

## TOP TEN LIST

# TEN BEST WAYS TO ENSURE A BAD HEALTH DATABASE OR “DIRTY DATA, DONE DIRT CHEAP”

With apologies to David Letterman, and thanks for editorial assistance to Elizabeth Kirby and for their insights to the following Internet contributors:

Robert Meyer, NC Center for Health Statistics

Kimberlea Hauser, U. of South FL

Greg Alexander, U. of AL-Birmingham

Michael Soref, WI Bureau of Health Information

Randall Rempel, Veterans Health Administration, Washington DC

Mathias Forrester, TX Birth Defects Monitoring Division

Jerrold Jacobson, U. of WI Medical School

R.S. Kirby, January 2000

Top Ten List: Ten Best Ways to Ensure a Bad Health Database

## Number 10

**Include the same variables redundantly in as many tables in the application as possible.**

**For maximum effect, store the variables as numeric fields in one table and character in the next, under the same variable name.**

**Example: SEX (M/F) and SEX (1/2)**

## Number 9

**“Less is Better”**

**Collect less information than is necessary to understand the context of any given record.**

**You can always go back to the original reporting source for clarification, or re-abstract the case if more data fields are needed later.**

## Number 8

**Keep no documentation for your database application. Over time, the database will become inoperable and unfixable due to lack of information.**

**Complement this with sufficient turnover in database staff that no one knows how or why any of the tables were created, what the codes mean, or how specific fields and tables relate to one another.**

## Number 7

**As an aid in unduplicating records in the main table, use the first three digits of the Social Security Number as a unique identifier.**

SSNs are unique, aren't they? What? The first three digits refer to regions of the country? OK - I guess I'll use the first three digits of the telephone number instead. Or maybe three digit ZIP Codes?

## Number 6

**Design data fields to enable numeric and character data entry. This way, when the table is accessed dynamically or imported into another application you'll be sure to get a bad result - missing values most of the time.**

Clinical example: Key ICD-9-CM codes (ICD-10 if you are really advanced!) as numeric values in a text field in your spreadsheet. When importing into an MS Access table, the values will convert to numeric values, unless you pay attention. Is this a problem? Heck no - if you want "Dirty Data, Done Dirt Cheap"!

## Number 5

Select variables and design data collection forms without a clear vision of what the database will be used for.

You can always fine-tune later, drawing from the excellent suggestions you receive by email, phone, FAX and in your complaint box!

## Number 4

Design your “database” as a single, flat file which includes all of your variables, even if only a tiny fraction of the cases have values for any specific variable.

Why worry about data storage and computer efficiency? Run out of space - just buy another 20 Gb hard drive at Best Buy for \$200.

## Number 3

**Collect continuous variables (maternal age, educational attainment, body mass index, etc) only as categorical values developed for another database or research project with a completely different purpose.**

**Example: instead of collecting the actual number of years of formal education completed, collect in the following categories: less than 6, 6-9, 10-11, 12-13, and 14 or higher. Be sure to code the categories ordinally to increase the probability of inappropriate use of these data in linear regression analyses.**

## Number 2

**In fields with pull-down menus of optional values, make sure that 'Other', 'Not Specified', or similar catch-all categories have the highest frequencies.**

**Administrators and politicians love to explain how "All Other Causes" was the leading reason for hospital admissions, trauma-associated deaths, or the leading cause of death.**

Top Ten List: Ten Best Ways to Ensure a Bad Health Database

## Number 1

Want to improve on the blissful imperfection of MS Office 2000 Professional Edition? Create an animated paperclip, and have it appear on the screen and sappily say to your data entry staff:

“It looks like you’re trying to key data - want some help?”

Better yet, fail to provide a manual override, so the animated paperclip ALWAYS appears.

Top Ten List: Ten Best Ways to Ensure a Bad Health Database

## Y2K Bonus

For consistency as you deal with the Y2K crisis in your health databases, use Roman numeral formats for ALL numeric fields, not just for dates.