Epidemiology 212, Spring 2016
Publishing and Presenting Clinical Research

Lecture 2: Figures & Tables

Visual Display of Data

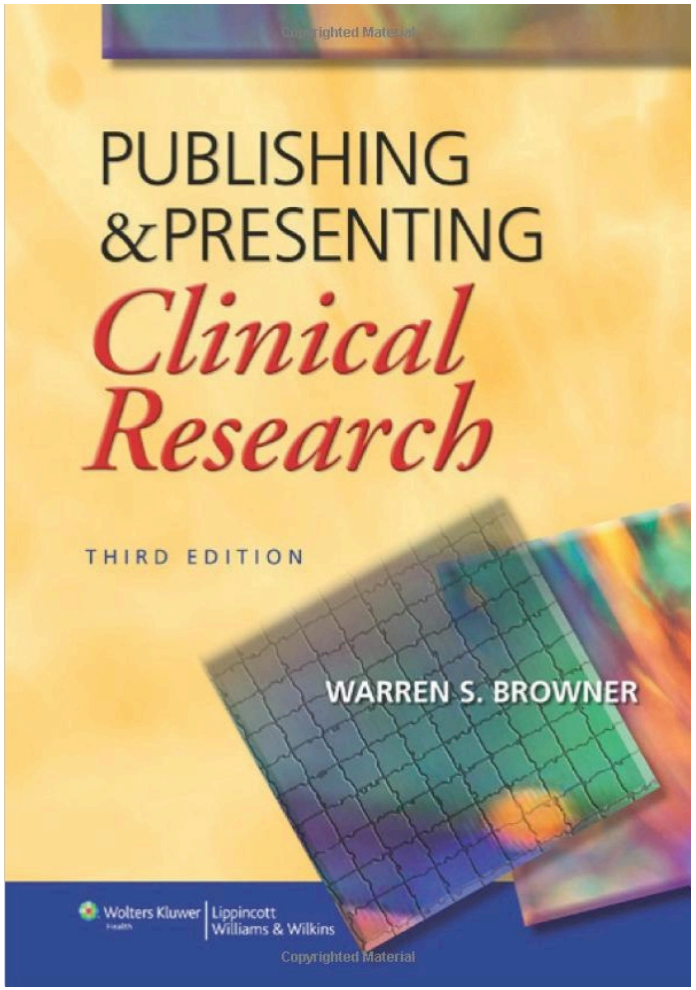
April 19, 2015

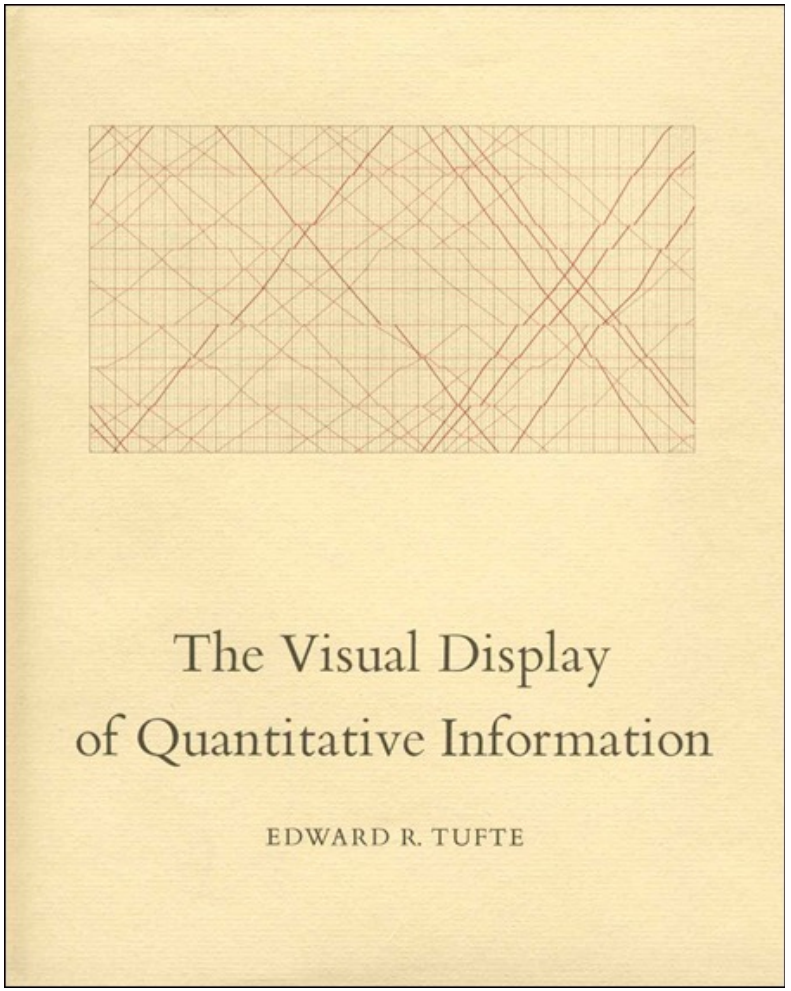
Vivek Jain, M.D., M.A.S.
Assistant Professor of Medicine
HIV/AIDS Division
San Francisco General Hospital, UCSF

Goals

- Highlight points from Warren Browner: *Publishing and Presenting Clinical Research, 4th Edition*
- Explore the best use of tables and figures to graphically display quantitative data
- Conveying maximum information in easiest format to understand
- Design considerations for different displays
- Making design itself invisible
- Illustrations

Books


PUBLISHING
&PRESENTING
Clinical
Research
THIRD EDITION
WARREN S. BROWNER
Wolters Kluwer
Lippincott
Williams & Wilkins
Copyrighted Material


The Visual Display
of Quantitative Information
EDWARD R. TUFTE

Tables

- When to use:
  - display comparison
  - enumerate a list that cannot be in text form
- When not to use
  - when data dimensionality >2 and can't be easily handled
  - when longitudinal data is better shown as a line, graph, etc. rather than numbers

Tables: design considerations

- Title: describe succinctly but with detail—not too little, not too much

Table title	Comments
Table 1. Baseline Characteristics of the Participants	Standard Table 1 title; Could include n=X sample size
Table 2. Clinical Characteristics of the Patients at Baseline	Could include n=X sample size
Table 2. Adverse Events up to Day 91.	Adverse clinical events up to Day 91 in aerosolized vs. subcutaneous vaccine groups
Table 3. Description of the air pollutants at the 39 schools	Great
Table 5. Population and school characteristics by school traffic	Great
Table 6. Difference in cognitive development, at baseline and 12-mo change, by school air pollution exposure (high/low group or interquartile range increase) in 2,715 children and 10,112 tests from 39 schools.	Changes in cognitive development by air pollution exposure at 39 schools (n=2,715)

Low et al. *NEJM*, 2015; Robinson et al., *NEJM*, 2015; Sunyer et al. *PLOS Medicine*, 2015

Tables: design considerations

- **The eye reads across → set up comparisons horizontally**
  - often, table can make sense both ways
  - decide what is the primary item to be compared
- **Use horizontal lines** to ease/guide interpretation
- **Avoid using any vertical lines**; use only horizontal lines

- **Labeling**
  - Label columns and rows in good size font
  - Give units
  - Be clear on groups, no acronyms
  - Avoid "Arm A/Arm B", or "Intervention/Control", "Group 1/2", etc.
  - Eliminate footnotes, asterisks, p-values in the footer
  - Think about centering (+/-)
  - Merge cells and group labels

- In manuscripts, don't double space tables.. print all on one page for ease
- Make one or two tables formatted the way you like, then re-use these

Example: not easy to look at

Improve the title

Comparison of Selected Characteristics of the Two Groups of Subjects.

Say what intervention was; include N

Label as control & include N

Give effect size & CI, then p-value

Female row duplicates info

Acronym

Characteristic	Group I	Group II	P
Male	265/401= 66%	197/402= 49%	*
Female	136/401= 34%	205/402= 51%	*
CHD	43/401= 11%	48/402= 12%	NS
Income in dollars ($)	$59,323 ± $13,358	$61,482 ± 16,552	NS
Age (years)	48 ± 16	49 ± 17	NS
<30 years	80/401 (20%)	84/402 (21%)	NS
31-65 years	45%	45%	NS
>65 years	35%	34%	NS
*p < 0 .05. NS = nonsignificant.			

Give p-value right here not in footnote

Use a sans serif font

Eliminate vertical lines

Improved Version

Table 1. Baseline clinical and socioeconomic characteristics of exercise intervention (n=401) and no-exercise control (n=402) participants.

Characteristic	Exercise Intervention (N = 401)	No-exercise Control (N=402)	P-value
Male, n (%)	265 (66)	197 (49)	<0.01
Coronary heart disease, n (%)	43 (11)	48 (12%)	0.6
Median age (years ± IQR)	48 ± 16	49 ± 17	0.4
<30 years, n (%)	80 (20)	84 (21)	<0.001
31-65 years, n (%)	180 (45)	181 (45)	0.5
>65 years, n (%)	140 (35)	137 (34)	0.5
Median 2014 Income ($ ± IQR)	$59,323 ± $13,358	$61,482 ± 16,552	0.7

Tables of Regression Output

- Predictors: describe, and display as: "age (per 10 years)", e.g.
- No regression coefficients
- Favor 95% CI's over p-values (but often must include both)
- Problematic version:

TABLE 6.24 Independent Predictors of Coronary Heart Disease among 2,124 Middle-aged Subjects, Using Logistic Regression Models*

Predictor	Regression Coefficient	Standard Error	*P*
Sex	0.51	0.22	0.01
Age	0.05	0.01	< 0.0001
Serum cholesterol	0.3	0.15	0.05
Systolic blood pressure	0.7	0.3	0.02
Smoking	1.1	0.3	< 0.0001

* This is not a good table;Table 6.25 is much better.

