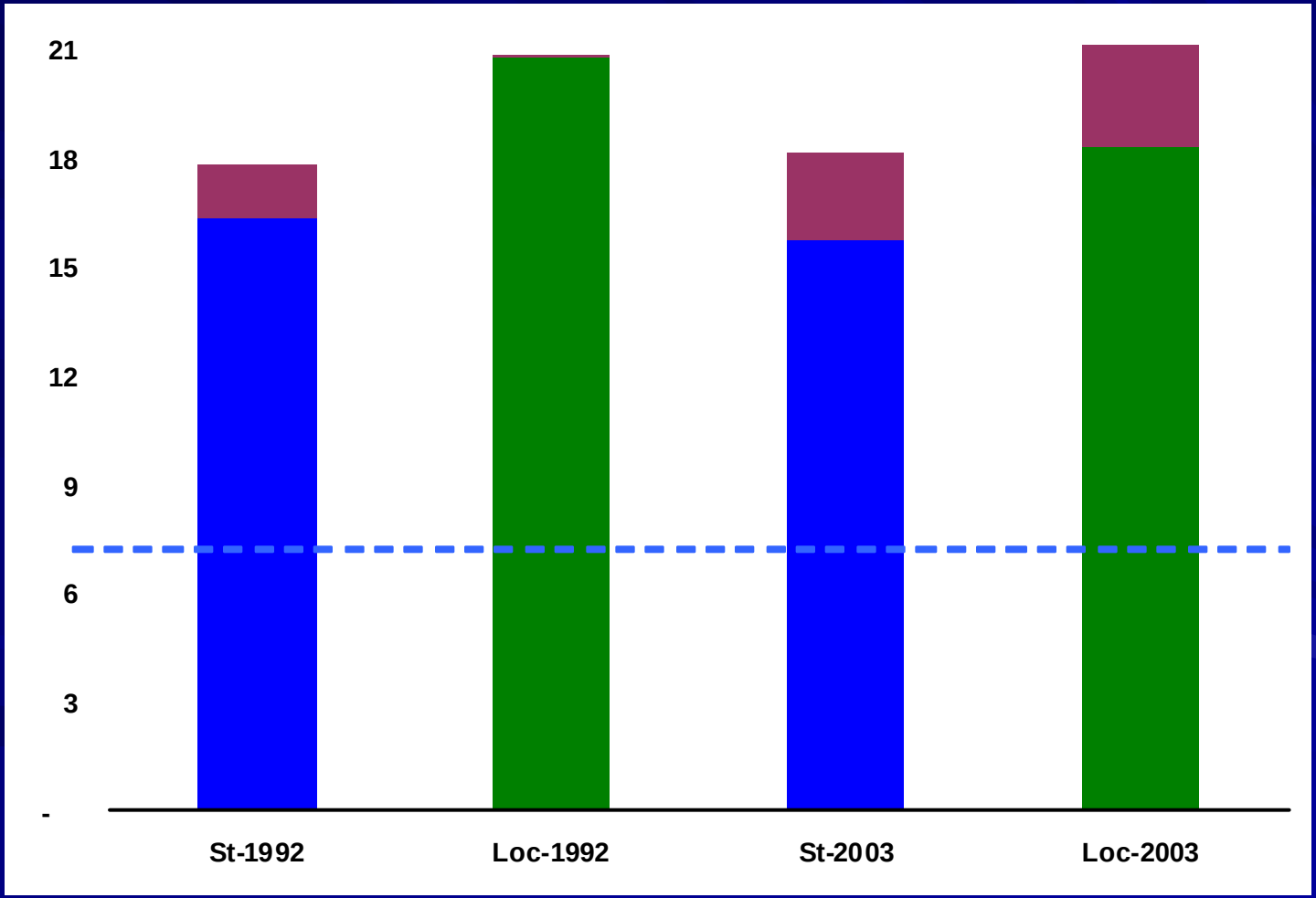
Impact of Reported and Imputed Asian Preterm Birth Rates 1992 and 2003



