Accreditation of Medi-Cal, Healthy Kids and Healthy Families Program.

Balancing Anonymity and Analytic Value in Surveys of Patients in Vulnerable Populations: The Mathematics and Logic of Selecting and Coding Variables to Control Risk While Maximizing Information From CAHPS Surveys in a Large Urban Medicaid Health Plan, 2008-2014

Session: 5034.0 Statistical Methods and Applications in Community Health
Section: Applied Public Health Statistics
Topic: CAHPS Surveys of Patients

November 19, 2014

S. Rae Starr
Healthcare Outcomes & Analysis
L.A. Care Health Plan, Los Angeles CA

L.A. Care HEALTH PLAN

Presenter Disclosures

S. Rae Starr

The following personal financial relationships with commercial interests relevant to this presentation existed during the past 12 months:

I am employed as a statistician at L.A. Care Health Plan –
the Local Initiative Health Authority of Los Angeles County, California.

L.A. Care is a public entity competing with commercial insurers in the Medicaid and S-CHIP markets in L.A. County.

Notes:
CAHPS® is a registered trade name of the Agency for Healthcare Research and Quality (AHRQ).
HEDIS® is a registered trade name of the National Committee for Quality Assurance (NCQA).

Outline

I. Learning Objectives.

II. Background: CAHPS Surveys -- Anonymity or Actionability: Nature of the Health Survey Dilemma.

III. Terminology, Techniques, and Tools for Assessing Re-Identification Risk in Release of a Limited Data Set.

IV. Example of Risk Assessment For a Simple Dataset.

V. Real World Example: CAHPS 2014 Sampling Frame Files.

VI. Discussion: Learning Objectives.

VII. Avenues For Further Research On This Topic.

I. Learning Objectives

1. Explain the importance of anonymity in surveys of populations particularly vulnerable to disclosure.
2. Describe the tradeoffs between anonymity and applied value to make health care surveys actionable.
3. Discuss analytic implications imposed by common rules for protecting anonymity.
4. Describe specific risks from releasing various demographics in response data from surveys.
5. Discuss degrees of anonymity and their appropriate application.
6. Explain the calculation of risks to anonymity in common demographics.
7. Identify factors that play into the calculation of risk against a predetermined threshold.
8. Describe best (and worst) practices for using anonymous and non-anonymous data from vulnerable populations.
9. Discuss how to adapt from a one-shot survey context into a multi-year data strategy for quality improvement analyses.

II. CAHPS Surveys of Health Care Service Quality

The Consumer Assessment of Healthcare Providers and Systems (CAHPS) is a survey of patient experience with health care services, and includes patients' ratings about the quality of services.

The CAHPS family of surveys is developed and maintained under the auspices of the Agency for Healthcare Research and Quality (AHRQ).

Domains measured: Health Plan CAHPS collects patients' ratings of health care, health plans, PCPs, and specialists; and collects composite measures on access to services, speed of access; provider communication; customer service; coordination of care; health promotion, and shared decision-making, etc.

End uses: Surveys within the CAHPS family of surveys are commissioned by health plans and public agencies for regulatory oversight; accreditation; and comparison by health care consumers.

- CMS uses a variant of CAHPS to survey Medicare members. CAHPS measures are used in calculating Star Ratings, which affect *reimbursement payments to health plans; sanctions; and even termination of contracts.*
- The National Committee for Quality Assurance (NCQA) uses a variant of CAHPS. The *scores determine 13% of a health plan's Accreditation score*.
- *Ratings are published* for use by payers and consumers in picking health plans.

Context: Use of CAHPS for CQI at L.A. Care Health Plan

- CAHPS provides a lens for viewing service quality as experienced by patients in a large county with a complex provider network (n=3,000 PCPs, 7,000 specialists).
- Large, diverse membership in Los Angeles, California.
- Status at the start of the period covered in this briefing (2009 forward):
  - Mostly Medicaid, urban, **2/3rd** pediatric, often Spanish-speaking.
  - Roughly **21%** of Medicaid managed care population in California.
  - Roughly **2.1%** of Medicaid managed care population in the U.S.
  - Roughly **1-in-14** L.A. County residents is an L.A. Care member.
  - Mostly Medicaid, some S-CHIP, SNP, and special programs.
  - Serving **10** distinct language concentrations ("threshold languages"): Spanish, English, Armenian, Korean, Cambodian, Chinese, Russian, Vietnamese, Farsi, Tagalog.
  - Mostly urban and suburban; 1 semi-rural region in the high desert.

Why More Penetrating CAHPS Analysis Matters

A separate purpose of the surveys, is to guide continuous quality improvement (CQI) activities in health plans and agencies.

CAHPS surveys used by agencies are typically anonymous, to reduce bias in patients' responses due to fear of retaliation for giving ratings.

Anonymity rules balance (well or poorly) between risk of bias and analytic value. (Paper in APHA 2013 Ethics SPIG addressed policy pros and cons.)

The manner in which anonymity is defined and implemented can impair analysis for quality improvement. Usefulness is mediated by several factors:

- *Credible causal analysis* requires variables and categories that are fine-grained enough to cleanly distinguish causes.
- *Drilldowns to affected sub-populations* must be fine-grained enough to be operationally useful to departments owning touch-points with patients, doctors.
- Entities being rated (doctors, clinics, etc.) are *not necessarily predisposed to trust that patients' opinions* about services are accurate or clinically relevant.
  - To be adequately compelling in a low-scoring clinical environment, CAHPS analyses must generally be tightly defined to make conclusions inescapable.
  - But categories that are sufficiently aggregated to protect patient anonymity, are often viewed as too general to be "actionable".

Impact of Overprotection on Analytic Rigor in Quality Improvement

Types of causal analysis that are harmed or limited under common methods for protecting anonymity:

- Analysis of causes: Indicators of whether or not a patient has a given *condition or set of conditions.*
- Evaluation of programs: Indicators of the effectiveness of giving a patient a particular *treatment* or set of *interventions and programs.*
- The feasibility and precision of multivariate analysis is hampered:
  - By outright prohibitions due to cell-size rules.
  - By coarse coding and aggregating continuous variables (age, months of coverage, number of visits, height and weight in fractional form);
  - By coarse groupings of categorical variables (ratings, demographics: ethnicity, language, geographical region, clinics and provider groups.

Cumulative nature both of risks and restrictions:

- CAHPS surveys examine numerous topics, each of which may require different analytic variables of various types.
  - The risk to anonymity is cumulative in a given survey dataset.
  - Adding one variable eventually means discarding another.
  - Choices about which variables to include *today,* prevent adding new variables to address *tomorrow's* health problems.

III. Terminology, Tools, Techniques: Assessing Re-Identification Risk

Definitions:

- Direct identifiers: Variables that uniquely identify individual cases.
- Concepts: Anonymity, uniqueness, identification, de-identif., risk of re-identif.
- Quasi-identifiers: Variables (often demographics) that can uniquely identify an individual (usually in combo with other vars in the limited dataset or in commonly avail. datasets).
- Thresholds: For *ex ante* anonymizing of CAHPS survey datasets released to health plans, k=10 is a common threshold.

Defining risk:

- The literature provide a good grammar for discussing risk and protection:
  - **k-anonymity:** each combo of quasi-identifiers has at least k unique cases
  - **ℓ-diversity:** sensitive values are each well-represented (not sparse)
  - **t-closeness:** distribution of sensitive attributes in each quasi-identifier category should be close to distribution in population.
  - …
- Less grammar exists for utility, or "opportunity cost" of each disclosed category:
  - Conceptually similar to a degree-of-freedom problem.
  - Degrees of freedom: "The number of values in a study that are free to vary." http://www.investopedia.com/terms/d/degrees-of-freedom.asp
  - "The number of independent ways by which a dynamic system can move without violating any constraint imposed on it[.]" http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)

Techniques For Anonymizing Datasets

Non-deterministic methods (employ randomness):

- Post-randomization (PRAM): Changing ("perturbing") some values of a categorical variable into other categories, per pre-defined probabilities in a transition matrix.
- Adding noise: Adding random noise to the values of a continuous variable to prevent matching against external datasets.
- Data swapping: Masking data by exchanging values of confidential variables, between records, without modifying the original values.
- Data shuffling: A hybrid approach blending data swapping and stochastic perturbation by way of ranks. (Sarathy and Muralidhar, 2011: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/08_Sarathy-Muralidhar.pdf)

Adapted largely from Templ, Meindl, Kowarik (2014): http://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf, p. 14. Data used for screenshots in this briefing, are populated from L.A. Care examples. The images are from sdcMicroGUI and sdcMicro, two modules distributed for use with the R statistical package. Any proprietary rights to the techniques and graphical layout of output from the software, are gratefully acknowledged as copyrighted to and by Templ, Meindl, Kowarik, above, under terms of R distribution. The examples in slides 12 to 21 are my own. Any flaws in execution or interpretation are mine, not the software authors'.

