Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Analysis of healthcare utilization data

Some practical considerations for investigators in palliative care

Peter May, PhD
Research Fellow in Health Economics, Centre for Health Policy & Management, Trinity College Dublin, Ireland
Visiting Research Fellow in Geriatrics & Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

February 2nd, 2016

Declaration

No financial interests to declare

This Webinar

Objective: To provide practical guidance for the analysis and reporting of healthcare utilization data, with a focus on (hospital) costs

Overview:

1. Introduction
2. Five considerations in data analysis
3. Concluding remarks

References for further reading detailed throughout

1. **Introduction**
2. Considerations in data analysis
3. Concluding remarks

Introduction

Why analyze utilization data?

Formally, we are interested in utilization analysis because:

- Health demands are infinite
- Resources to provide healthcare are finite ("scarce")
  - ➢ Decisions in allocation to be made

In practice the reason is the same as for any other type of study:

- Ensuring that the most effective care is made available
- Economic perspective is often useful (& typically essential at a systems/policy level)

Introduction

Why this webinar?

Utilization data are awkward:

- Unusual properties for statistical analysis
- Often deceptively complex to interpret

Practical consideration of how to organize and analyze data

(Not considered: where to get data)

Typically we estimate how x impacts y, given *varlist*, where:

y=dependent utilization variable (e.g. costs, admissions)

x=exposure (e.g. palliative care, hospice enrolment)

varlist=baseline independent variables

1. Introduction
2. **Considerations in data analysis**
3. Concluding remarks

1. Introduction
2. **Considerations in data analysis**
    1. **Determining cost data**
3. Concluding remarks

2.1: Determining cost data

Understanding your dependent variable

Count utilization data are self-explanatory:

– (Re)admissions (how many); length of stay (days)

$$$ data are more complicated:

Precision	**the cost of *what*?**	**To whom?**
Max	Direct measurement	Patient & their families
	Activity-based estimates	Provider, e.g. hospital
	Charges	Insurer, e.g. Medicaid
Min	Estimated charges	

2.1: Determining cost data

Understanding your dependent variable

Advice:

The cost of *what*?

- Take most precise sources available
- Report clearly how data were determined
- Where data were not directly measured, this is an important issue to be discussed under Limitations

The cost to whom?

- Take the broadest perspective available
- Where perspective is limited to specific parties, this is an important issue to be discussed under Limitations

2.1: Determining cost data

Understanding your dependent variable

Warning:

- Charges ≠ Costs

Further reading:

- ***For determining costs (what?), see VA HERC*** ***www.herc.research.va.gov/include/page.asp?id=determining-costs***

- ***For more detail on perspective (to whom?) and general principles in health economic evaluation, see papers by Russell; Weinstein; Siegel (JAMA, 1996) & book by Gold (1996)***

2.2: Standardizing cost data

$1in Time Square ≠ $1 in Alaska; $1in 1945≠ $1 in 2015

Where costs come from more than one site and/or more than one year, it is essential that raw data are standardized prior to analysis:

- Standardize by year using (for example) Consumer Price Index
- Standardize by region using (for example) Medicare Wage Index

E.g. Unadjusted average cost data from two hospitals (2001-2015):

	2001	2007	2015
New York, NY	**$9021**	**$10390**	**$11872**
Lexington, KY	**$6503**	**$7111**	**$7995**

2.2: Standardizing cost data

Standardize by year using Consumer Price Index

Consumer Price Index (using 1982 as 100; bls.gov):

2001: 177.1 2007: 207.3 2015: 233.7

Standardize data to a single year (usually final year of collection):

	2001			2007		
	Unadjusted	CPI	CPI-Adjusted	Unadjusted	CPI	CPI-Adjusted
NY	**\$9021**	/(177.1/233.7)	=**\$11904**	**\$10390**	/(207.3/233.7)	=**\$11713**
KY	**\$6503**	/(177.1/233.7)	=**\$8581**	**\$7111**	/(207.3/233.7)	=**\$8017**

Thus, all costs **in amber** are in 2015 dollars.

