

The Metamorphosis of a Study Design

Marge Scerbo, CHPDM/UMBC
Craig Dickstein, Intellicisions Data Inc.

Abstract

In a perfect world, there would be perfect data, perfect analysts, and perfect programmers creating perfect outcomes to every possible study. Unfortunately, one, two, or all of these factors are usually imperfect. Data are, especially data in large volumes, rarely flawless. Researchers and analysts designing studies may have great ideas of studies to undertake, but may have little idea of whether it can be done or how to do it. Programmers may be incredibly facile with the software, but rarely comprehend all the intricacies needed to complete a study. Thus, study designs are not often etched in stone. Most likely, they are the outcome of a long and tedious process of checks and balances.

This paper will take the reader through the process of developing a study design, using SAS software to provide results on which to base outcomes. A health care policy issue will be used as the basis for the discussion, but the ideas should carry across many industries.

Introduction

A good programmer analyst must work with a variety of methodologies within a single project. The 'programmer' portion of the brain is organized, methodical, and logical. The 'analyst' is quite a different story; patience, foresight, and a depth of understanding of both the data and the outcome, beyond merely understanding code structures, is required. In a sense, the analyst must be a mind reader and a magician.

Health care data are a world unto themselves. There are vast amounts of administrative (billing) data produced daily. Except for the payment and/or collection of bills, these data are largely underused or misused.

Note that the study discussed in this paper is fictional, and that none of the data can be associated with any state or institution.

Background

Health care billing data are in three primary formats: HCFA-1500, Pharmacy, and UB-92.

HCFA-1500 records contain professional service information provided by an 'individual' practitioner. The place of service for these claims and encounters can encompass many venues, including doctor's offices, laboratories, and

hospitals. Usually, one record is analogous to one procedure in a physician's office, and so one visit may include multiple records.

Pharmacy data are the cleanest, most efficient, and easiest to manipulate. Most pharmacy data are collected at the 'point of sale' (POS), right in the drug store. These records contain information about the drug, the prescription, the provider, and the patient. One record is essentially one prescription.

On the other hand, although they represent the largest percent of health care dollars, UB-92 data are not easy to use. These files are uniquely produced by facilities: acute care hospitals, hospices, nursing homes, emergency rooms, and outpatient clinics. These data are far more complicated when used for analysis. There is no clear definition for an inpatient data record; this depends on the database design of the keeper of the data.

Initial Study Design Proposal

HMOs (Health Maintenance Organizations) and MCOs (Managed Care Organizations) are based on the premise that by providing good preventive care and case management, fewer facility charges will be incurred and overall cost will decline. In this vein, HMO and MCO management is always looking at the bottom line for possible savings.

One HMO administrator was asked to identify methods of saving money on hospitalizations. He reviewed the various monthly and quarterly reports relating to this subject. After several weeks of review, he decided that it might be possible to save money by transferring patients from a hospital into a nursing home setting more quickly than was presently the case.

Hence, a new study is launched...

Under the present HMO contract with a large corporation, the HMO was responsible to pay for the first 31 days in a nursing facility.

A meeting took place between the administrator, an analyst, and a programmer. In order to best understand what analysis should be undertaken to get the best results, the analyst prepared a list of questions (the administrator's responses are shown in *italics*):

- What type of data should be studied?

- *Inpatient*
- As inpatient data are requested, can you specify exactly what type of facility should be studied? Inpatient facilities include acute care hospitals, rehabilitation centers, chronic hospitals, hospices, and other special settings.
 - *Acute Care Hospitals*
- Nursing homes are classified as three different types: ICF (Intermediate Care Facility), SNF (Skilled Nursing Facility), and ICF-MR (Intermediate Care Facility for the Mentally Retarded). Are you interested in all types or specific ones?
 - *ICF only*
- For acute care services, can we narrow the population studied by other criteria?
 - *Include only those patients that can be identified as still hospitalized for over 5 days.*
- Are these same criteria to be imposed on the ICF population?
 - *No*
- To continue, what type of information would be useful?
 - *The number of patients*
 - *The total dollar amount*
 - *The total number of days*
 - *The average cost per day*
- In addition, what information do you want from the ICF file?
 - *Same as that from the acute care file*
- Finally, is there a specific set of dates of services you are interested in?
 - *Whatever file is the most current and yet the most complete.*

The analyst identified calendar year 2000 data as the most complete and the most current. The first step proposed in the study design is to count the number of patients at the close of calendar year 2000 who remain in the hospital and the number of patients who were in an ICF during calendar year 2000.