Techniques For Anonymizing Data (Cont.)

Measuring <u>information loss</u> and impact on data utility:

- Methods for continuous variables are based on analyzing differences between original values and the perturbed values, either directly, or by comparison of co-variances from the original and perturbed variables.
- Although statistical comparisons are suggested, comparison against benchmarks is most strongly recommended.

Quantitative implications: **The methods of anonymization vary in the degree and type of their effects on the utility (usefulness) of the protected data in applied analysis.**

- Each method can harm statistical properties of the data for modeling and analysis.
- Adding uncorrelated noise, for example, can preserve the mean, while corrupting variances and correlation coefficients. Adding correlated noise avoids the latter.
- Likewise, censoring or top-coding of data can make OLS regression estimates biased and inconsistent.

Tools for Calculating Re-Identification Risk

Software tools exist to aid in calculating risk and choosing a strategy to reduce information loss in the data protection process.

- The same tools used to de-identify data, can help end users determine which combination of techniques best preserve data utility.

Formal methods:

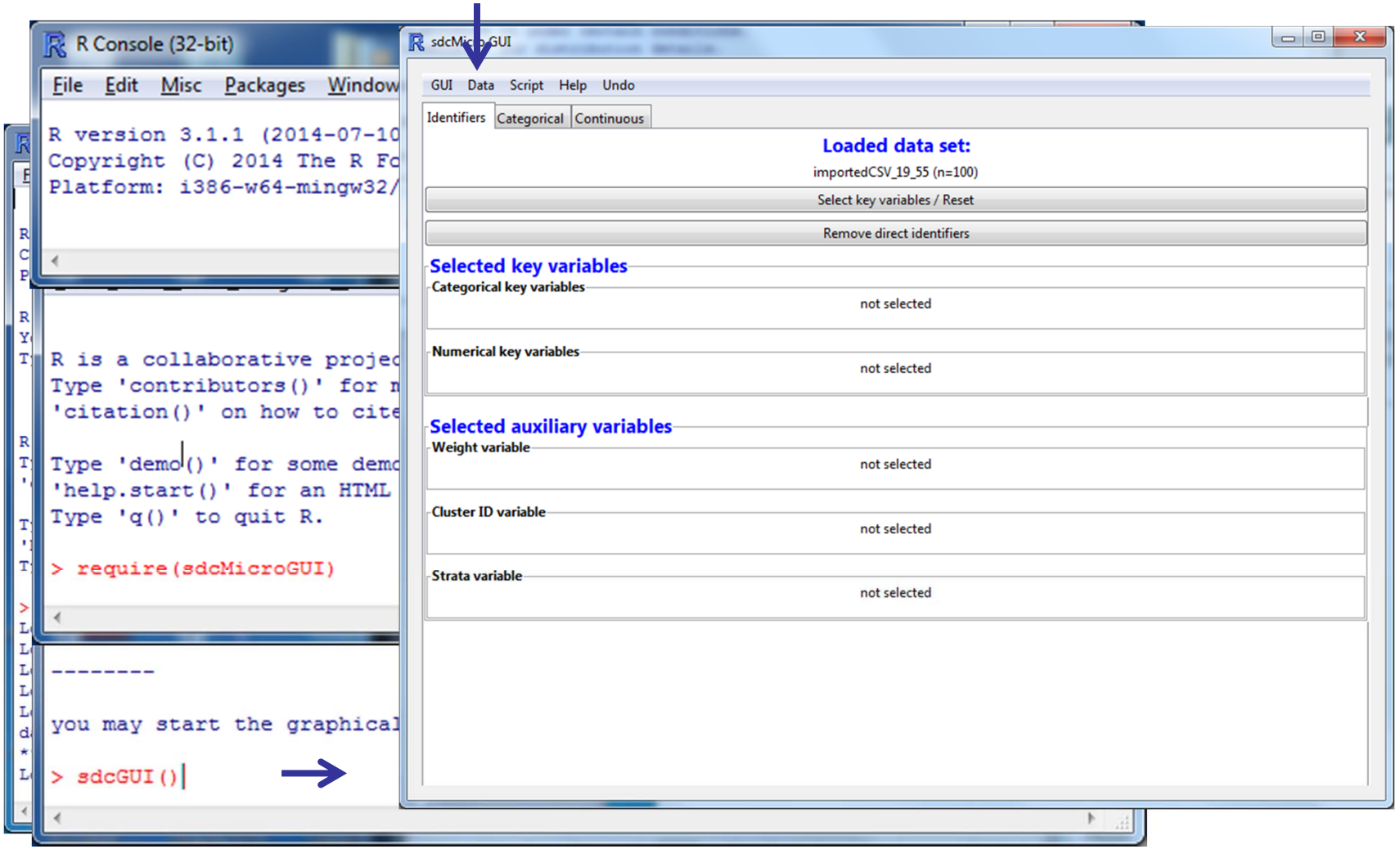
- Data Intrusion Simulation (DIS).
- Special Uniques Detection Algorithm (SUDA and SUDA2).
- Templ *et al* 2014 provide an elegant summary: http://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf.

Non-commercial software packages:

- DAS software is available from the Federal Committee on Statistical Methodology (FCSM): Runs under a popular licensed commercial statistical package, but requires particular modules:
  http://fcsm.sites.usa.gov/committees/cdac/cdac-resources/
  Use "Contact Us" link and request the 17mb Zip file via email.
- sdcMicro and sdcMicroGUI run under the "R" statistical package and are downloadable at http://cran.r-project.org/web/packages/sdcMicro/index.html. Runs both DIS and SUDA risk calculation methods.

Note: This research did not canvass or review anonymization software alternatives. Mention of the software above, or use of exploratory examples in the briefing, is not offered as a review or endorsement or promotion of any particular software package or product.

IV. Example of Risk Assessment For a Simple Dataset

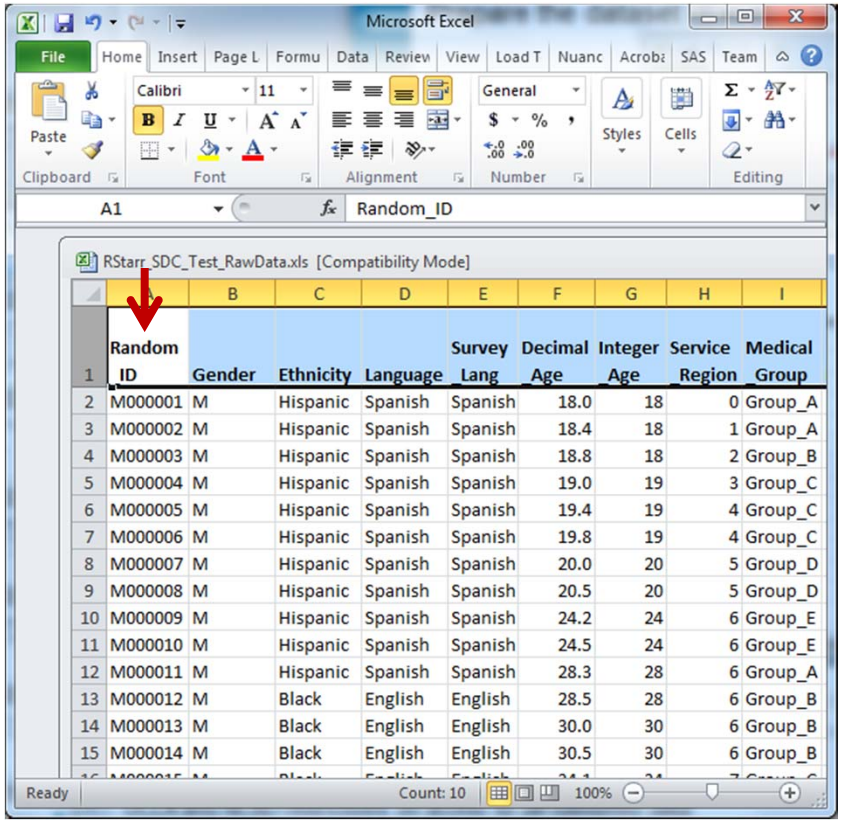


Source: M. Templ, B. Meindl, A Kowarik, "IHSN GUI Tutorial for sdcMicroGUI (and sdcMicro)", Nov. 2013, program documentation, http://www.data-analysis.at .