Tables of Regression Output

- Improved version:

TABLE 6.25 Independent Predictors of Coronary Heart Disease among 2,124 Middle-aged Subjects

Predictor	Relative Risk*	95% Confidence Interval	*P*
Male sex	1.7	1.1 – 2.6	0.01
Age (per 10 years)	1.6	1.4 – 2.0	< 0.0001
Serum cholesterol (per 20 mg/dL)	1.3	1.0 – 1.8	0.05
Systolic blood pressure (per 10 mm Hg)	2.0	1.1 – 3.6	0.02
Current smoker (vs. never smoked)	3.0	1.7 – 5.4	< 0.0001

* Relative risks approximated with odds ratios from a logistic regression model.

For discussion

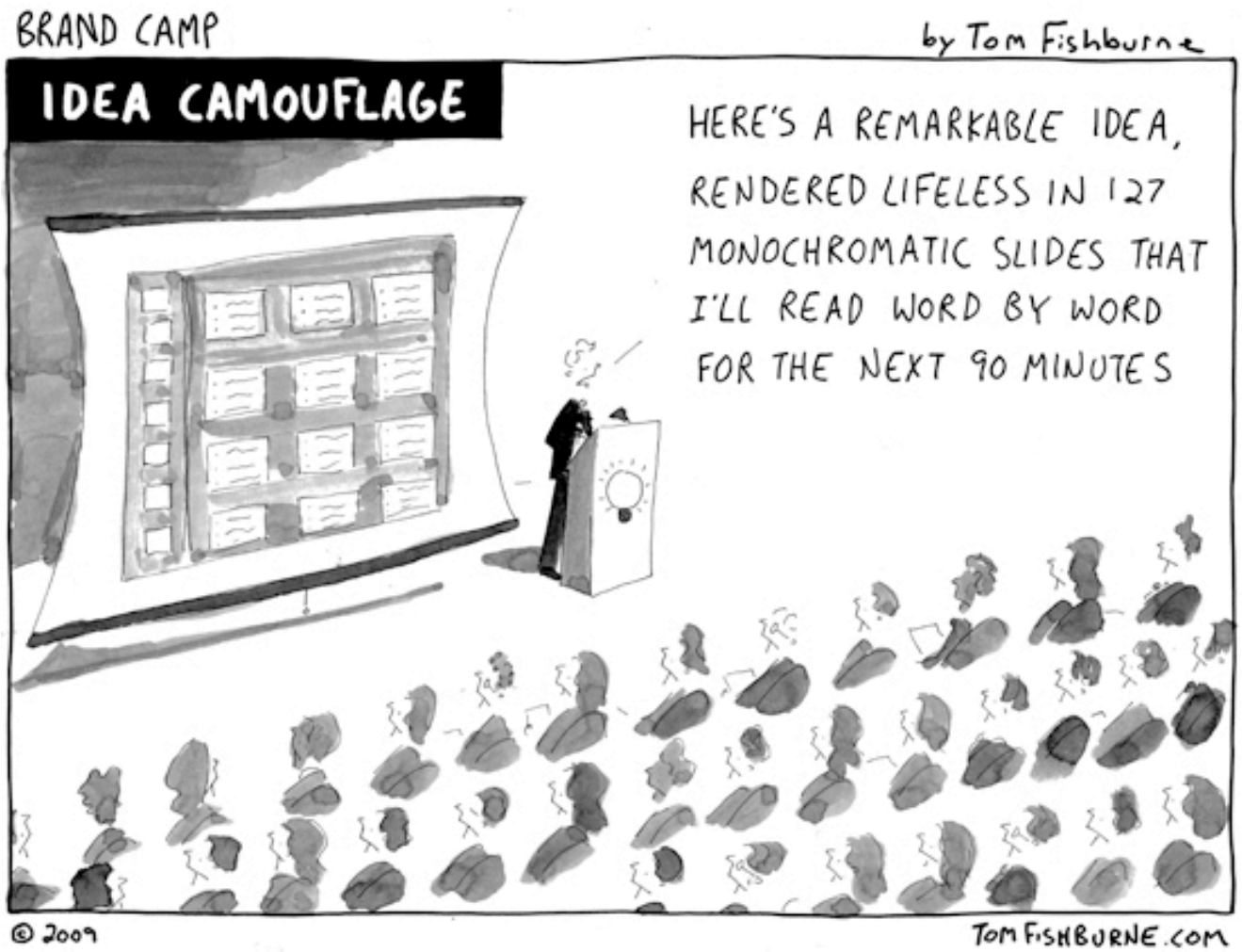
- Good title
  Overall organization good
  Categorical variables clear
  Referent is very clear
  OR=1.0 included (or can say ---)

- Vertical lines
  Needs top and bottom lines
  Too many thick lines
  Could separate OR from 95% CI
  P-values could go one decimal, especially the ones <0.05
  "Non-academic" description → footnote
  No need to bold the significant p-values

Table 5. Adjusted odds ratios of prescribing prophylactic antibiotics for confirmed postnatal hydronephrosis.

	Odds Ratio (95% Confidence Interval)	p value
Geographic region		
Northeest (reference group)	1.0	
Midwest	1.68 .64-4.43	.293
South	1.08 .49-2.39	.840
West	.530 .237-1.18	.0121
Practice type		
Non-academic (reference group) includes private practice, community clinic, hospital-based, managed care, and other)	1.0	
Academic	1.80 .58-5.60	.312
No. of cases encountered per year		
0-5 (reference group)	1.0	
≥6	1.29 .64-2.6	.48
Years in practice (for every 1 year increment)	1.015 .98-1.048	.35
Belief from literature of antibiotic effects		
Not up-to-date (reference group)	1.0	
Equal benefit and harm	3.29 1.67-6.48	**.001**
Beneficial	6.18 2.43-15.71	**<.0005**
Harmful	3.20 .84-12.19	.088
Specialist consultation readily available		
Yes(reference group)	1.0	
No	7.6 1.4-41.0	**.018**

Figures: Reveal your Data



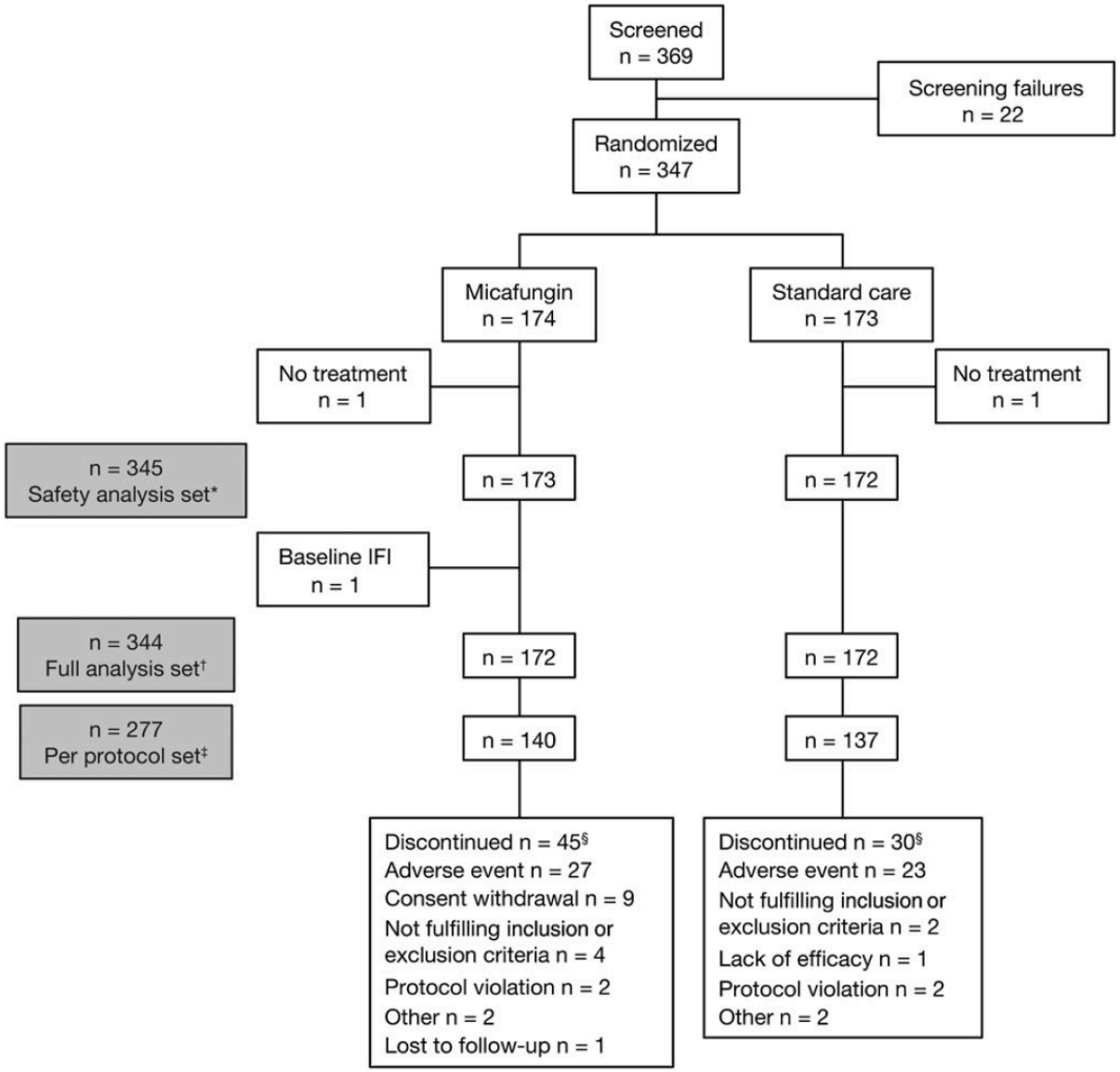
http://tomfishburne.com.s3.amazonaws.com/site/wp-content/uploads/2011/12/090309.ideacamo.jpg