The initial programming request includes the following:

- Obtain a frequency count on discharge status in the UB-92 year 2000 acute care summary file
- Select those patients who are still hospitalized
- Select those patients who have been hospitalized over 5 days
- Calculate their cost per day

- Calculate the cost per day for patients who are in an ICF
- Calculate the difference in costs
- Produce tabular reports on the demographic (age and gender) identifiers of selected populations

These reports are to be delivered within one week of the initial proposal.

Overview of the Data

UB-92 data are both complicated and extensive. These data files are not comparable to a hospital medical record, which contains notations on every drug, laboratory test, physician visit, procedure, etc. that is incurred during a patient stay. Rather, these files contain billing data. Services are collapsed into revenue codes with units of service and charges attached. For example, multiple laboratory services may be grouped under the revenue code 300, 'Laboratory, General Classification'. In addition, one inpatient hospital stay may in fact be billed across several UB-92 records, some with charges and others with adjustments, some across a particular date range and others across the date range from admission to discharge.

This HMO's IS department, in order to better utilize UB-92 files, had written SAS programs to create discharge summaries; where all possible records associated with a unique patient stay were stored in one large record. Charges, units, and days are totaled under this file structure design. These data sets have been validated for quality and are sorted by the recipient identifier (RECIPID) to allow for ease in merging processes. In this situation, the discharge summary data set for acute care inpatient discharges (ACUTE00) and a separate discharge summary data set for nursing home facilities (LTC00) will be used. Both of these data sets also contain two important fields for analysis: the length of stay (LOS) field containing the total number of days of the stay and the payment (PAYMENT) field containing the total cost of the stay.

The demographic information on each patient (AGE, GENDER) is stored in an annual enrollment file that contains one record per patient.

Preliminary Analysis

In order to satisfy the first request, the programmer ran a frequency analysis of the discharge status (DISCHSTATUS) in the acute care discharge summary file (ACUTE00). There exists a format for the discharge status

(STATUS.) that is used to produce more readable results:

```
proc freq data = acute00;
  tables dischstatus;
  format dischstatus status.;
run;
```

The results are shown in Table A. This frequency analysis demonstrates that 1,021 patients were transferred to an ICF in calendar year 2000 while 844 patients were still in the hospital at the end of the year. Code 30 is defined as 'still in the hospital' and code 5, 'transferred to ICF'.

The analyst, stepping beyond the exact specifications, request frequency counts on the length of stay field found in the acute care file. Rather than producing a report that showed the count of each length of stay, the programmer grouped the days to provide a more concise report:

```
proc format;
  value days
    0-5      = '0-5'
    6-high   = '> 5'
run;
proc freq data = acute00;
  tables los;
  format los days.;
  where dischstatus = 30;
run;
```

The output of this report is shown in Table B. It clearly shows that of the 844 patients still in the hospital, *only 32.7% of the patients (276 people) had stays of over 5 days.* This table serves several purposes, the most important being a checkpoint for the files to be created in the next steps of the study.

The programmer then proceeds to create the requested data set containing only those patients still in the hospital and with a stay greater than 5 days:

```
data stillin;
  set acute00 (where = (los gt 5
    and dischstatus = 30));
run;
```

The LOG reports:

NOTE: The data set WORK.STILLIN has 276 observations and 59 variables.

This data step allows the analyst a double check that 276 patients fell into this category, as was

reported on the discharge status frequency analysis. In addition, this data set can now be used for further studies.