Example (Cont.): Data in CSV File

Prepare the dataset (Excel -> CSV, commercial stat packages, etc.).

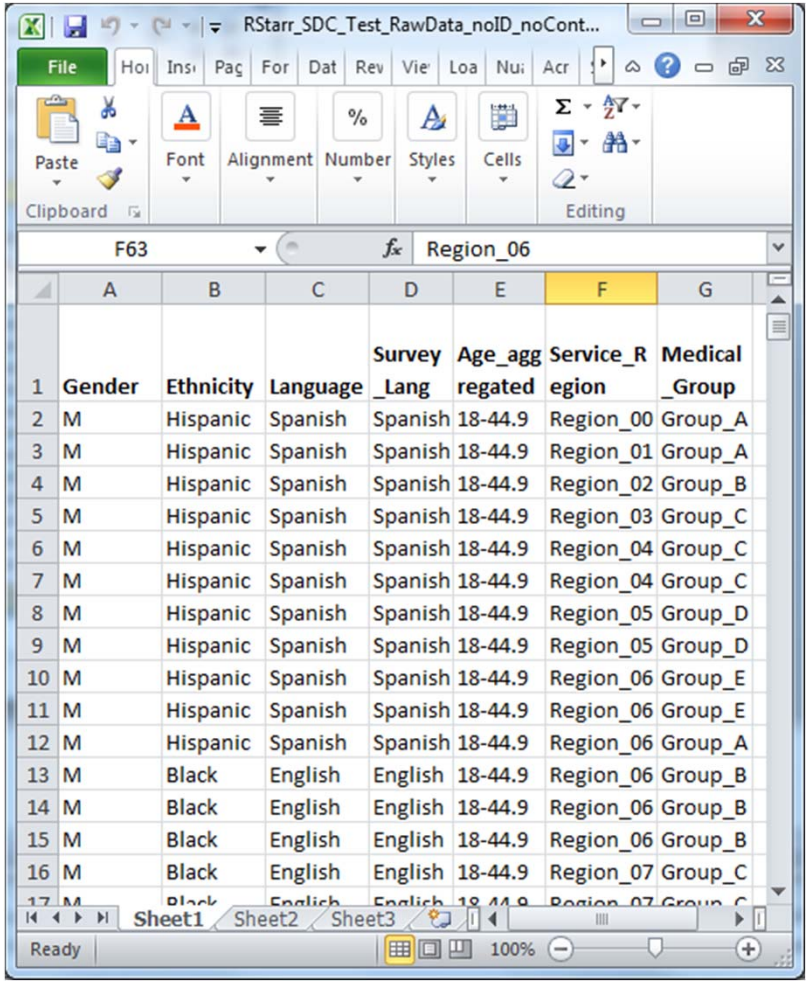
- View this contrived n=100 example as a small rural clinic or public agency doing a patient experience survey.
- Main threats for re-identification are the Random_ID, the continuous numeric values, and the sparse service region variable.



	A	B	C	D	E	F	G	H	I
1	Random ID	Gender	Ethnicity	Language	Survey_Lang	Decimal_Age	Integer_Age	Service_Region	Medical_Group
2	M000001	M	Hispanic	Spanish	Spanish	18.0	18	0	Group_A
3	M000002	M	Hispanic	Spanish	Spanish	18.4	18	1	Group_A
4	M000003	M	Hispanic	Spanish	Spanish	18.8	18	2	Group_B
5	M000004	M	Hispanic	Spanish	Spanish	19.0	19	3	Group_C
6	M000005	M	Hispanic	Spanish	Spanish	19.4	19	4	Group_C
7	M000006	M	Hispanic	Spanish	Spanish	19.8	19	4	Group_C
8	M000007	M	Hispanic	Spanish	Spanish	20.0	20	5	Group_D
9	M000008	M	Hispanic	Spanish	Spanish	20.5	20	5	Group_D
10	M000009	M	Hispanic	Spanish	Spanish	24.2	24	6	Group_E
11	M000010	M	Hispanic	Spanish	Spanish	24.5	24	6	Group_E
12	M000011	M	Hispanic	Spanish	Spanish	28.3	28	6	Group_A
13	M000012	M	Black	English	English	28.5	28	6	Group_B
14	M000013	M	Black	English	English	30.0	30	6	Group_B
15	M000014	M	Black	English	English	30.5	30	6	Group_B

Example (Cont.): Transform the Data -- Aggregation

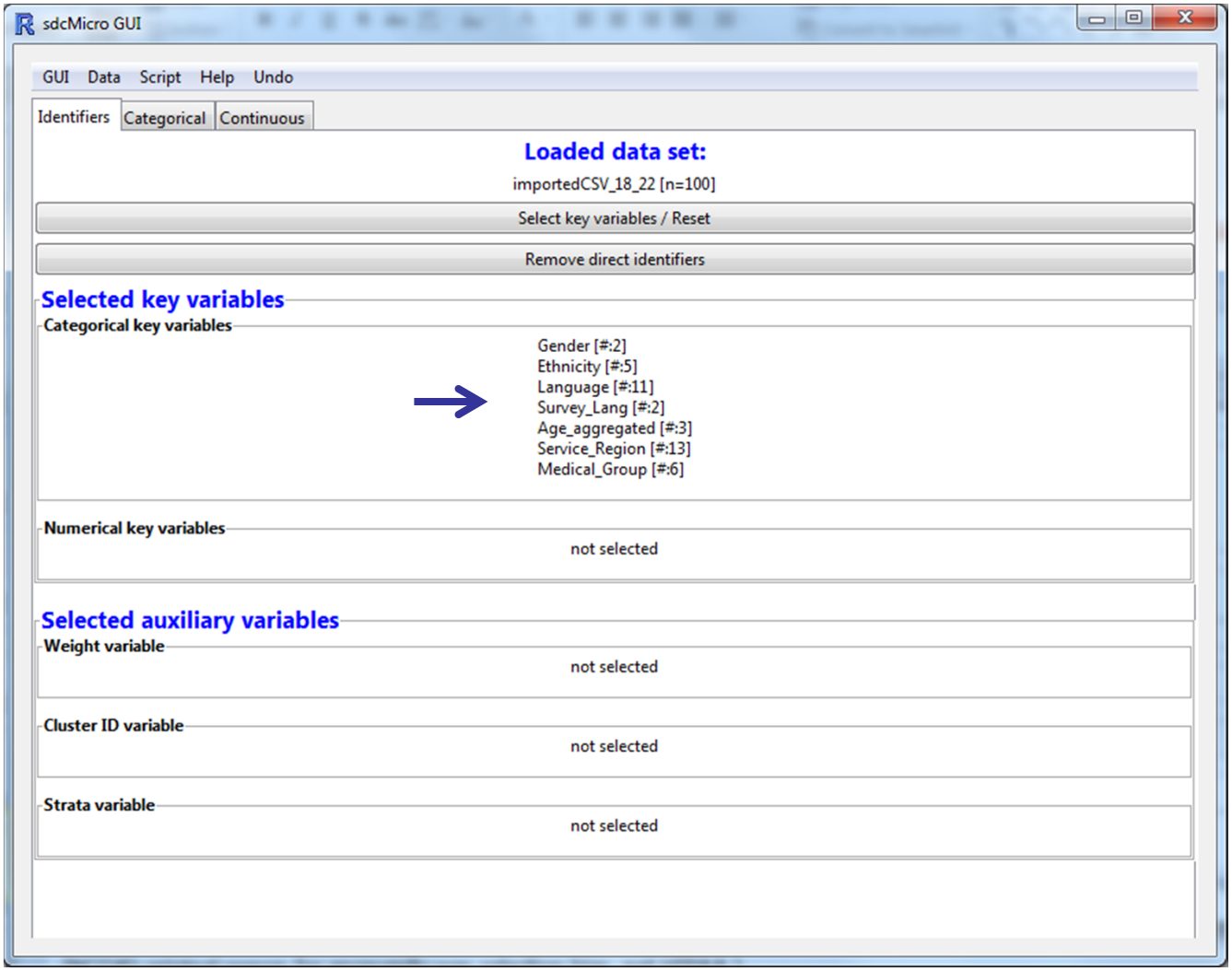
Can address these threats using the software tool, or can fix in the spreadsheet or in R or other statistical package at user's option:



	A	B	C	D	E	F	G
1	Gender	Ethnicity	Language	Survey_Lang	Age_aggregated	Service_Region	Medical_Group
2	M	Hispanic	Spanish	Spanish	18-44.9	Region_00	Group_A
3	M	Hispanic	Spanish	Spanish	18-44.9	Region_01	Group_A
4	M	Hispanic	Spanish	Spanish	18-44.9	Region_02	Group_B
5	M	Hispanic	Spanish	Spanish	18-44.9	Region_03	Group_C
6	M	Hispanic	Spanish	Spanish	18-44.9	Region_04	Group_C
7	M	Hispanic	Spanish	Spanish	18-44.9	Region_04	Group_C
8	M	Hispanic	Spanish	Spanish	18-44.9	Region_05	Group_D
9	M	Hispanic	Spanish	Spanish	18-44.9	Region_05	Group_D
10	M	Hispanic	Spanish	Spanish	18-44.9	Region_06	Group_E
11	M	Hispanic	Spanish	Spanish	18-44.9	Region_06	Group_E
12	M	Hispanic	Spanish	Spanish	18-44.9	Region_06	Group_A
13	M	Black	English	English	18-44.9	Region_06	Group_B
14	M	Black	English	English	18-44.9	Region_06	Group_B
15	M	Black	English	English	18-44.9	Region_06	Group_B
16	M	Black	English	English	18-44.9	Region_07	Group_C

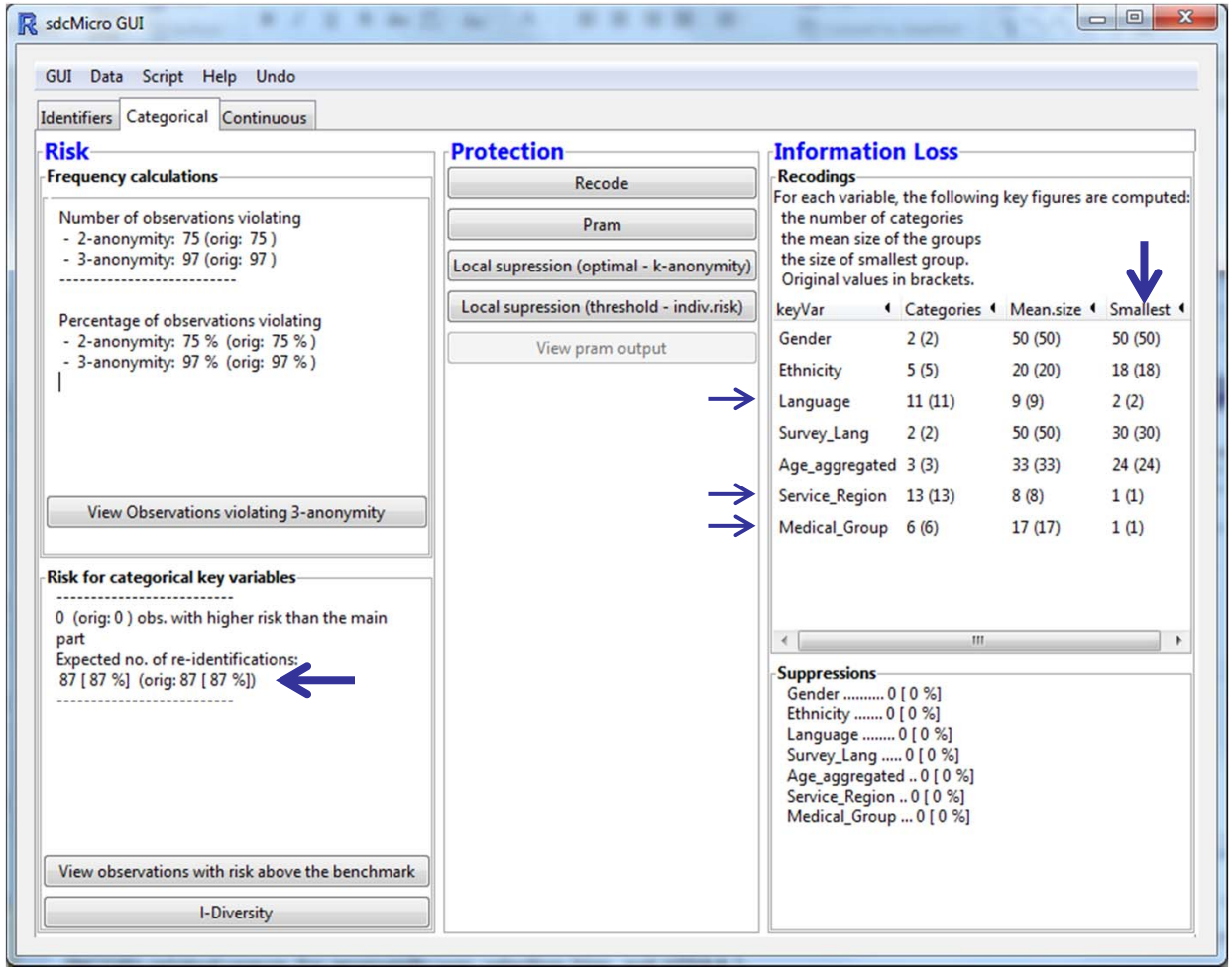
Example (Cont.): Load the Data Into the Risk Evaluation Software

Demographic categories within a notional sampling frame file (n=100):



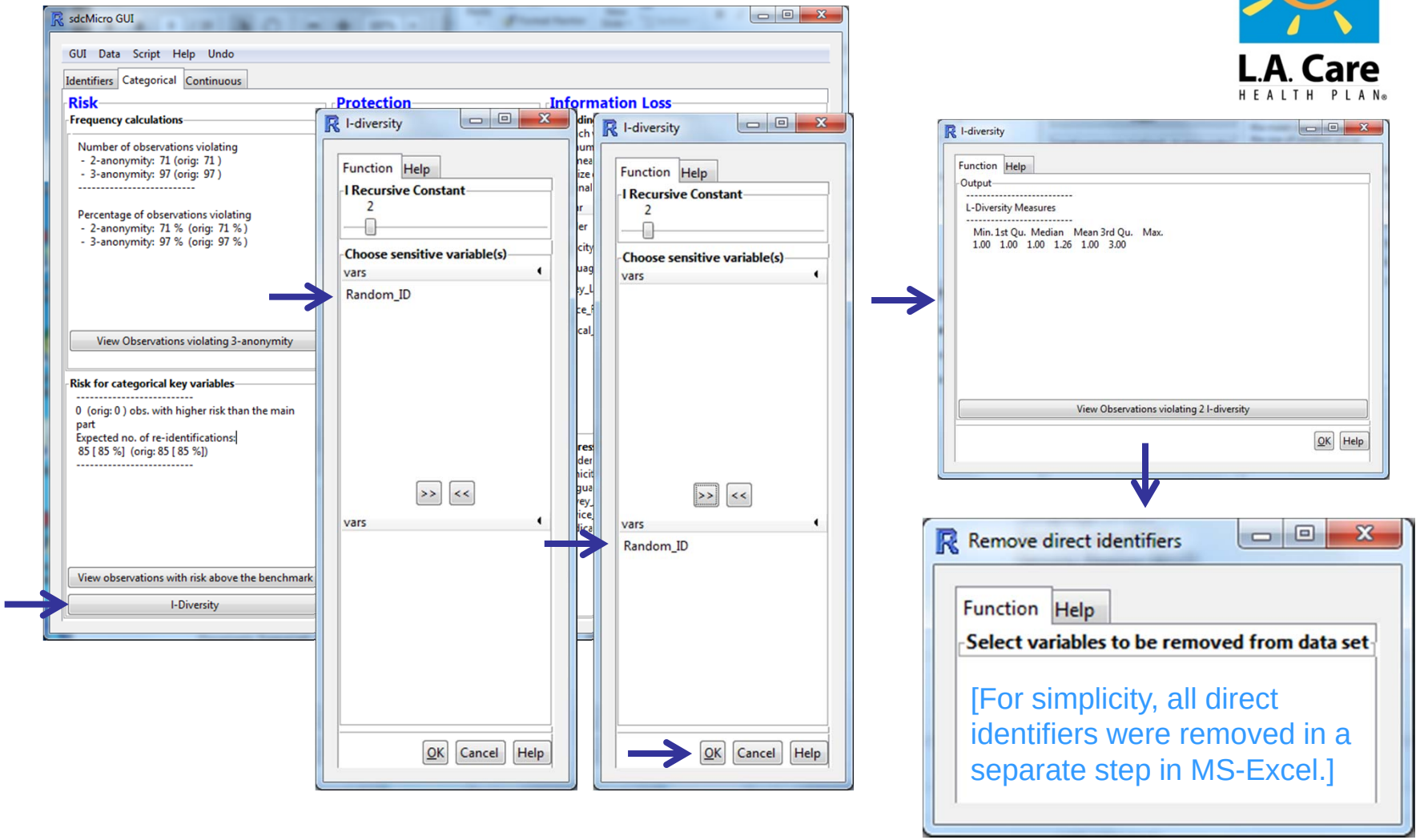
Example (Cont.): Get Initial Re-Identification Risk Calculation