2.2: Standardizing cost data

Standardize by region using Medicare Wage Index

Medicare Wage Index (cms.gov):

NY: 1.3014 KY:0.8829

	2001			2007			2015		
	CPI-adjusted	MWI	Fully standardized	CPI-adjusted	MWI	Fully standardized	CPI-adjusted	MWI	Fully standardized
NY	**$11904**	/1.30	=**$9157**	**$11713**	/1.30	=**$9010**	**$11872**	/1.30	=**$9132**
KY	**$8581**	/0.88	=**$9751**	**$8017**	/0.88	=**$9110**	**$7995**	/0.88	=**$9085**

Thus, all costs **in green** are in 2015 dollars and standardized by geographical location, and may be pooled for analysis.

(Repeat for all years for which data were collected)

2.2: Standardizing cost data

Advice:

Always standardize cost data by year and region

- Bigger time spans & more sites = more important to standardize

Report methods of standardization in Methods

1. Introduction
2. **Considerations in data analysis**
    3. **Defining the sample**
3. Concluding remarks

2.3: Defining the sample

Appropriate approaches to utilization outliers and length of stay (LOS)

Healthcare utilization data are typically right-skewed

A complex minority of patients account disproportionately for:

- Admissions
- Hospital days
- Cost of care to insurers and health systems

Various strategies to simplify analysis are observable

- ‘Controlling for’ outlier status by using LOS as an independent variable
- Remove high-cost/long-stay outliers prior to analysis. E.g. estimate treatment effect for patients who stayed in hospital <=1 month

2.3: Defining the sample

Appropriate approaches to utilization outliers and length of stay (LOS)

However, there are good reasons not to

1. 'Control for' outlier status by using LOS as independent variable

- LOS is **<u>not</u>** an independent variable where utilization is the dependent variable!
- LOS is associated with both treatment (LOS = indicator of need) and outcome (LOS ≈ cost of stay)

2. Remove high-cost/long-stay outliers prior to analysis

- Estimated effects for a sample defined by outcome are not scientific (endogeneity) and not useful (we still have to pay for outliers)

2.3: Defining the sample

Appropriate approaches to utilization outliers and length of stay (LOS)

Advice:

Employ LOS a dependent variable. It is a utilization outcome that treatment can impact.

Never use LOS as an independent predictor either in regression on costs or as a covariate in propensity scoring. **This is an error.**

Never compare estimated effects of an intervention on utilization for different samples defined by LOS. **This is an error.**

2.3: Defining the sample

Appropriate approaches to utilization outliers and length of stay (LOS)

Advice:

Incorporating intervention timing may mitigate outliers (see 2.4)

In the presence of extreme high-utilization outliers distorting results, consider alternative strategies:

- Can outliers be identified by baseline data?
- Is latent class analysis appropriate?

Where extreme outliers remain a decisive issue in analysis, report results with and without these subjects

Further reading:

- ***For a detailed discussion of all points raised in '2.3', see May et al (2016a)***
- ***For an accessible use of latent class analysis, see Conway & Deb (2005)***

1. Introduction
2. **Considerations in data analysis**
    4. **Defining the treatment variable**
3. Concluding remarks

2.4: Defining treatment variable

The importance of timing

Palliative care is often not a default option:

- Patients referred to PCU or PCCT
  - ➢ Therefore, timing often differs between patients: some first receive PC on day 1, others on day 99

Utilization outcomes are **additive:**

- – If evaluating cost of an episode of care, costs accrued from the point of admission form part of the dependent variable
- – Ditto an evaluating of length of stay: each day from admission is in your outcome of interest

2.4: Defining treatment variable

The importance of timing

Therefore, timing is very important

A consultation (or PCU admission) on the 99th and final day of a hospital admission cannot impact utilization equally to an intervention on day 1

- Grouping all hospital-based PC in utilization analyses risks a false negative (May et al. 2016a)
- E.g. Does hospital-based PC impact LOS? Literature is not clear but has rarely included timing

2.4: Defining treatment variable

Appropriate approaches to utilization outliers and length of stay (LOS)

Advice:

Incorporate timing where appropriate

Think very carefully about how to do so (more complicated than it looks!)