Figures: General Principles

- **When to use**
  - demonstrating overall effects, comparisons
  - showing trends, longitudinal variation
- **When not to use**
  - can data be displayed in text form? or as a table?
  - when precise numerical display is needed
- **Design features**
  - dimensionality of data
  - "data-to-ink" ratio: always want to maximize this
  - use a simple design, large and easy-to-read labels
  - ensure all axes are labeled, use sans serif fonts for everything
  - sketch out by hand before making the graphic on computer

Diagrams

- **Study subject flow, CONSORT diagrams**
- PowerPoint versatile; Word can make these but not as well
- Note **vertical** vs. horizontal flow
- Generally want to avoid unnecessary "boxing" of words, but here it is helpful…



Saliba et al. *Clin Infect Dis*, 2015

Scatterplots

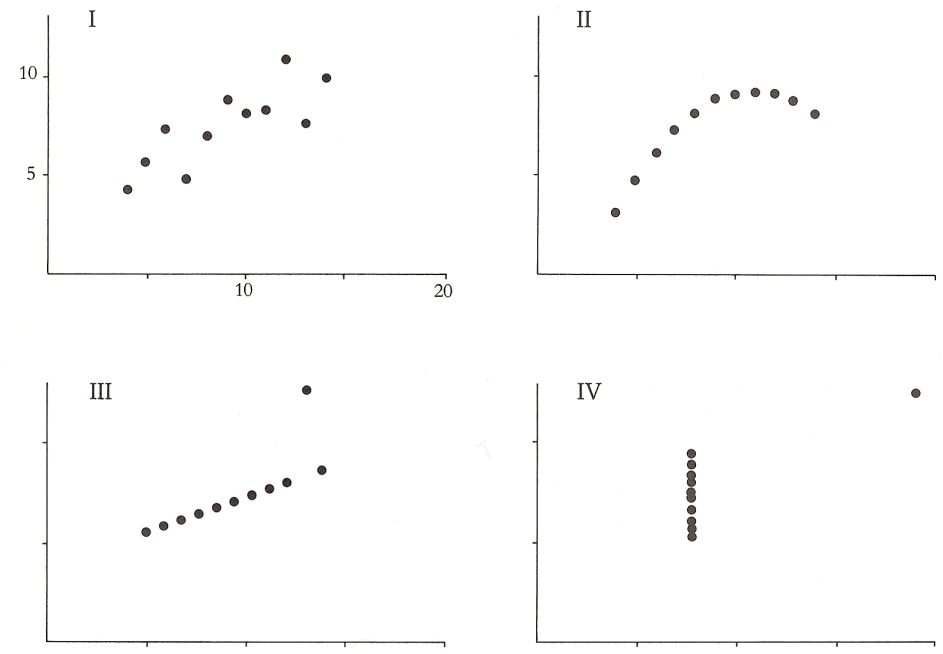
- Often very useful
  - shows all discrete data points, including interesting outliers
  - essential first step in examining data

Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations. Consider Anscombe's quartet: all four of these data sets are described by exactly the same linear model (at least until the residuals are examined).

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

And yet how they differ, as the graphical display of the data makes vividly clear:

F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (February 1973), 17-21.

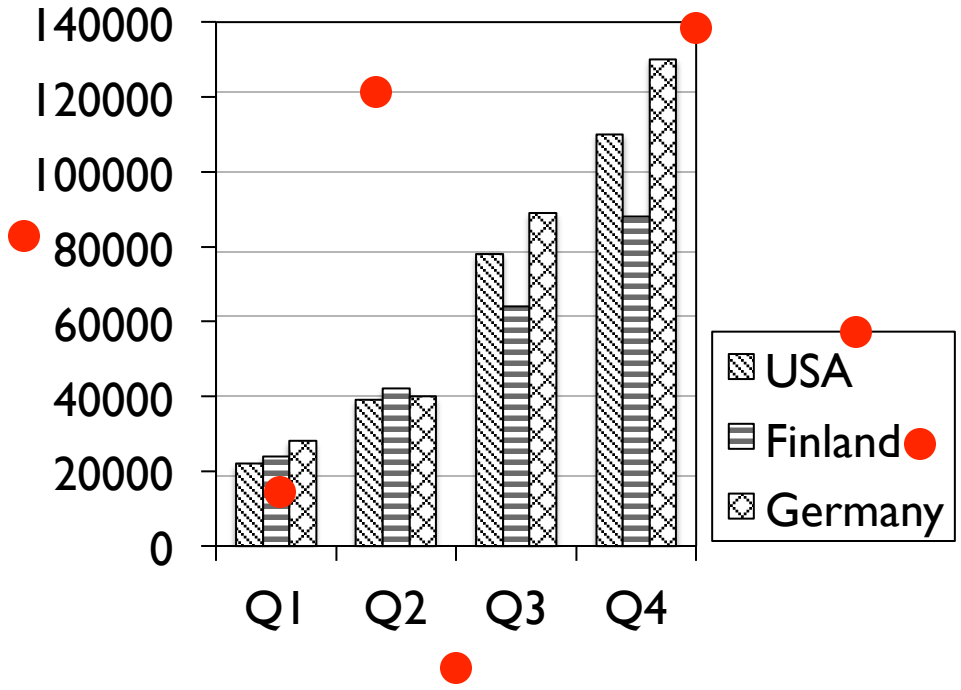


Tufte, E, *Visual Display of Quantitative Information*, 2nd Ed., p.14-15

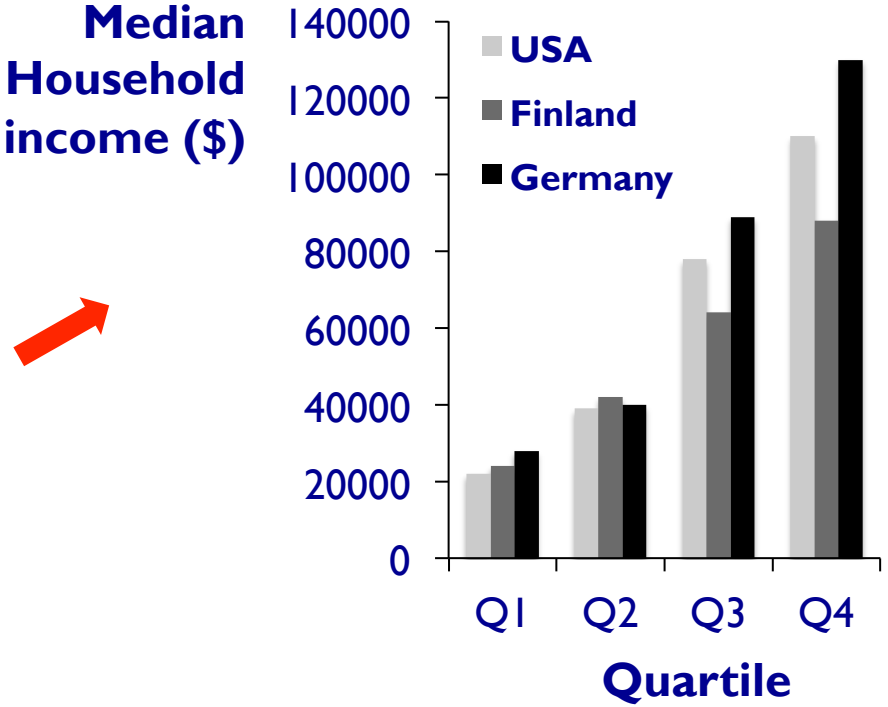
Bar Graphs

- Ubiquitous, often very useful
- Stacked bar: useful for showing proportions, trends
  - be careful with y-axis
- Avoid color unless needed; avoid moire effects: use grayscales