The average cost per day of these 276 patients is then calculated using PROC SUMMARY. (Note that PROC MEANS also provides similar output.)

```
proc summary data = stillin;
  var payments los;
  output out= inptcost
    sum= inptdol inptdays;
run;
data results;
  set inptcost;
  costperday = inptdol/inptdays;
run;
proc print;
  var _freq_ inptdol inptdays
    costperday;
run;
```

Output of this routine (Table C) demonstrates the following:

- The number of patients (_FREQ_) is 276
- A total cost (INPTDOL) of \$15,779,690
- A total number of days (INPTDAYS) of 6,053
- Thus, the average cost per day is \$2,606 (INPTDOL/INPTDAYS)

A similar process is then used to calculate the number of people in ICFs. Since there are several types of long term care facilities contained in this file and the request specified that only ICFs were to be included, a provider type (PROVTYPE) code of 57 is used to identify ICFs:

```
data nursinghome;
  set ltc00 (where= (provtype = 57));
run;
```

NOTE: The data set WORK.NURSINGHOME has 5983 observations and 43 variables.

The total costs are calculated thus:

```
proc summary data = nursinghome;
  var payments los;
  output out = nhcost
    sum = nhdol nhdays;
run;
data results;
  set nhcost;
  costperday = nhdol/nhdays;
run;
proc print;
```

```

var nhdol nhdays _freq_
costperday;
run;

```

Table D.1, containing the output of this procedure, shows:

- The number of patients (_FREQ_) is 5,983
- A total cost (NHDOL) of \$876,121,178
- A total number of days (NHDAYS) of 1,143,242
- Thus an average cost per day is \$767 (NHDOL/NHDAYS)

Thus, the difference of cost per day between the inpatient hospital and the nursing home is \$1,839 (\$2,606 - \$767).

Warning – Inconsistency Detected

Although the programmer had followed the specifications, certain inaccuracies were apparent when this information was presented to the analyst:

- There are 1,021 patients identified in Table A as transferred from an inpatient hospital to an ICF. The results of the ICF study showed 5,983 patients in nursing homes. What caused this discrepancy?
- The HMO administrator had stated that the HMO is required to pay only up to 31 days in a nursing home facility. Should this be included in the design?

Before completing the final tasks listed in the initial study design, updates were required to the study.

Revised Study Design

The study design is now revised to include additional subset criteria:

- Include only patients who can be identified as having transferred from a hospital to an ICF (discharge status of 5)
- Of those patients, select only those with a length of stay of 31 days or less

Additional Analysis

The first step to be completed is the selection of those patients who were transferred from a hospital to an ICF:

```

data icf;
merge acute00 (in = hosp keep=

```

```

recipid dischstatus
where=(dischstatus=5))
nursinghome (in = ltc
where=(los le 31));
by recipid;
if hosp and ltc;
run;

```

NOTE: The data set WORK.ICF has 1021 observations and 44 variables.

This process identifies the desired 1,021 patients who have been transferred from a hospital to a nursing home. The next step is to summarize the payments for this group of patients:

```

proc summary data = icf;
var payments los;
output out = icfcost
sum = icfdol icfdays;
run;
data results;
set icfcost;
costperday = icfdol/icfdays;
run;
proc print;
var _freq_ icfdol icfdays
costperday;
run

```

The output (Table D.2) contains this information:

- The number of patients (_FREQ_) is 1,021
- A total cost (ICFDOL) of \$8,592,300
- A total number of days (ICFDAYS) of 5,242
- Thus, an average cost per day of \$1,639 (ICFDOL/ICFDAYS).

Consequently, the difference in cost per day between the inpatient hospital and the nursing home is \$967 (\$2,606 - \$1,639). Note that the number of patients now matches the number identified in the original frequency analysis (276 patients still in the hospital and 1,021 patients transferred from a hospital to an ICF).

With the populations of both facilities correctly identified, the programmer needs to add the demographic information to the newly created data sets:

```

data stillin;
merge stillin (in = hosp)
enrollment00 (keep= recipid
age gender);
by recipid;
if hosp;
run;

```

Once the new variables have been added to the data set, the programmer produces the demographic tables requested using this simple code:

```
proc format;
  value ages
    0 - 18 = 'Children'
    19 - high = 'Adults';
run;
proc freq data = stillin;
  tables age * gender / list;
  format age ages. gender $sex.;
  title 'Demographic Identifiers
        for Still in Hospital';
run;
```

The demographic studies can be found in Tables E and F.