Examples in the literature:

- Exclude later consults from analysis (May 2015; May 2016b)
- Interaction terms in regression (McCarthy 2015)
- Time from first PC to death (Scibetta 2016)

Some disagreement on validity of some methods

Further reading:

- *Papers cited above, or please contact me to discuss (peter.may@tcd.ie)*

1. Introduction
2. **Considerations in data analysis**
    5. **Choice of appropriate model**
3. Concluding remarks

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Distributions typically pose problems for statistical analysis:

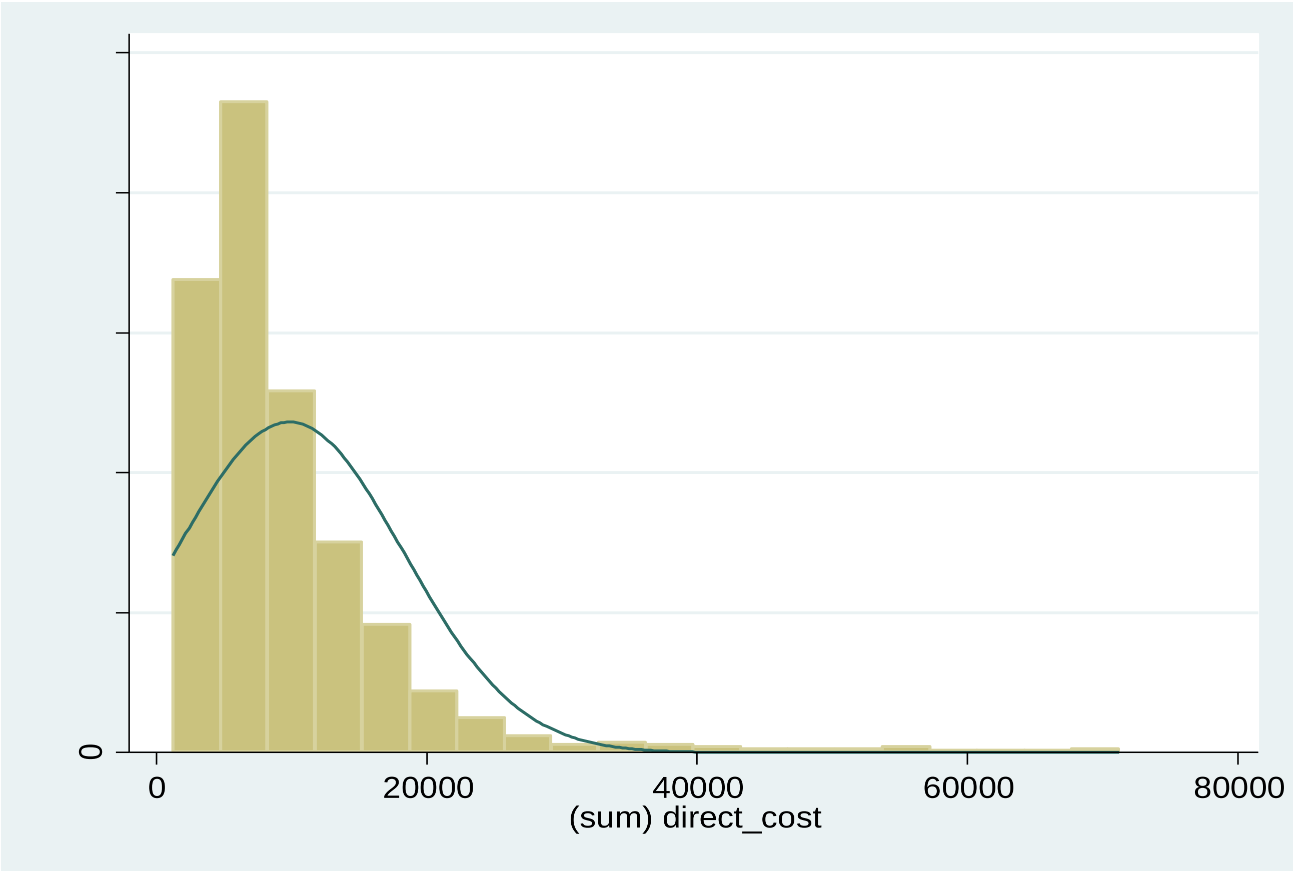
- **Non-negativity:** by definition never less than zero
- **Mass of zero-value observations:** in data drawn from populations, a large number of cost data-points will be zero
- **Positive skew:** a minority of patients incur a disproportionately high level of costs, skewing the distribution right
- **Heteroscedasticity:** variability of costs is unequal across a range of values for important predictors
- **Leptokurtosis:** clustering of cost observations for a large number of patients with similar care trajectories may result in high ‘peaked-ness’ of distribution