Original

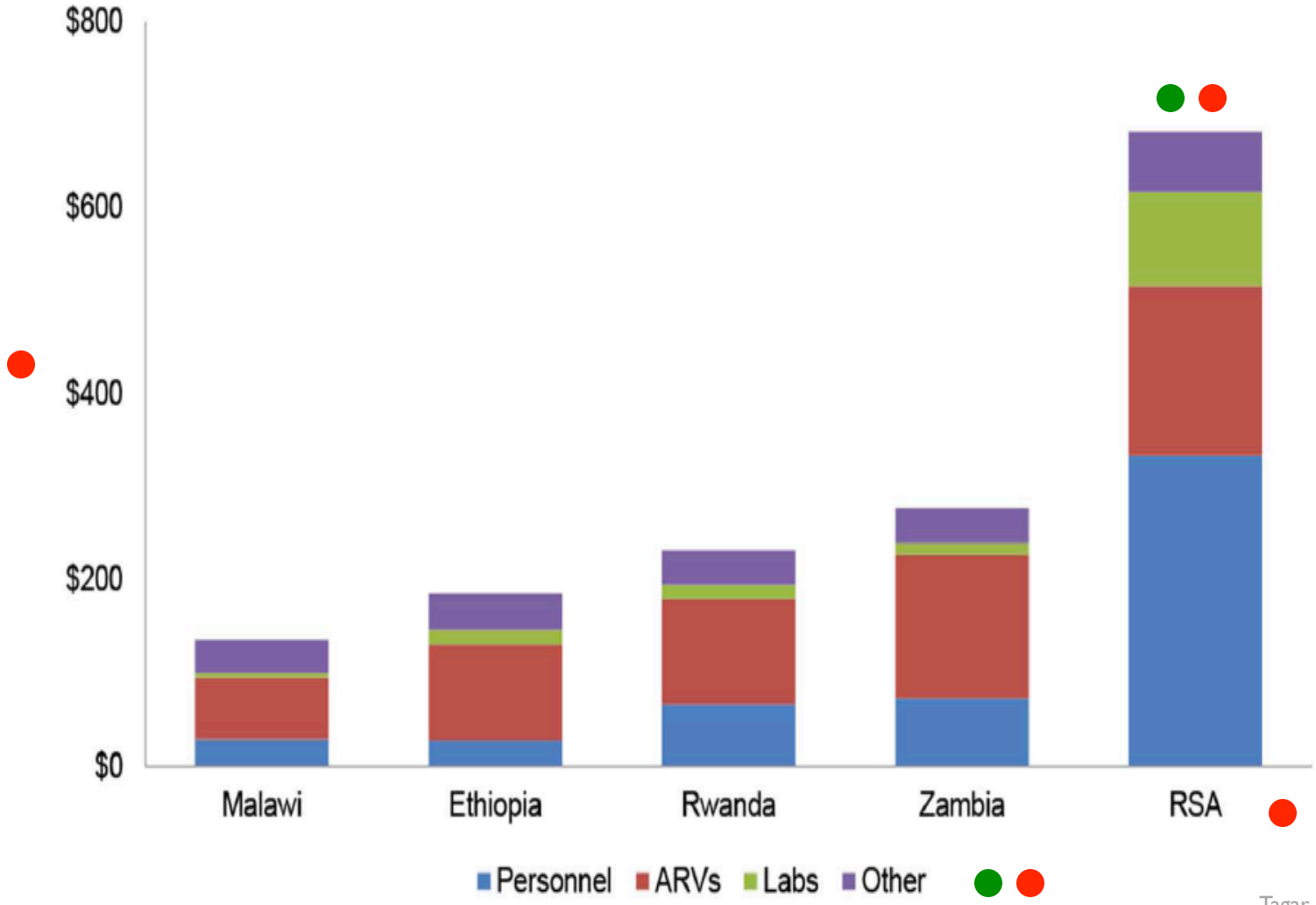


Improved



Stacked Bar 1: components

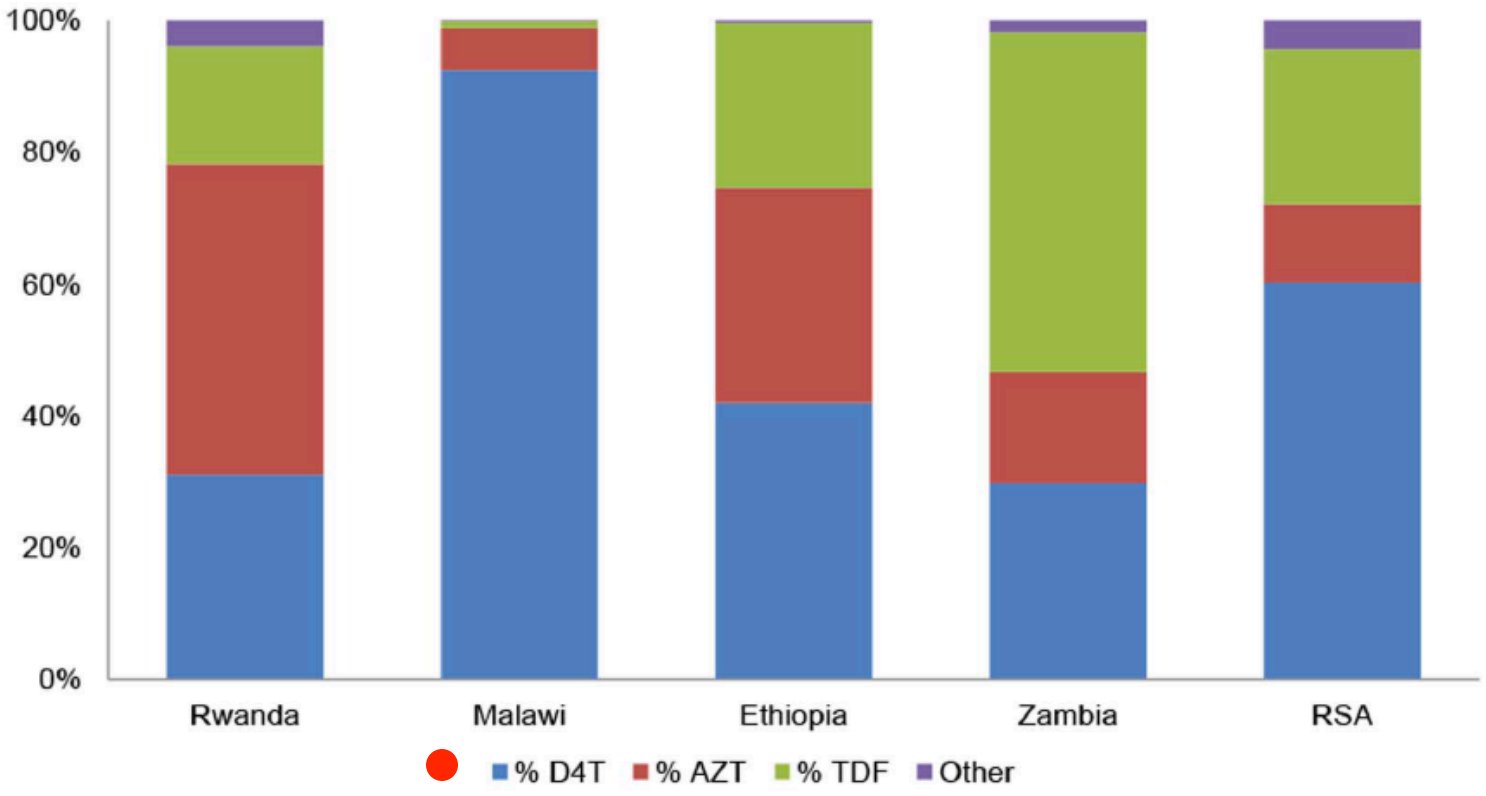
Figure 2. Simple Average Cost per ART Patient-Year by Country by Cost Category (USD).
doi:10.1371/journal.pone.0108304.g002



Tagar et al. *PLOS ONE*, 2014

Stacked Bar 2: adds to 100%

Figure 3. Regimen Selection and Average First Line ARV Costs by Country. Regimen breakdowns are from the cost data year, and do not reflect changes in regimen selection that countries have made since 2009–2011, such as the continued phase-out of d4T.
doi:10.1371/journal.pone.0108304.g003

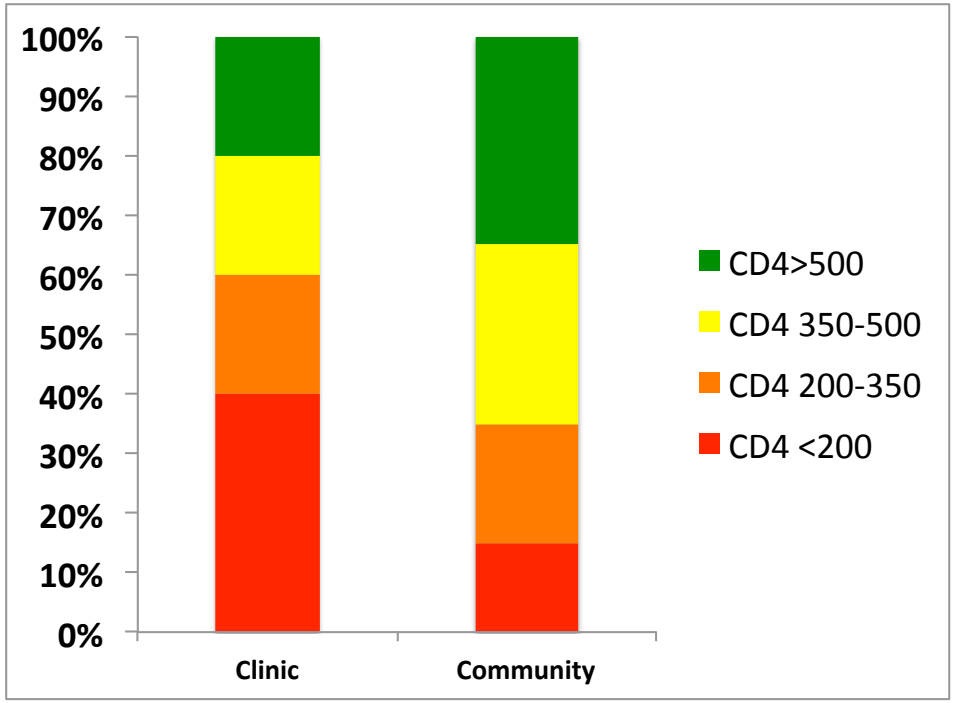


	Rwanda	Malawi	Ethiopia	Zambia	RSA
Avg. Adult 1st Line ARV Cost	$119	$68	$105	$147	$160