Initial statistics show high risk of re-identification (87.0%), and illustrate which variables have sparse categories in this notional dataset:



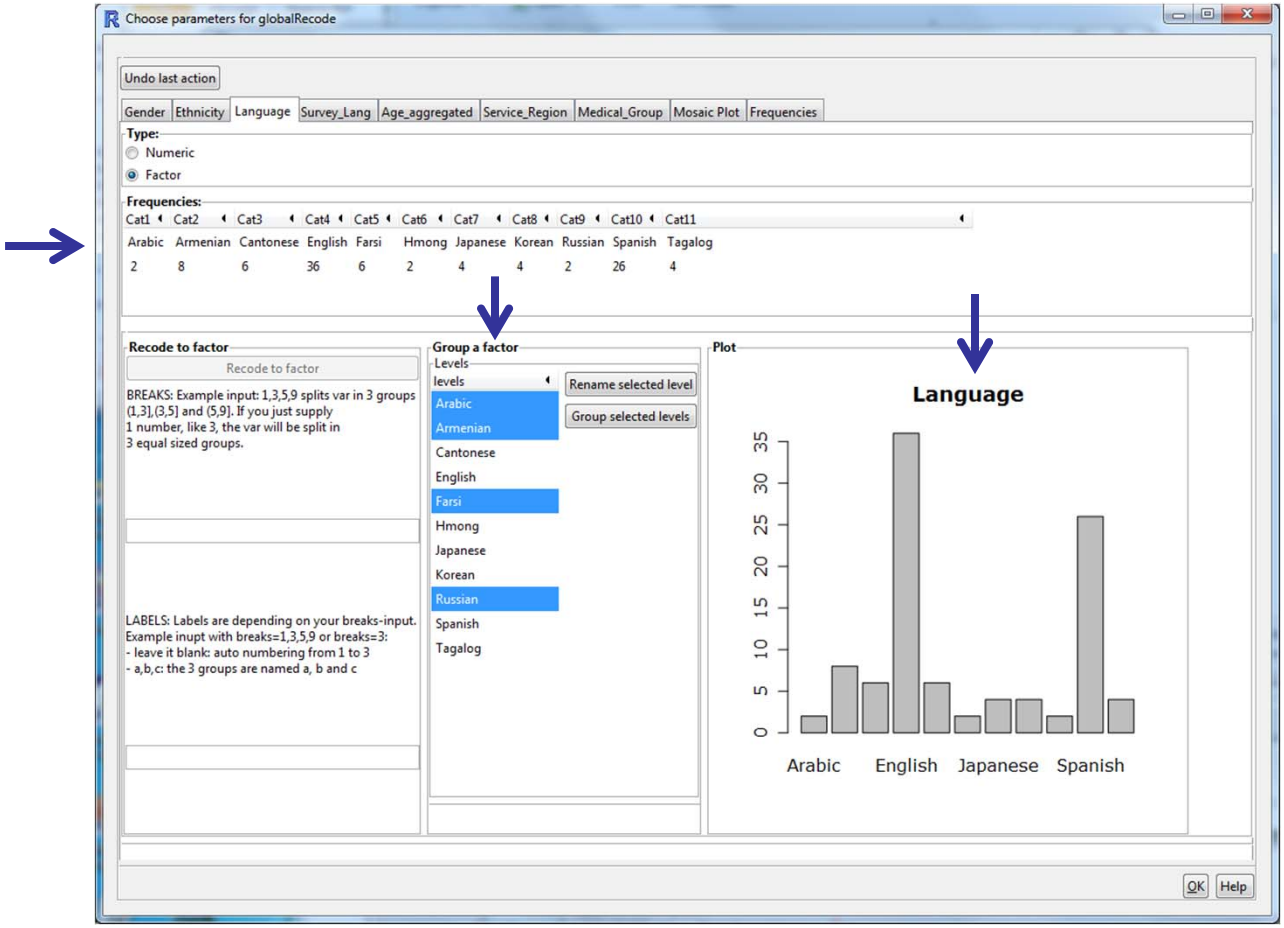
Example (Cont.): Remove Direct Identifiers

Begin modifying the dataset: Remove direct identifiers.



Example (Cont.): Aggregate Sparse Categories

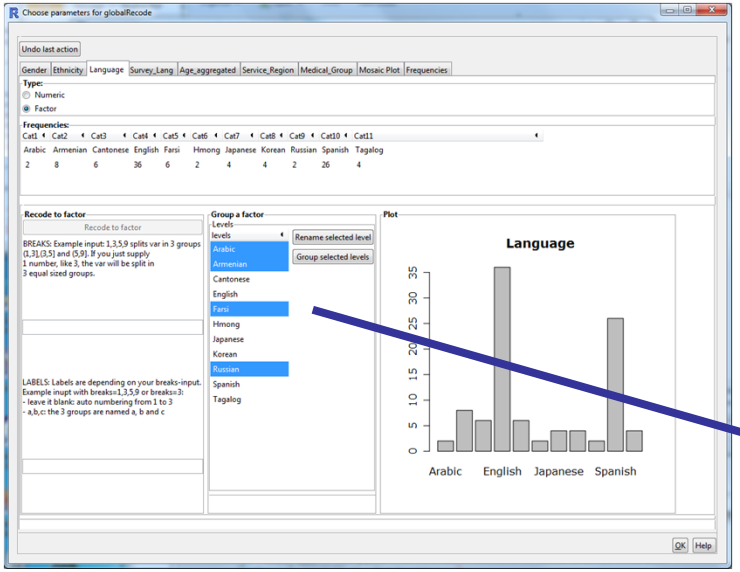
Begin modifying dataset – recode/aggregate sparse categories:

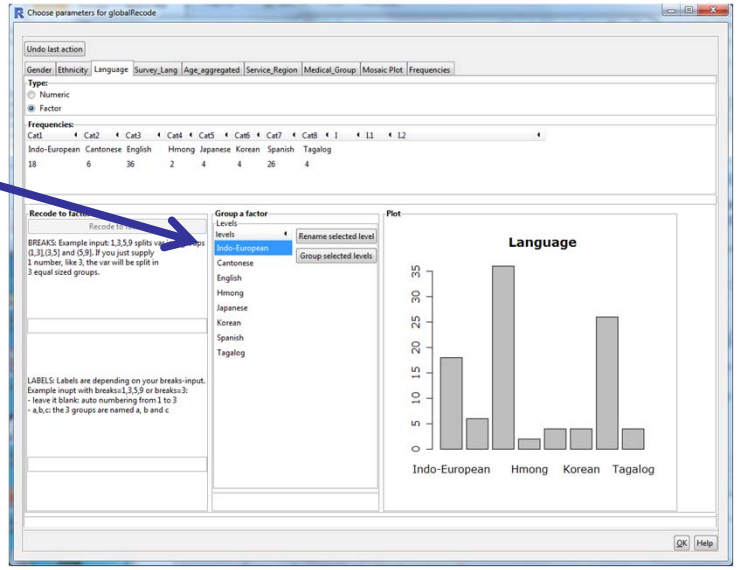


For large and complex datasets, data operations can be efficiently done in a statistical package prior to loading the data into the risk assessment software

Example (Cont.): Aggregate Language

Begin modifying dataset – recode/aggregate sparse categories:

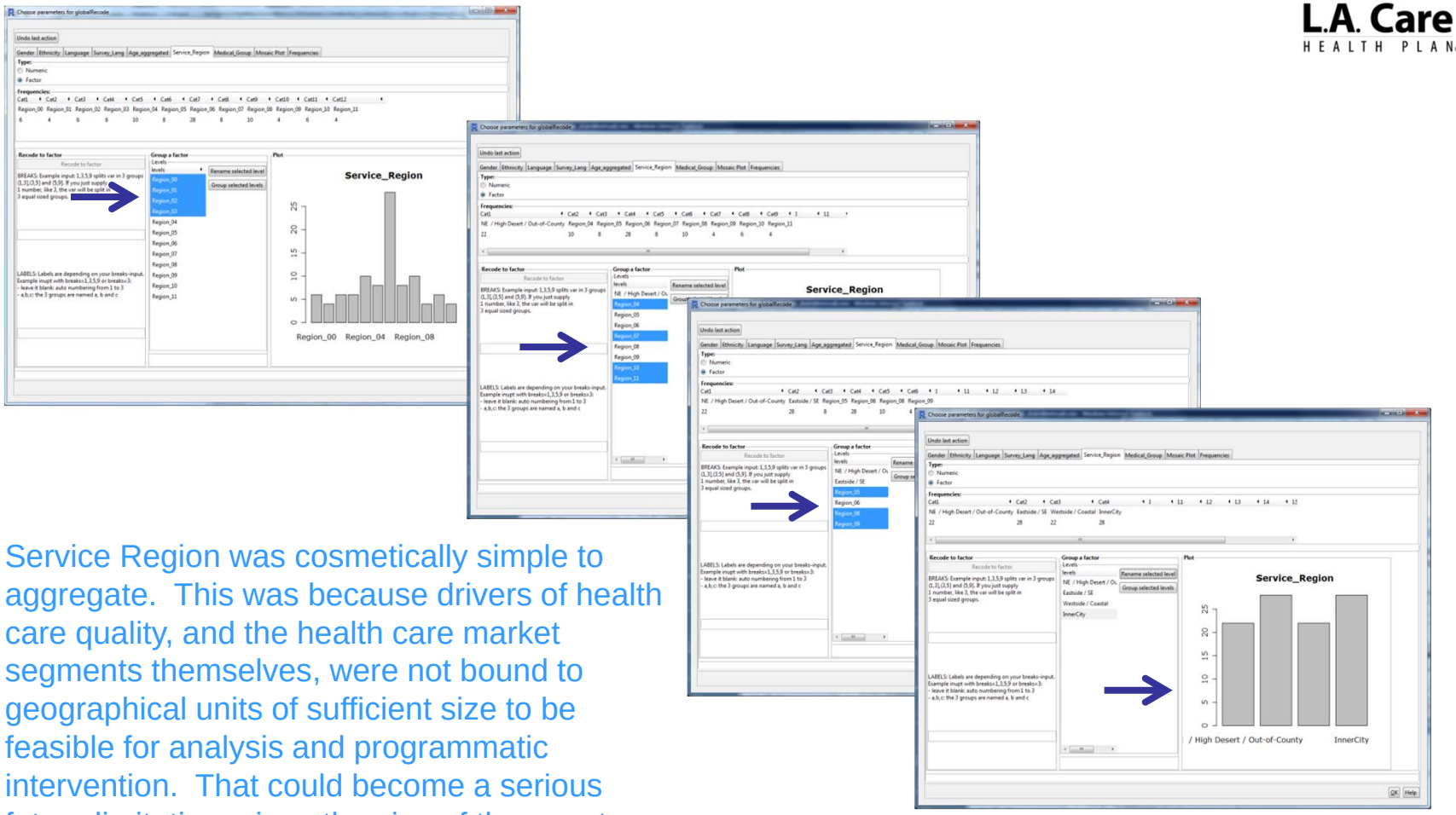




Even after aggregating along semantically-defensible lines, the Language variable remains sparse.

Example (Cont.): Aggregate Service Region

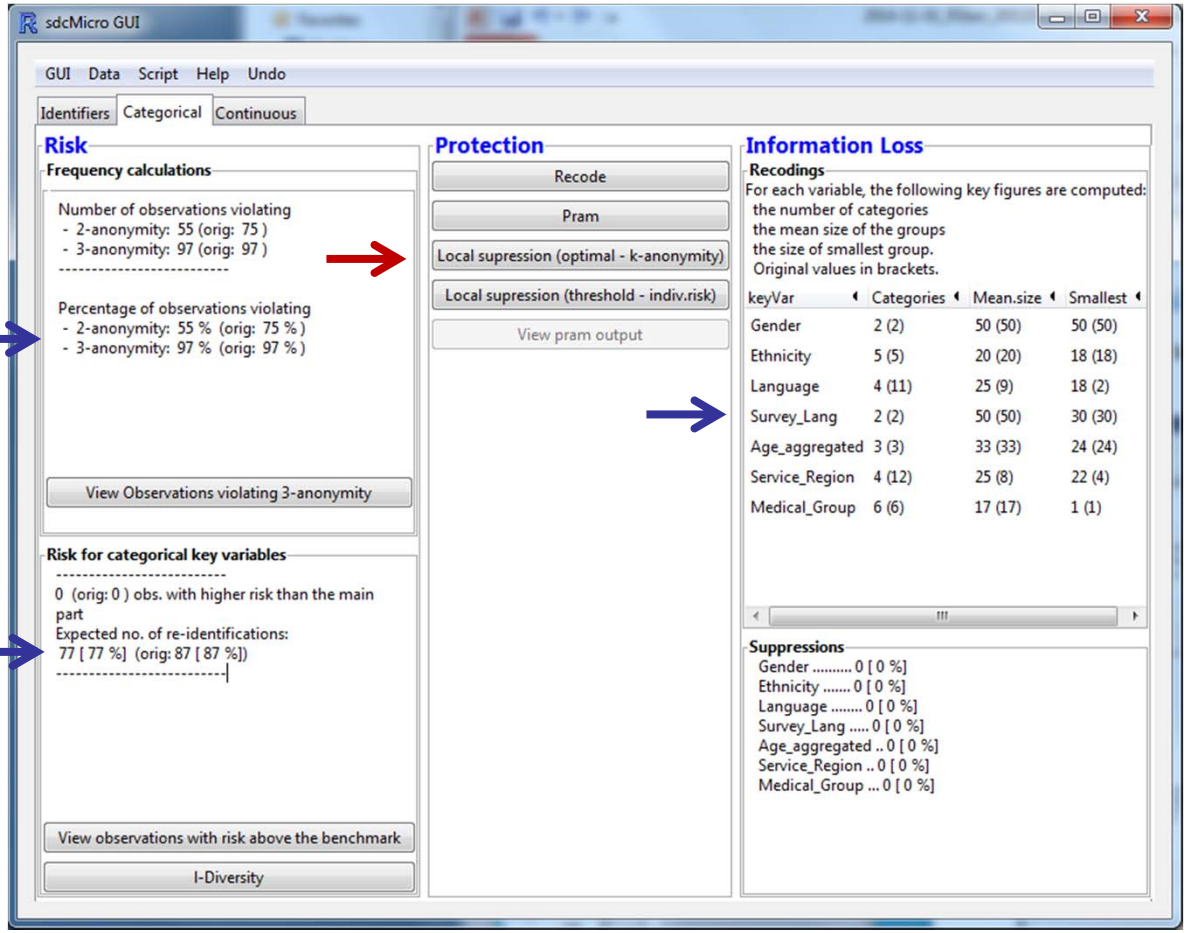
Begin modifying dataset – recode/aggregate sparse areas:
NE / High Desert / Out-of-County; Eastside; Inner City; Westside/Coastal.

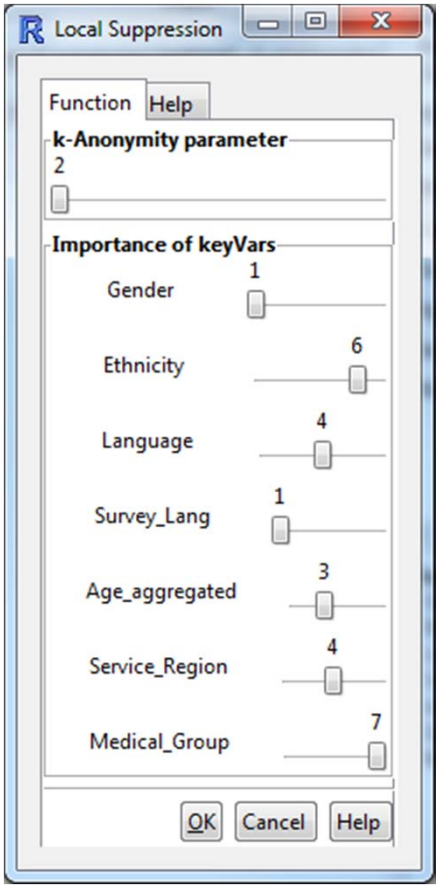


Service Region was cosmetically simple to aggregate. This was because drivers of health care quality, and the health care market segments themselves, were not bound to geographical units of sufficient size to be feasible for analysis and programmatic intervention. That could become a serious future limitation, given the size of the county.