Review of the Materials Produced

A meeting took place with the original team and a clinical specialist to provide further insight. Each report was studied carefully. Although there were clearly differences between the per day cost of a hospital and a nursing home, the clinician did not believe there was enough information to implement policy changes. The demographic tables (E and F) did not provide any significant decision-making information.

Without further studies to assess the diagnoses involved or the clinical appropriateness of transfers from a hospital to an ICF, no action plan could be set in place. Was it possible there were other areas of study that might provide clearer results?

After discussion, additional studies were requested that included:

- Produce a report of the 15 top primary diagnoses associated with those patients still hospitalized
- Produce a report of the 15 top primary diagnoses associated with those patients in ICFs
- Create a list of the top five costliest hospitals for those patients still in the hospital
- Create a list of the top five costliest nursing homes to which hospital patients were transferred

Additional Coding

In order to complete the final programming request, the programmer created various SAS programs. The first task was to create a report

showing the top 15 diagnoses for patients still in the hospital as defined by the study design:

```
proc freq data = stillin order = freq;
  tables primarydx / noprint
    out = acutedxs;
run;
proc print data = acutedxs (obs = 15);
  var primarydx count;
  title 'Top 15 Diagnoses for
        Still in Hospital';
  format primarydx $icdcode.;
run;
```

The results of these studies are displayed in Table G. This code can be edited to create the same report (Table H) on patients in an ICF.

The next step is to produce a listing of the top 5 hospitals and nursing homes. This code was written:

```
proc summary data = stillin nway;
  class hospital;
  var payment;
  output out = hospcost
    sum = totcost;
run;
proc sort data = hospcost;
  by descending totcost;
run;
proc print data = hospcost (obs=5);
  var hospital totcost;
  format totcost dollar15.;
  title 'Top 5 Hospitals';
run;
```

The results of this analysis are shown in Table I, with the corresponding information on ICF costs in Table J.

Final Study Design

The final study design is written to include all the steps necessary to produce the required results:

- Obtain a frequency count on discharge status in the UB-92 year 2000 acute care summary file
- Select those patients who are still hospitalized
- Select those patients who have been hospitalized over 5 days
- Calculate their cost per day
- Identify patients who have been transferred from a hospital to an ICF (discharge status of 5)

- Of those patients include only those with a length of stay of 31 days or less
- Calculate the cost per day for patients who are in an ICF
- Calculate of the difference in per day costs between hospital and ICF facilities
- Produce tabular reports on the demographic (age and gender) identifiers of the selected populations
- Produce a report of the 15 top primary diagnoses associated with those patients still hospitalized
- Produce a report of the 15 top primary diagnoses associated with those patients in ICFs
- Create a list of the top five costliest hospitals for those patients still in the hospital
- Create a list of the top five costliest nursing homes to which hospital patients were transferred

Conclusion - Outcome of Study

As shown, this study was implemented as simple frequencies and iteratively enhanced as each resulting table was available. Since this analysis might have direct impact on patients' treatments, it was important that clinical input was requested.

During the wrap-up meeting of the study team, one surprising and unexpected result was identified. Although there could be no final decision to move patients after 5 days from a hospital to a nursing home, it was clear from Table I (Top 5 Hospitals), that one hospital, University Center, accrued the largest dollar amount for patients hospitalized for over 5 days.

The clinician was aware of the variations of specialties across the hospitals, but could not explain the wide variation in total costs. It was determined that a meeting between the HMO and hospital administrators was needed. At that time, discussions as to the length of stay of patients in the hospital were discussed. The possibility of medical record review was proposed to assess this issue.

In addition, there were still several questions pending. Is this truly the final study or is this all that is available in the time allotted? Can any requirements be placed on physicians concerning length of stay that are clinically sound?

Should additional analysis take place? For example:

- Does the type of condition and the status of the patient control the outcome?

- What other factors may affect overall length of stay?
- Do patients recover more quickly in hospitals?