Tagar et al. *PLOS ONE*, 2014

Table vs. Stacked Bar Graph

CD4+ Group (cells/uL)	Clinic (n)	Community (n)
>500	200	350
350-500	200	300
200-350	200	200
<200	400	150



Simulated data

Small Multiples

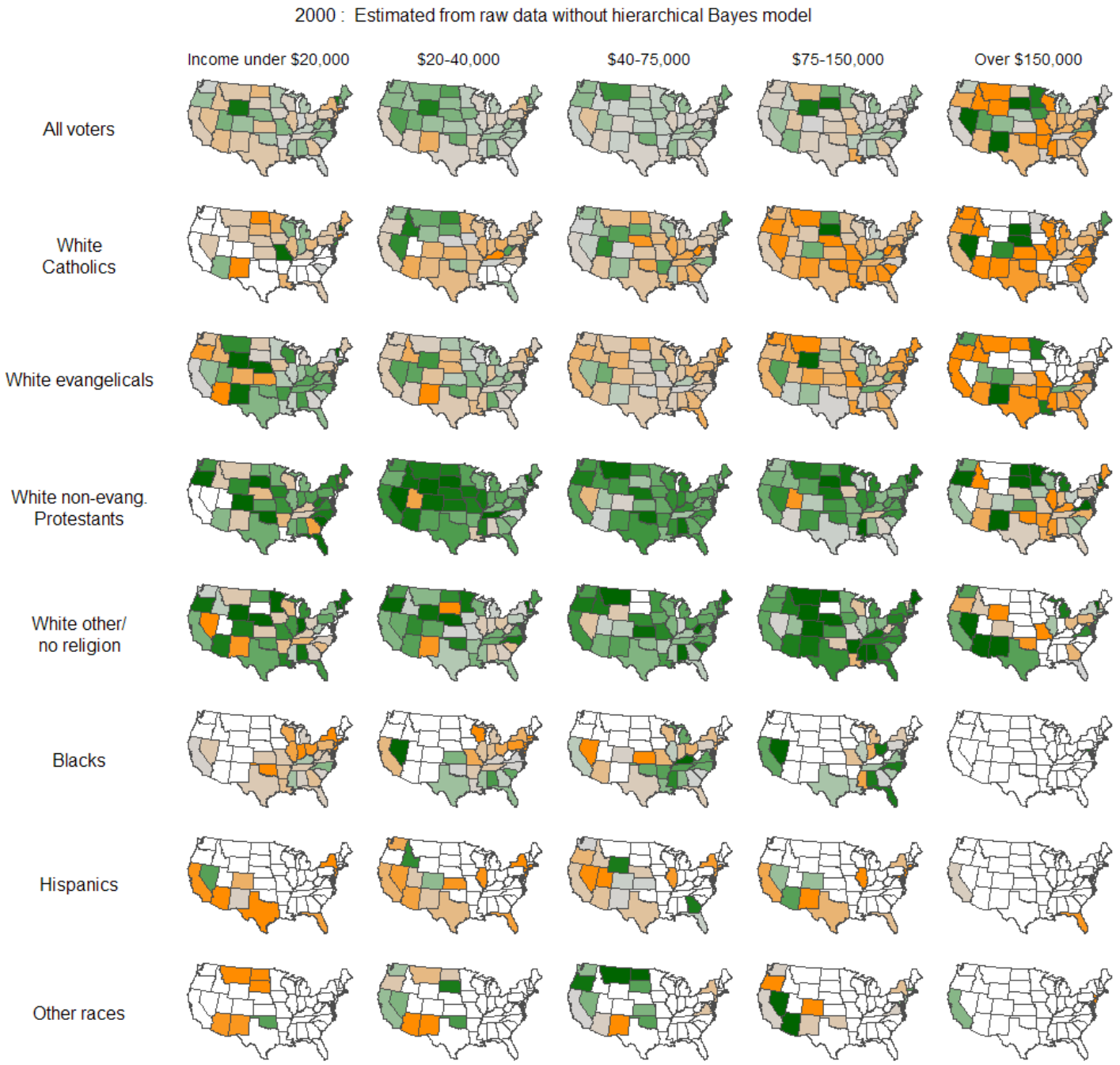
The eye trains on the first complex image; rest of content is then rapidly absorbed.

Orange = *more* in favor of school voucher program compared to 45% national average

Green = *less* in favor of vouchers than average

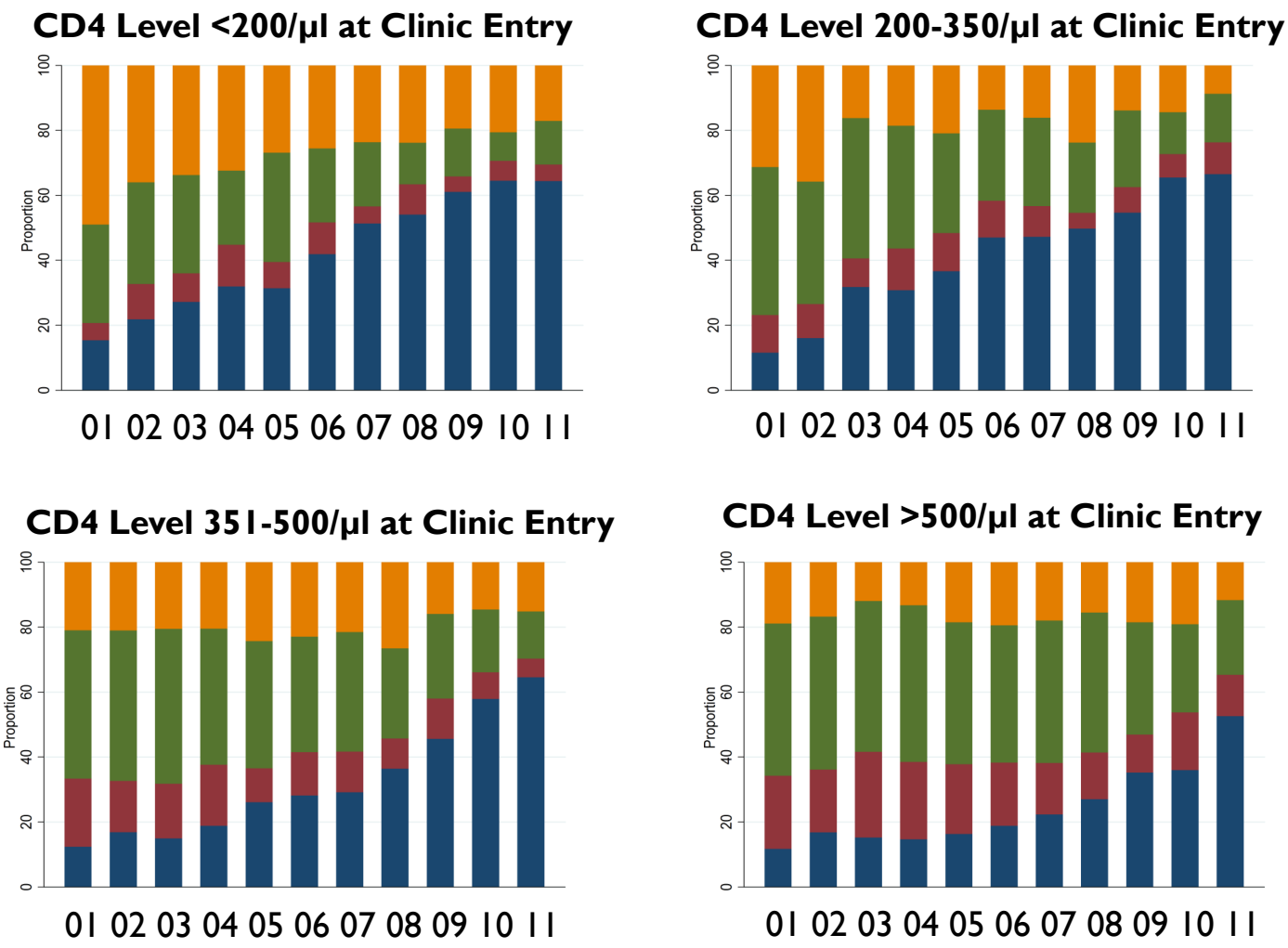
- Overall + 7 subgroups
- 5 income levels
- 50 US states
- 8 levels of support

→ very rich information



Compared to the Bayes maps, these are very noisy, and it is difficult to try to interpret the patterns.