➢ **<u>Linear regression (OLS) is seldom appropriate</u>**

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Total direct cost of hospital admission



Skewness: 3.2

(0 for normal distribution)

Kurtosis: 17.7

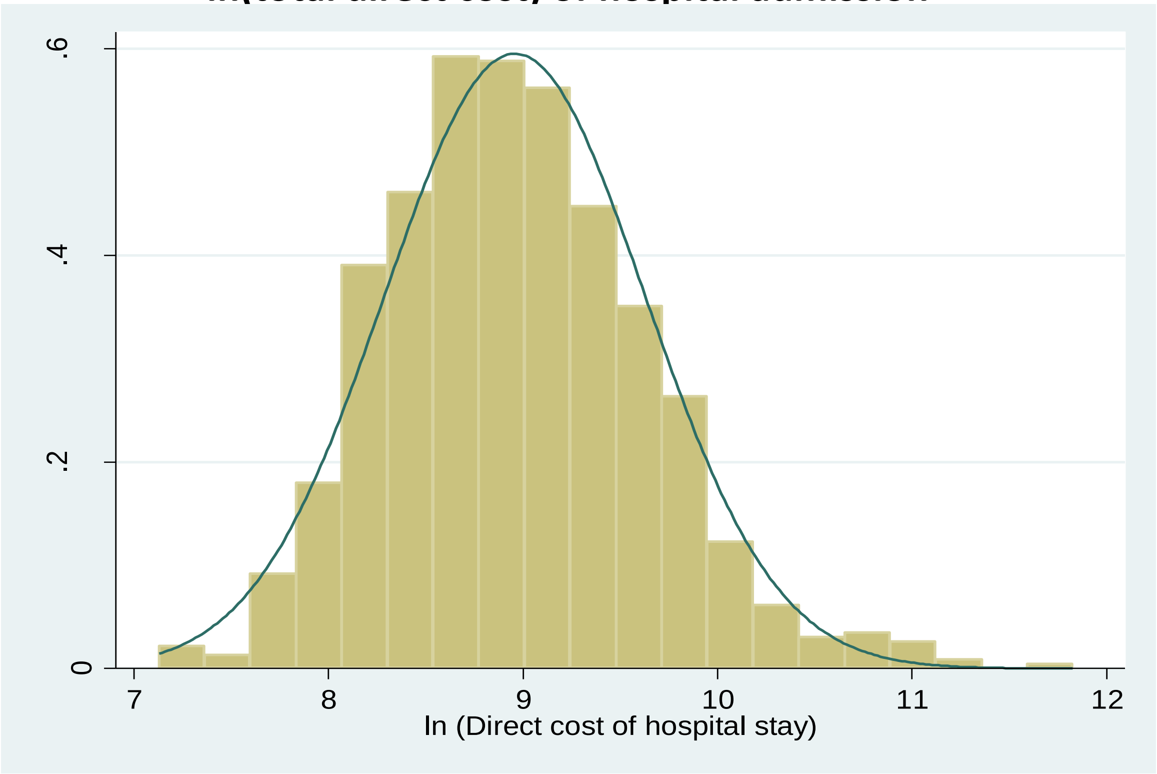
(3 for normal distribution)

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

The ‘old’ way to address this was log-transformation, which generally mitigates skew, heteroscedasticity & leptokurtosis

ln(total direct cost) of hospital admission



Skewness: 0.4

(0 for normal distribution)

Kurtosis: 3.4

(3 for normal distribution)

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

However, beware the 'retransformation problem':

"Although [log-transformed] estimates may be more precise and robust [than estimates using highly skewed distributions of untransformed costs], no one is interested in log model results on the log scale per se*.*

"Congress does not appropriate log dollars. First Bank will not cash a check for log dollars. Instead, the log scale results must be retransformed to the original scale so that one can comment on the average or total response to a covariate x*.*