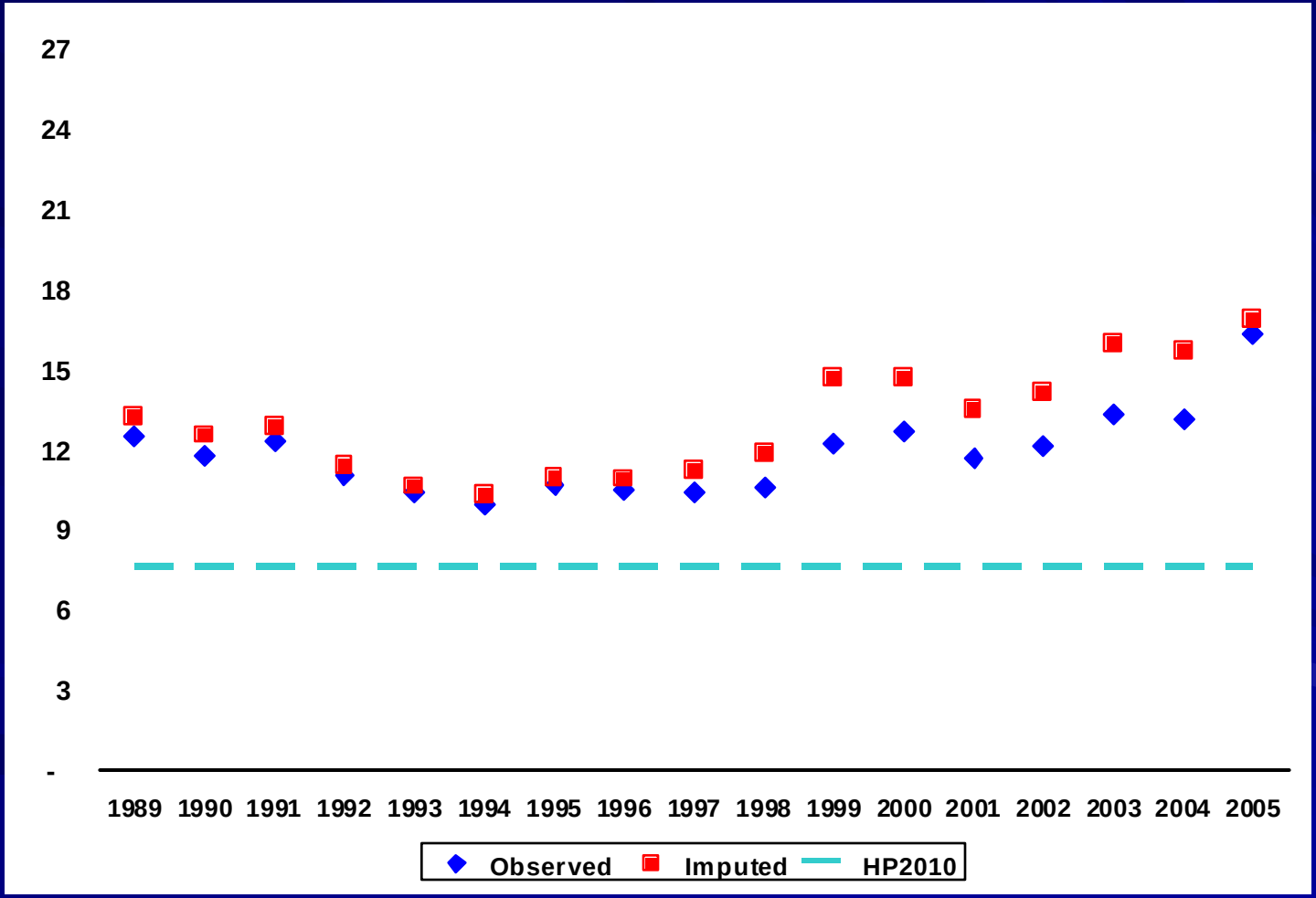
Trends in Local Black Preterm Birth Rates Observed and Imputed 1989-2005



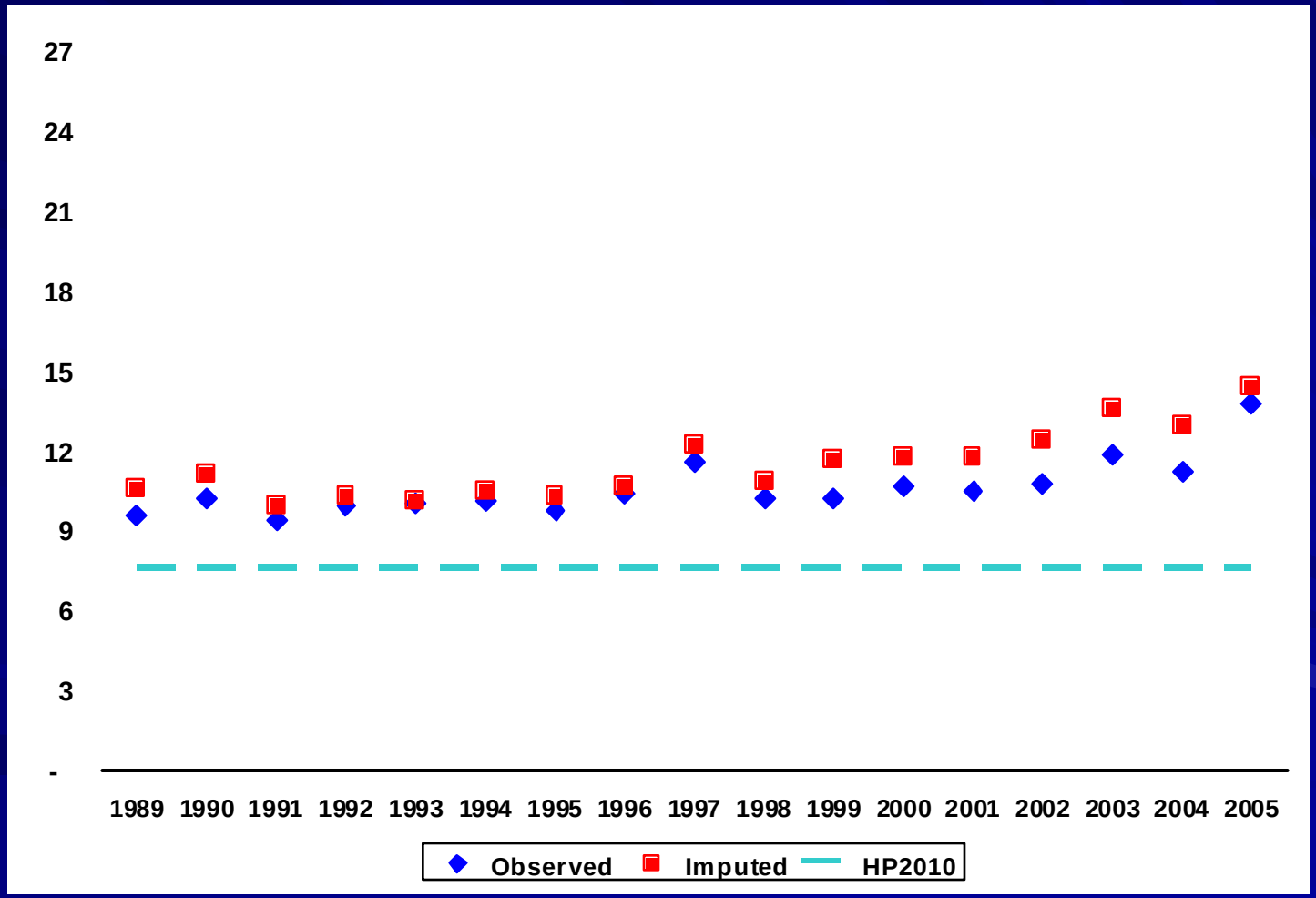
Impact of Edits on Black Preterm Rates for State and Local 1992 and 2003