Stacked Bar Graph with Small Multiples: Dense Information



Graph shows:

1) rising viral suppression of HIV over a decade (blue)

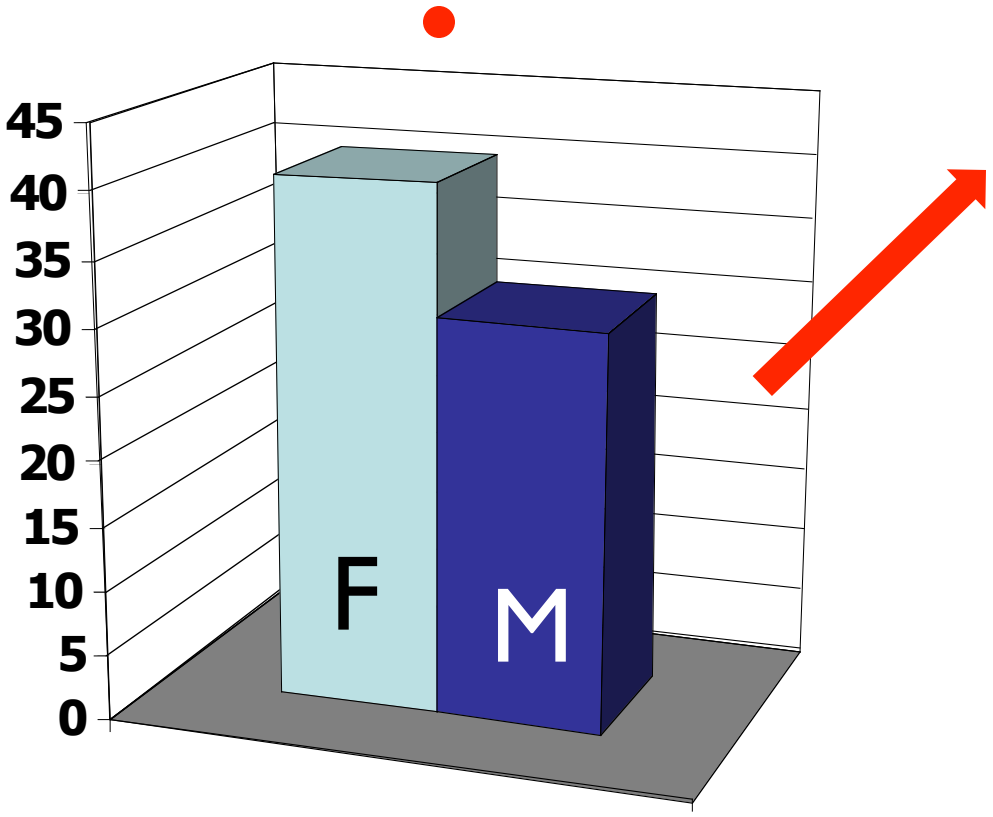
2) this occurs in 4 different patient groups (CD4 counts)

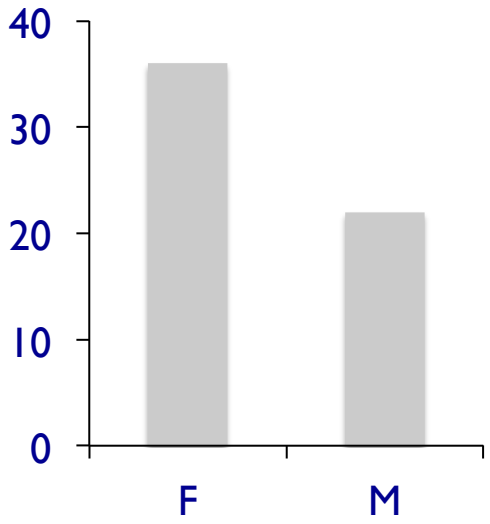
3) in high CD4+ patients, this occurs later, when guidelines changed favoring treatment in this group.

Geng et al. *Clin Infect Dis*. 2012.

3D Bar Graphs: rarely indicated

Fig. 1. Proportion of men and women who reported voting



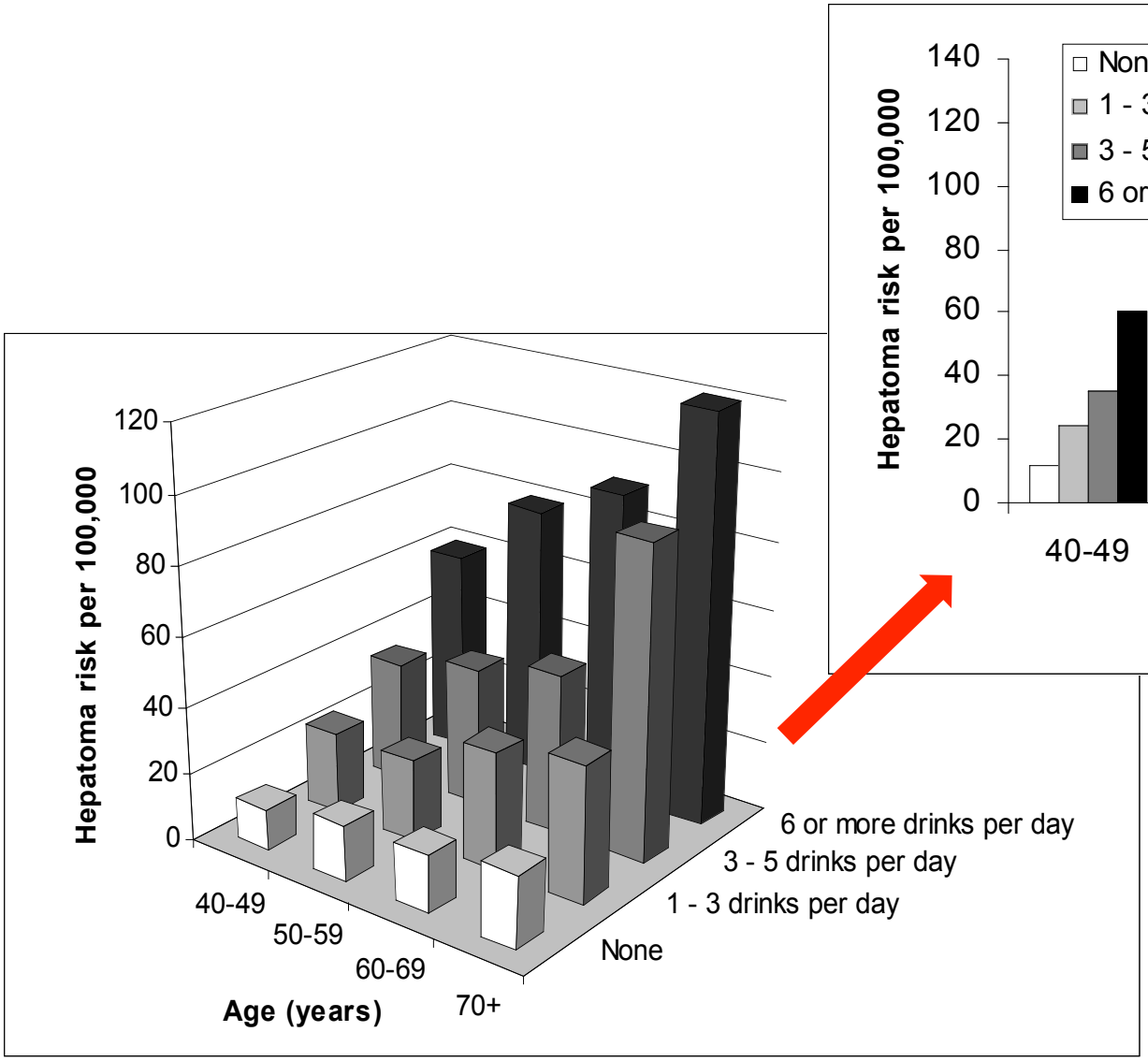


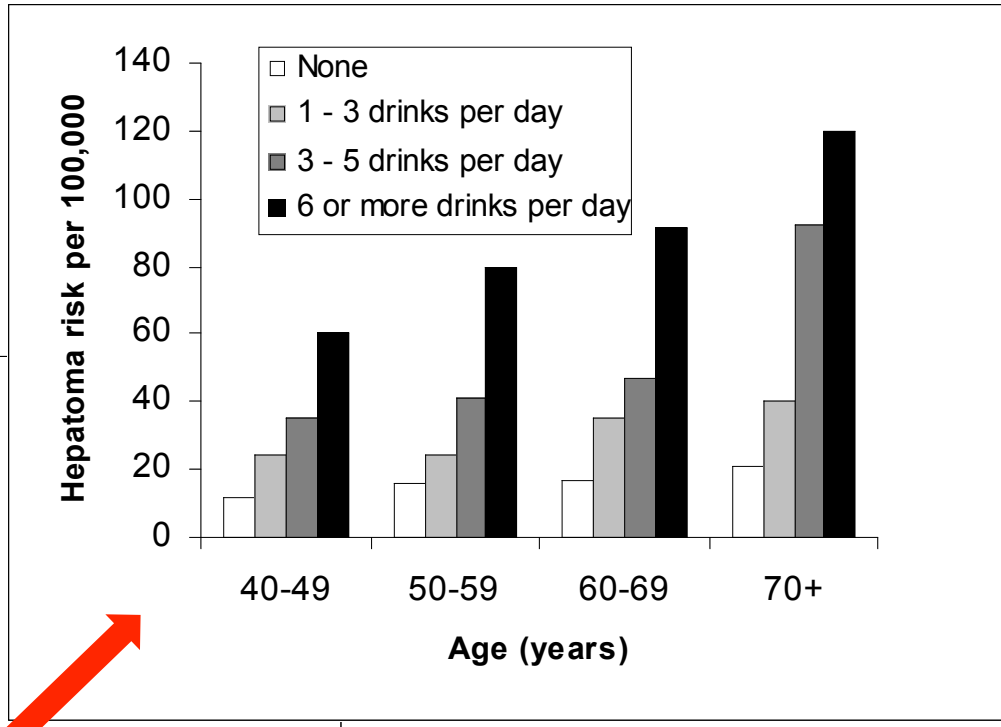
Female **Male**

36% 22%

"Overall, 36% of women and 22% of men reported voting."