"There is a very real danger that the log scale results may provide a very misleading, incomplete, and biased estimate of the impact of covariates on the untransformed scale, which is usually the scale of ultimate interest." - Manning (1998)

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Consider instead non-linear alternatives to OLS:

Generalized linear model

<u>Family</u>	<u>Link</u>
Gaussian	Identity
Poisson	Log
Gamma	Power
Inverse Gaussian	

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Consider instead non-linear alternatives to OLS:

Generalized linear model

<u>Family</u>	<u>Link</u>
Gaussian	Identity
Poisson	**Log**
Gamma	Power
Inverse Gaussian	

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Consider instead non-linear alternatives to OLS:

Generalized linear model

Family	Link
Gaussian	Identity
Poisson	Log
Gamma	Power
Inverse Gaussian	

Exponential conditional mean models

Generalized gamma models

Extended estimation equations

Finite mixture models

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Software is freely available online to evaluate model performance:

- For GLMs only, Stata *glmdiag.do* from UPenn (http://www.uphs.upenn.edu/dgimhsr/stat-cstanal.htm)
- For all models, Stata *AHE_2ed_Ch_3&12.do* from University of York (http://www.york.ac.uk/economics/postgrad/herc/hedg/software/)

- These test the appropriateness of specific models to a given distribution
- No model is dominant
  - Evaluating models prior to analysis is essential to maximize accuracy of estimated effects

2.5: Choice of appropriate model

Awkwardness of healthcare utilization data

Advice:

- Consider and describe data carefully prior to analysis
- Avoid use of OLS and OLS ln(y) with healthcare utilization data
- Consider nonlinear alternatives
  - Use available software to understand and evaluate options
  - Report briefly this process in Methods

Further reading:

- *The York* .do *file accompanies a book: Jones et al. (2013a)*
- *For an overview of why model choice matters, see Jones (2010)*
- *For more technical analyses, see Jones et al. (2013b); Garrido et al. (2012)*
- *Again, I am happy to help if I can (peter.may@tcd.ie)*

1. Introduction
2. Considerations in data analysis
3. **Concluding remarks**

Concluding remarks

Analyzing healthcare cost data

Utilization data are not always simple

- Challenges in statistical analysis
- Careful organization and interpretation required

1. Clarify & understand what $$$ data are
2. Standardize cost data for year and region
3. Consider impact of extreme outliers
4. Consider how you define your treatment/exposure
5. Move beyond linear regression in estimating effects

Concluding remarks

Analyzing healthcare cost data

Caveat: The guidance discussed here is far from comprehensive

- Additional complications in cost analysis
- 'Full' economic evaluation also incorporates patient & family outcomes

Evidence on utilization is

- Essential to maximize provision of effective care
- Sparse in the field of palliative and hospice care
  - Opportunities for high-impact studies

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You for your attention

E: peter.may@tcd.ie

References

References

Conway & Deb. 2005. J Health Econ. (24): 489–513.

Garrido et al. 2012. Health Serv Res, 47, 2377-97.

Gold et al. 1996. *Cost-Effectiveness in Health & Medicine*. New York: OUP.

Jones. 2010. ‘Models for health care’, Working Paper 10/01, HEDG, University of York.

Jones et al. 2013a. *Applied Health Economics*. 2nd ed., Oxford: Routledge.

Jones et al. 2013b. ‘A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare’, Working Paper 13/30, HEDG, University of York

Manning. 1998. J Health Econ. (17): 283-95.

May et al. 2015 J Clin Oncol. 33(25):2745–52.

May et al. 2016a. Health Serv Res [in press] . ‘Using length of stay to control for unobserved heterogeneity when estimating treatment effect on hospital costs with observational data: reliability, robustness & usefulness’

May et al. 2016b. Health Affairs, 35, no.1 (2016):44-53.

McCarthy et al. 2015. Health Serv Res.50(1):217–36.

References

continued

Russell et al. 1996. JAMA. Oct 9;276(14):1172-7.

Scibetta et al. 2016. J Palliat Med. Jan;19(1):69-75

Siegel et al. 1996. JAMA. Oct 23;276(16):1339-41.

Weinstein et al. 1996. JAMA. Oct 16;276(15):1253-8.

Appendix

Cost-consequence analysis