Local Hispanic Preterm Birth Rates Reported and Imputed 1989-2005



Local White Preterm Birth Rates Reported and Imputed 1989-2005



Conclusions

- Data quality for preterm birth rates varies by race/ethnicity, within and across counties, and over time.
- The shift of poor quality data from smaller to more populous counties has an increasing impact on the accuracy of state rates.
- Data quality issues result in significant under-estimates of California's preterm birth rates and erroneous comparisons with standards such as the HP 2010
- Before concluding that population-based rates are changing, it is important to evaluate and understand the impact of data quality

Political and Policy Implications

- Laws that mandate use of unedited data impact the accuracy and utility of health indicators calculated from those data
- Indicator values based on poor quality unedited data may lead to inaccurate assessments of policies and programs directed to alleviate a health problem
- Racial and ethnic disparities in data quality may result in underestimates of health disparities particularly in Black and Asian populations.

For Further Information:

- Linda Remy, MSW, PhD
  - Email: lremy@well.com
- Gerry Oliva, MD, MPH
  - Email: olivag@fcm.ucsf.edu
- Website: www.ucsf.edu/fhop

Improbable Values Birthweight

Year	Births	Improbable Birthweight (Grams)				
		Missing	Lt 250	Gt 4999	Total	Pct
1992	600,838	104	57	1,325	1,486	0.25
1993	584,483	85	82	1,281	1,448	0.25
1994	567,034	78	64	1,135	1,277	0.23
...						
2001	527,371	5	70	965	1,040	0.20
2002	529,241	7	77	929	1,013	0.19
2003	540,827	13	82	944	1,039	0.19

Improbable birthweight distributed unevenly

- In 1992, county-level improbable BWT values ranged from 0.0%-2.4%, median 0.27%.
- The 2003 range was 0.0%-0.9%, median 0.2%. The median was little changed, reflecting that jurisdictions tended to improve data quality on this measure with time.
- Several reabstraction studies have found BWT is one of the most reliably coded variables. Unlikely values are not a significant factor in calculating low birthweight rates.

Utility of Data Quality Reports

- Data quality reports help counties:
  - assess the potential impact of data errors on the accuracy of indicators
  - give health department staff information to work with providers and hospitals to improve data quality.
- For rural counties, a few missing or unlikely values can result in misleading conclusions about the quality and adequacy of prenatal care or the effectiveness of outreach.
- Using the statewide average to gauge problems may not be helpful for a state as large and diverse as California.

Between 1992 and 2003:

- The number of births to California residents dropped 11.1% from 600,838 to 540,827.
- Improbable BWT (missing, less than 250 grams and more than 4999 grams) dropped from 1,486 to 1,039 births.
- Most of the decrease was associated with BWT greater than 4999 grams.
- In 1992, improbable BWT represented 0.25% of records. In 2003, this was 0.19%.

Implications

- Local jurisdictions must carefully review their data quality reports for a given indicator, both to understand the impact of quality data on results in a given year or trends over time.
- FHOP has prepared a new spreadsheet presenting annual preterm birth data quality. It is available on the website.
- Local jurisdictions are advised to consult that spreadsheet before reporting their preterm birth rates.

California Preterm Birth Rates Reported and Imputed 1989-2005