3D Bar Graphs: Better Example





Stacked Line Graph

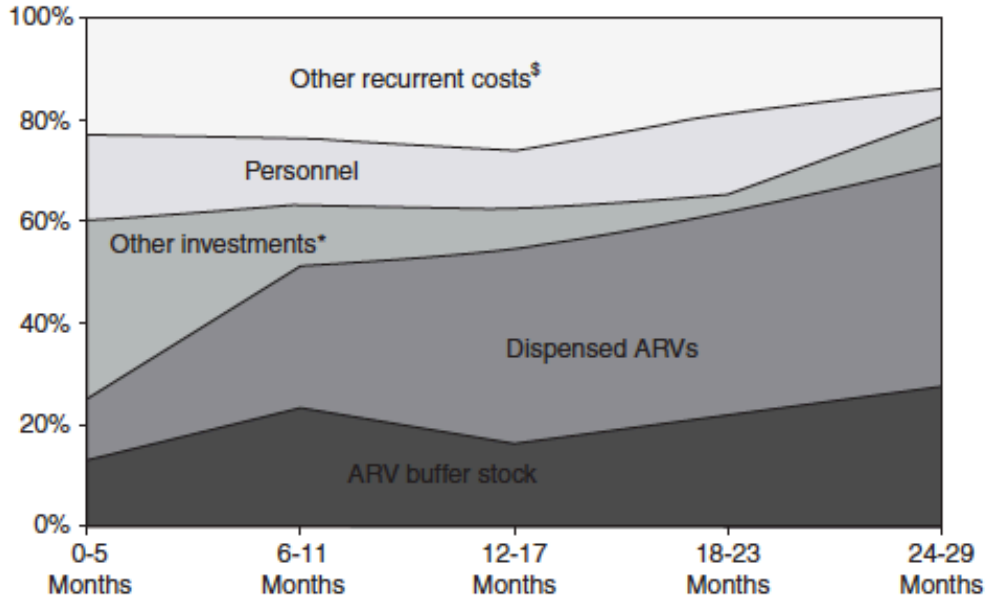
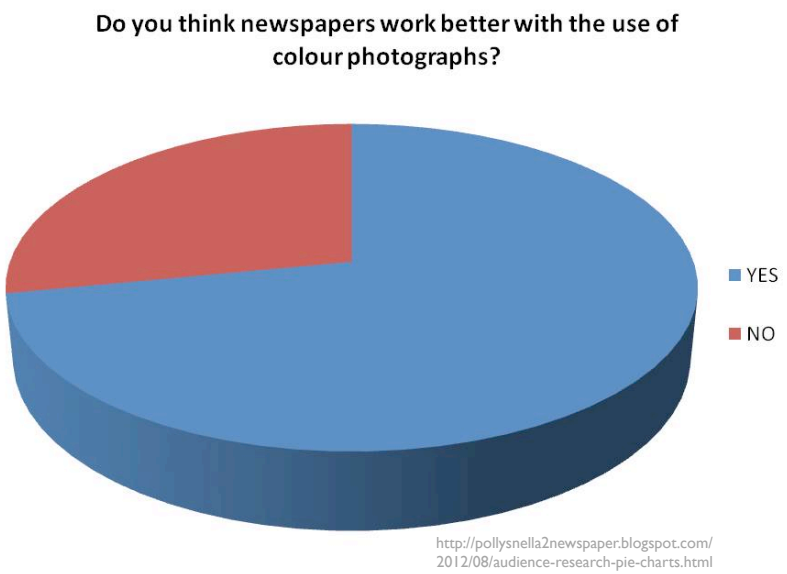


Fig. 3. Distribution of total financial costs in successive 6-month periods, from start of HIV treatment scale-up in each site through 2006–2007[δ]. *Other investments include construction/renovation, equipment and training. [$]Other recurrent costs include non-antiretroviral (ARV) medications, laboratory supplies, other supplies, building rental, travel expenses, utilities and contracted services. [δ]Figure excludes Botswana sites.

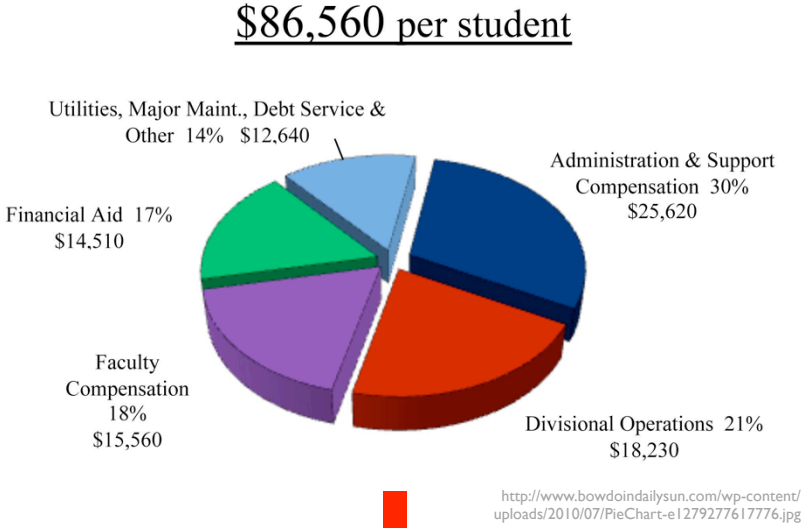
Menzies et al. *AIDS*, 2011

Pie Charts: rarely indicated



http://pollysnella2newspaper.blogspot.com/2012/08/audience-research-pie-charts.html

"74% of respondents said they preferred newspapers with color photos."



http://www.bowdoindailysun.com/wp-content/uploads/2010/07/PieChart-e1279277617776.jpg

Component	Cost
Financial Aid	$15,560
Division operations	$18,230
Administrative compensation	$25,620
Debt service	$12,640
Financial aid	$14,150

Line Graphs

- Very useful for showing longitudinal data: within-subject or within-group data over multiple timepoints

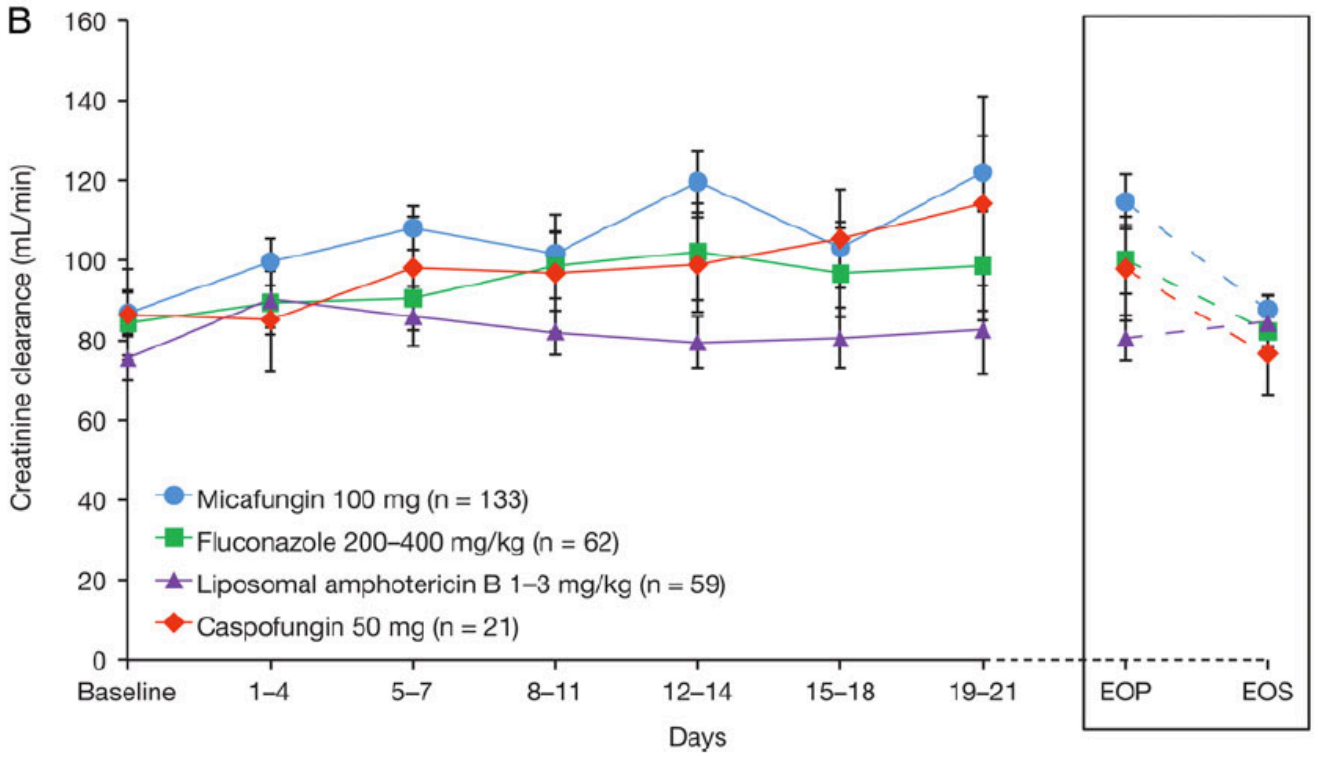
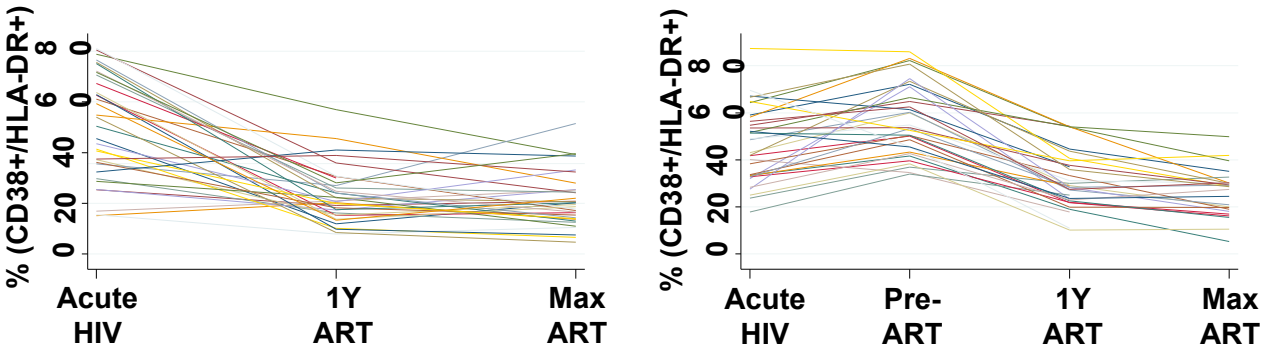