Example (Cont.): Consider Next Steps – Suppression

After “informed recoding” consider statistical tools.
Despite dramatic reduction in sparse categories,
re-identification risk (77%) has not improved much.





V. Real World Example: CAHPS 2014 Sampling Frames

Process:

- Identify which variables are important to decision-makers.
- Omit unnecessary variables.
- Suppress sparse variables. (State, Phone_flag.)
- Review possible redundant variables. (E.g. Language ⇔ Ethnicity.)
- Aggregate sparse values. (Language, Ethnicity, Service_Region.)
- Perform local suppression.
- Perform perturbation on dependent variables (preserving means, variability).

Considerations:

- A realistic threshold for CAHPS survey work is k=10*, but the software's authors note that smaller k-values are also used in practical work*.
- k=10 in health care survey work, is likely based on perceptions about survey respondents' tolerance for risk of re-identification.
- That perception should be empirically tested among patients, weighing the risks to anonymity, against the benefits of finding root causes of bad service.

Potential fallbacks for privacy policy:

- Have survey firm run privacy-sensitive analyses, providing no limited dataset.
- Planned phase-in and phase-out of released variables after "y" years.
- Purge archival data to keep dep. vars while replace old IVs with improved IVs.

Case Study: Demographics Released in CAHPS Limited Data Set

Demographics deemed releasable under 2010 HIPAA guidance:

"What we received."

	Categories	Likely risk	Risk of uniqueness (% blinded by survey firm)	Analytic value to Health Plan in that form	Analytic value if aggregated	Preferred course (not avail.)
Survey language	2 NS	Low				**Keep.**
Gender	2 NS	Low				**Keep**.
Ethnicity	21 S	High	x (7.7%)	Low	Med	Aggregate this.
Language	24 S	High	x (10.1%)	Low	Med	Aggregate this.
State	50 S	High		Low	Low	Omit.
Flag -- address present	1 NS	Low		Low	Low	Constant (no effect).
Flag -- phone # present	2 NS	Mod		Low	Low	Omit aft. 3 yr.
Decimal age (18-80)	S	High		Med	High	Aggregate this.
Integer age (cap at 80)	81 M	Med		Med	Med	Aggregate this.
Age group	13 M	Low		Med	Med	**Keep.**
Service region	12 M	Med	x (5.5%)	Low	Med	Aggregate this.
Aid Code	41 S	High	x (26.7%)	Low	Med	Aggregate this.
CCC flag	3 M	Med		Med	Med	**Keep.**
CCC prescreen code	2 M	Med		Low	Low	Omit.
Dual-eligible	2 M	Med		Low	Low	Omit (redundant).
Line-of-Business	5 M	Low	X (0%)	High	High	Aggregate stray uniques.
Pseudo medical group	12 M	High	X (13.4%)	Med	Med	Omit.
Provider Group / Clinic	138 M	High	X (16.3%)	High	Med	Aggregate this.
1115 Waiver SPD	3 M	Mod		Mod	High	**Keep.**

S=Sparse M=mixed NS="non-sparse" but are non-sparse only if non-missing (since M.D. is also an identifier). #=Largely fixable by aggregating or omitting "stray uniques" (patients in categories too sparse for programmatically useful intervention). #=Can be aggregated or removed without impairing analysis for quality improvement analysis.

Impact on Utility – (Downside of "Anonymization on Autopilot")

Demographics deemed releasable under 2010 HIPAA guidance:

"What it means."

Out of 414 categories, 79 were not valuable, and >90 could have been aggregated without harm.

	Categories (degrees of uniqueness ≈ degrees of freedom)	Risk of uniqueness	Blinded by survey firm (X=impacts utility x=less impact)	Analytic value to Health Plan in this form	Analytic value if aggregate	Preferred action would be:
Survey language	**2 NS**	**Low**				**Keep.**
Gender	**2 NS**	**Low**				**Keep.**
Ethnicity	21 S	High	x (7.7%)	Low	Med	Aggregation (not blinding).
Language	24 S	High	x (10.1%)	Low	Med	Aggregation (not blinding).
State	50 S	High		Low	Low	Omit in favor of other demogs.
Flag -- address present	1 NS	Low		Low	Low	Constant (no effect if omit).
Flag -- phone # present	2 NS	Mod		Low	Low	Study for 3 years then omit.
Decimal age (18-80 only)	~ S	High		Med	High	Aggregate this.
Integer age (capped at 80)	81 M	Med		Med	Med	Aggregate this.
Age group	**13 M**	**Low**		**Med**	**Med**	**Keep.**
Service region w/i county*	12 M	Med	x (5.5%)	Low	Med	Aggregate this.
Aid Code	41 S	High	x (26.7%)	Low	Med	Aggregate this.
CCC flag	**3 M**	**Med**		**Med**	**Med**	**Keep.**
CCC prescreen code	2 M	Med		Low	Low	Omit in favor of other demogs.
Dual-eligible	2 M	Med		Low	Low	Omit – (is in Medicare CAHPS).
Line-of-Business	5 M	Low	X (0%)	High	High	Aggregate stray uniques.
Pseudo medical group	12 M	High	X (13.4%)	Med	Med	Omit in favor of other demogs.
Provider Group / Clinic	138 M	High	X (16.3%)	High	Med	Aggregate by custom rules.
1115 Waiver SPD	**3 M**	**Mod**		**Mod**	**High**	**Keep.**

"What Was Lost" – Actionable Variables Tacitly Crowded Out of Analysis

Variables left out due to generic choice of demographics w/o user input:

- **Provider Group / Clinic:**
  - Large stand-alone medical groups. County clinics. Inner city IPAs as an aggregate.
- **Disease cohorts:**
  - Large groups include: Asthma, diabetes, depression (adult), obesity.
  - Risk score stratum (risk of ER visit, hospitalization, or risk of higher costs (as proxy for health status)).
- **Intervention groups:**
  - Members whose doctors or medical groups are in an incentive program.
  - Patients receiving disease management services.
  - Patients receiving case management services.
- **Members seeking or receiving services directly or indirectly rated on CAHPS**
  - Members using Family Resource Centers.
  - Members calling Call Center.
  - Members using language services.
  - Members whose doctors are Board Certified.
  - Members whose doctors have and use Electronic Health Records and HIT.
  - Members whose doctors are contracted in staff model health plans, or medical groups, or IPAs.
- **Home circumstances and support system:**
  - 2010 Census economic data on member's census tract.
  - Information from Initial Health Assessments (IHAs).
- **Continuity of care:**
  - Member's enrollment history with the health plan. (Retention analysis.)
  - Member's assignment history with the doctor or clinic. (Retention: "Good service ties to revenue.")

Careful selection guided by a statistician could have salvaged some of these.

Demonstration Based On CAHPS 2014 Sampling Frame Files

	Adult	**Child**
L.A. Care population size:	n=338,175	n=677,536
Before any modification:		
• Categories:	288 + 2 contin. vars.	288 + 2 continuous vars.
• Quasi-identifiers (vars):	12	13
• **Re-identification rate:**	***34.36%***	***32.54%*** **(High risk.)**
After severe manual aggregation:		
• Categories:	35	35
• **Re-identification rate;**	***3.51%***	***2.04%*** **(Much better.)**
Before suppression:	State, Phone_Flag	State, Phone_Flag
After local suppression (k=10):		
• **Re-identification rate:**	***1.39%***	***0.84%*** **(Not quite done.)**
• Mainly affected:	Provider_Group	Provider_Group
• (Coincidentally similar	Ethnicity	Service_Region
• in Adult and Child,	Service_Region	Ethnicity
• but at different rates.)	Language	Language

Next steps to block direct re-identification of anyone would include perturbation: Post-Randomization Method (PRAM), micro-aggregation, adding noise, shuffling.

Realization: To get robust sets of analytic variables, anonymity protection (even with large populations) requires radical aggregation and perturbing response data.

VI. Discussion: Learning Objectives

1. Explain the importance of anonymity in surveys of populations particularly vulnerable to disclosure.
   - Patient populations include vulnerable populations (children, persons with disabilities, groups that may be exposed to discrimination related to health conditions, etc.).
   - Health data contain very private and sensitive information (e.g. mental illness, physical disabilities, sexual behavior, addiction recovery, etc.).
   - Misuse of the information could impact coverage and benefits.
   - Misuse of the information could harm the patient vis-à-vis outside parties (spouses, employers, customers, other insurers, etc.).
2. Describe the tradeoff between anonymity and applied value to make health care surveys actionable.
   - High protection means low information.
   - High information means low protection.
   - In improving quality of health care and services, *targeting is key.*
   - Overprotecting patients can mean that corrective actions can't reach them.
   - Overprotection suppresses patient voice.