Complete study designs are imperative to valid analyses. They are created over time and may in fact present unexpected results. Programmers and analysts who can work within the variation of needs and personalities fill an invaluable role in any organization.

For more information on study designs, check the case study discussed in *Health Care Data and the SAS System*.

References

Scerbo, M., Dickstein, C., and Wilson, A. (2001). *Health Care Data and the SAS System*, Cary, NC: SAS Institute, Inc.

Contact Information

For more information contact:

Marge Scerbo
CHPDM/UMBC
1000 Hilltop Circle
Social Science Room 309
Baltimore, MD 21250
Email: scerbo@chpdm.umbc.edu

Craig Dickstein
Intellicisions Data Inc.
P.O. Box 502
Weare, NH 03281
Email: cdickstein@att.net

Table A - Frequency of Patient Status, CY2000

Status	Frequency	Percent	Cum. Frequency	Cum. Percent
Disch/Trans to Home or Self Care	87,219	85.4	87,219	85.4
Disch/Trans to Other Hospital	1,688	1.6	88,907	87.0
Disch/Trans to SNF	5,619	5.5	94,526	92.5
Disch/Trans to ICF	1,021	1.0	95,547	93.5
Disch/Trans to Other Institution	2,420	2.3	97,967	95.9
Left Against Medical Advice	1,397	1.3	99,364	97.2
Patient Died	1,901	1.8	101,265	99.1
Still in the Hospital	844	0.8	102,109	100.0

Table B - Formatted Frequency of Total Days - Still in Hospital

Totaldays	Frequency	Percent	Cum. Frequency	Cum. Percent
0-5	568	67.3	568	67.4
> 5	276	32.7	844	100.0

Table C - Output of Summary on Acute Care Patients

# of Patients (_FREQ_)	Total Dollars	Total Days	Average Cost per Day
276	\$15,779,690	6,053	\$2,606

Table D.1- Output of Summary on ICF Patients

# of Patients (_FREQ_)	Total Dollars	Total Days	Average Cost per Day
5,983	\$876,121,178	1,143,242	\$767

Table D.2 - Output of Summary on ICF Patients - Amended

# of Patients (_FREQ_)	Total Dollars	Total Days	Average Cost per Day
1,021	\$8,592,300	5,242	\$1,639

Table E - Demographic Identifiers for Still in Hospital

Female		Male	
Children	Adult	Children	Adult
26	99	43	108

Table F - Demographic Identifiers of Transferred to ICF

Female		Male	
Children	Adult	Children	Adult
56	452	65	448

Table G – Top 15 Diagnoses for Still in Hospital

Primary Diagnosis	Count
Schizoaffective Disorder	32
Congestive Heart Failure	28
HIV Aids	23
Pneumonia	23
Respiratory Distress	23
Hypovolemia	16
Septicemia	16
Rehabilitation Procedure	15
Depress Psychosis	12
Extreme Immaturity	12
Food/Vomiting	12
Paranoid Schizophrenia	12
Staphylococcal Pneumonia	10
Bipolar Affective Disorder	9
Decubitus Ulcers	9

Table H – Top 15 Diagnoses for ICF

Primary Diagnosis	Count
Senile Dementia	240
CVA	198
Cardiovascular Disease	99
Alzheimers	76
Cerebrovascular Disorder	32
Hip Replacement	26
Hypertension	24
Psychosis	18
Paralysis Agitans	17
Multiple Sclerosis	14
Diabetes	14
Decubitus Ulcers	9
Presenile Dementia	7
Depressive Disorder	7
Neoplasm	5

Table I - Top 5 Hospitals

Hospital	Total Payments
University Center	\$2,187,634
Ben Franklin Hospital	\$1,506,123
Union Square Hospital	\$414,168
Childrens Center	\$361,421
St. Johns	\$310,448

Table J - Top 5 ICFs

Hospital	Total Payments
Freetown Rehabilitation Center	\$146,940
Morristown Center	\$121,461
Main Eldercare	\$98,821
Northeast Convalescent Home	\$96,503
St. Michaels Nursing Center	\$87,606