Discussion / Learning Objectives: Analytic Implications

3. Discuss analytic implications imposed by common rules for protecting anonymity.
    a. Analysis of causes: Indicators of whether or not a patient has a given *condition or set of conditions.*
    a. Evaluation of programs: Indicators of the effectiveness of giving a patient a particular *treatment* or *intervention / program.*
    b. The precision of multivariate analysis is hampered by coarse coding and aggregating of continuous variables (age, months of coverage, number of visits, height and weight in fractional form); and course groupings of categorical variables (ratings, demographics: ethnicity, language, geographical region, clinics and provider groups.
    c. The rational response of health plans to analytic restrictions, is to *do* ***off-season non-anonymous surveys*** *similar to agency surveys:*
        a. Overprotection means added cost, with burden and risk to members.
        b. Either mid-season (biannual) surveys or short small-sample monthly or quarterly tracking surveys pooled for analysis.
        c. Collision course: Survey firms note that response rates have slowly been declining nationally. Agencies have also cut survey size in response to concerns about respondent burden.
        d. Implication: Non-response bias and self-selection will become more prominent factors on survey-based assessments of service quality.

Discussion: Considerations in Risk Assessment

4. Describe specific risks from releasing various demographics in response data from surveys.

   Sensitivity of the information for the respondent.
   Vulnerability of the respondent to retaliation (young, infirm, mentally ill).

   Degree of *incentive* present for end-users to breach anonymity. (Rationally assess if benefits from retaliation against patients would justify the cost and risk of doing so.)

   Degree of *disincentive* present to prevent end-users from breaching anonymity. (Would the consequences of getting cost outweigh the cost and risk of retaliating?)

   Balance that against tangible risk from not improving quality of services.

   Main protection for patients is in keeping the value low for breaching anonymity:

   - The reward/risk ratio to a health plan in breaching anonymity and retaliating against an individual patient for a poor rating on a survey, is arguably quite low.
   - In contrast, the reward/risk ratio could be higher for an organization involved in litigation with an individual patient with exploitable information on a survey ("Plaintiff X gave this institution a high rating on the last survey.").

   It is equally important to grant that: (a) Survey anonymity policies are driven by ideals, less than proven empirical harm to patients who responded to surveys. But (b), the risk of harm from breaching anonymity is arguably non-zero, and some (unknown) degree of patients' response rates, candor, and accuracy of ratings on surveys are based on the promise of anonymity.

Discussion / Learning Objectives (Cont.): Calculation of Risk

5. Discuss degrees of anonymity and their appropriate application.

   Anonymity rules are typically calibrated by:

   - Degree of implied consent by the patient.
   - Role and "need to know" to perform consented duties.

   Potential enhanced criteria:

   - Conflict of interest vis-à-vis the content of the information.
     (Doctor will know the results of an exam, but need not necessarily know the patient's confidential rating of that service on a survey.)

6. Explain the calculation of risks to anonymity in common demographics.

   Produce a tree reflecting the full list of demographics desired back in the dataset. Populate that tree with the cases from the sampling frame. Tally the number of persons in the furthermost branches of that tree. Identify the smallest branch. If that branch has 10 persons, for example, 1/10 yields the risk: a 1-to-10 chance of correctly identifying one of those 10 persons from that list of demographics. The dataset would be compliant with a cell size rule of 10 or higher.

   When additional data sources are factored in, the risk math becomes non-trivial – particularly with CAHPS (many measures, many end users with access to various data sources).

   The patient's main protection is in keeping the value of breaching anonymity low.

Discussion / Learning Objectives (Cont.): Risk Drivers

7. Identify factors that play into the calculation of risk against a predetermined threshold.

The size of the patient population in categories in the sampling frame reduces risk.

Sparseness of categories in an aggregate variable increase risk.

Missing data can act as an identifier.

Approaches to reducing threats to anonymity in limited datasets from surveys:

- Removing sparse variables is an absolute cure, but limits analysis.
- Aggregating sparse categories until the cell size rule is satisfied, is a common strategy.
- Randomly perturbing the data can preserve statistical properties (average, variance) while protecting anonymity.

However, end-users may distrust statistically-perturbed data.
More knowledge about the technology may help survey firms and client health plans negotiate mutually-agreeable data release plans that are agency-compliant.

If survey firms and clients can discuss and use the same software tools and calculated measures of risk for sampling frame files – (presumably arriving at the same measures and conclusions) – that would help build confidence in limited datasets released at the end of the survey process).

Discussion / Learning Objectives (Cont.): Best Practices

8. Describe best (and worst) practices for using anonymous and non-anonymous data from vulnerable populations.

Best practices for survey firms creating limited data sets:

1. ***Notify*** *health plan users when masking techniques have been used on their data.*
2. ***Identify which techniques*** *have been used, and any threshold values used.*
3. ***Identify what percent of cases*** *in the dataset have been modified, and what percent of cases have been modified for each variable.*
4. *Provide* ***estimates of information loss*** *for each variable, and statements about which statistical properties were preserved or not preserved by the method chosen.*
5. *Survey firms and end-users sharing and using the same software tools would help.*

Negotiate terms of data release as part of the survey contracting process.

Sequester data within an independent evaluation department with a reporting venue above the highest level being evaluated.

Place survey function within that unit, *and sequester sensitive data there*.

Worst practices would include:

- Use of non-anonymous data for **personally-targeted member interventions**. (Even if intent is benevolent, breached privacy is unethical and worries subjects.)
- Aggregate to avoid appearance of releasing breachable data at the provider level.
  - Avoids providers trying to guess which demographic groups gave poor ratings.

Discussion / Learning Objectives (Cont.) – Multi-Year Context

9. Discuss how to adapt from a one-shot survey context into a multi-year data strategy for quality improvement analyses.

One test of an organization's sincerity in commitment to quality improvement, is the willingness to *simultaneously* have short- and long-term interventions and evaluation processes in play.

Identify the policy period:

- Number of years of relevant past data on hand.
- Number of years for program or intervention to reach steady state. (I.e. first full test of program effectiveness.)

It often takes 3-6 years for a program (and its evaluation process) to mature. Hence one gathers a pool of 3-6 years in a CAHPS limited dataset for program evaluation, within the k-anonymity rule negotiated with the survey firm.

Discussion / Learning Objectives (Cont.) – Multi-Year Context

9. Discuss how to adapt from a one-shot survey context into a multi-year data strategy for quality improvement analyses (cont.).

The analytic plan for anonymous survey data should accommodate long-term needs by allowing flexibility.

- Keep a few degrees-of-freedom in reserve for inclusion of new policy variables.
- Over time, can phase out some variables, while phasing other variables in, subject to the negotiated k-anonymity rule.
- Survey contracts can also include provisions to allow for destruction of limited datasets to be replaced by limited datasets containing the same response data for the same time series of years, but containing different independent variables suited to meet current analytic needs.

Analytically, low-scoring programs with few years of archival data are in the worst position in terms of data, resources, and client urgency.

VII. Avenues For Further Research On This Topic

Anonymity policy would benefit from research into the practices described in this briefing:

- What are common practices among CMS- and NCQA-certified survey firms in creating limited datasets for CAHPS and other patient experience surveys?
- What are common practices among health plans contracting for such surveys, in terms of release of data, and the specific content of limited datasets?
- What methods and software tools do survey firms use to anonymize limited datasets? How many simply use the HIPAA Safe Harbor rules? How many do calculations to assess risk? How calculate data utility remaining after completing anonymization?

Anonymity policy would benefit greatly from empirical research into the assumptions on which agencies' anonymity policies are based:

- Incidence and degree of harm from breaches of anonymity on health care surveys.
- Degree to which patients trust the promise of anonymity on health care surveys. (If not trusted, the restriction might not be providing accuracy or candor. Since separate, non-anonymous surveys of members are routinely allowed, anonymity policy on surveys primarily seeks to protect the validity of the survey results comparing health plans.)
- Degree to which patients are more candid and accurate in giving assessments about health care quality, tested both with and without the promise of anonymity.
- Degree to which patients are more candid about offering negative ratings of doctors or health plans, when being surveyed by the sponsoring institutions.

Contact Information

S. Rae Starr, M.Phil, M.OrgBehav
L.A. Care Health Plan
RStarr@LACare.org, rae_starr@hotmail.com
213-694-1250 x-4190