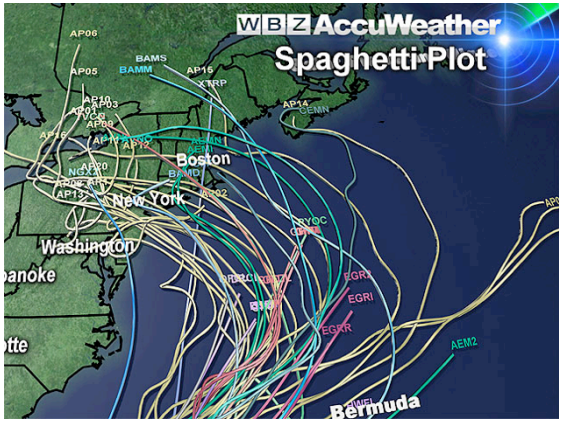
Figure 3. Mean glomerular filtration rate (*A*) and mean creatinine clearance (*B*) in patients not requiring renal replacement therapy, according to specific study drug. Abbreviations: EOP, end of prophylaxis; EOS, end of study.

Saliba et al. *Clin Infect Dis*, 2015

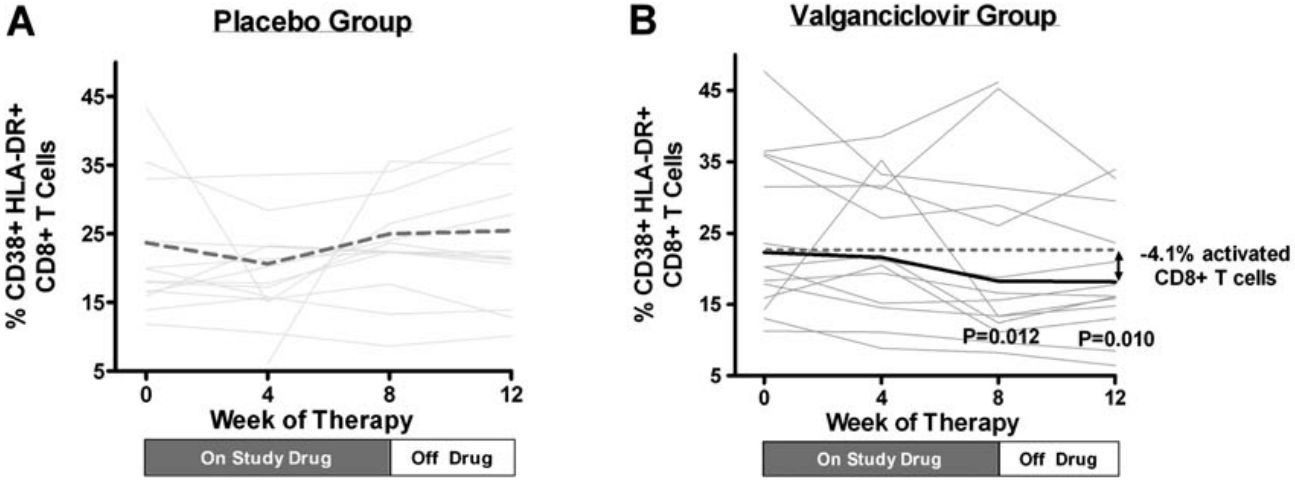
Spaghetti Plots



Jain et al., *JID*, 2013



https://cbsboston.files.wordpress.com/2012/10/spaghetti.jpg



Hunt et al., *AIDS*, 2011

Box Plots & Christmas Tree Plots

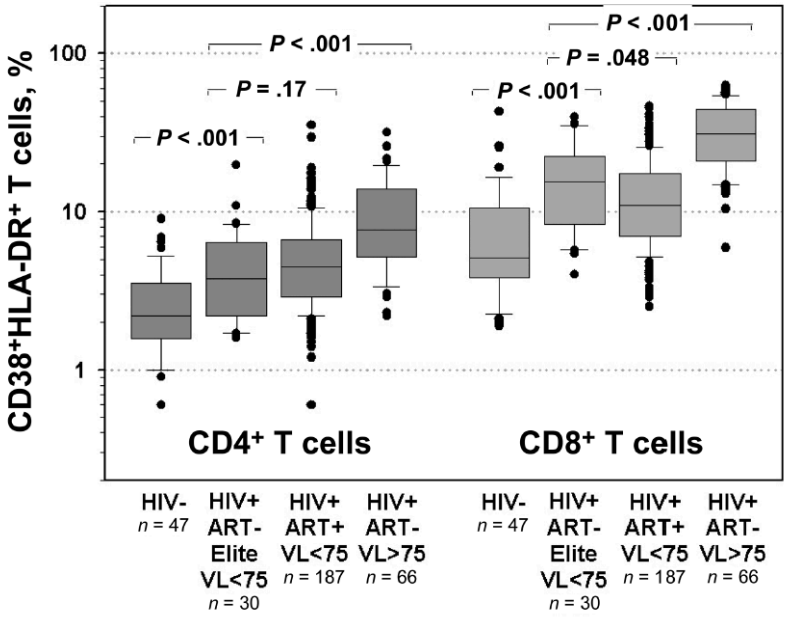
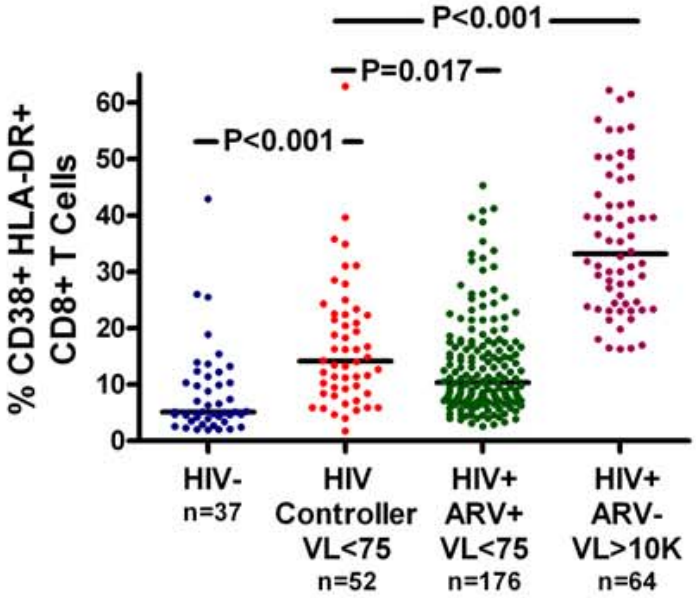


Figure 1. T cell activation among HIV-uninfected and HIV-infected participants by treatment status and level of viremia. The percentages of activated ($CD38^+HLA\text{-}DR^+$) $CD4^+$ T cells *(dark gray boxes)* and $CD8^+$ T cells *(light gray boxes)* are plotted among 47 HIV-uninfected participants, 30 controllers, 187 antiretroviral therapy (ART)–treated HIV-infected participants with plasma HIV RNA levels <75 copies/mL (VL<75), and 66 untreated HIV-infected participants with plasma HIV RNA levels >75 copies/mL (VL>75). The boxes span the 25th and 75th percentile values, the error bars span the 10th and 90th percentile values, each midline represents the median, and the dots represent individual observations below the 10th and above the 90th percentile values. Although controllers had lower $CD4^+$ and $CD8^+$ T cell activation levels than untreated HIV-infected participants with detectable viremia ($P < .001$, for both), they had higher $CD4^+$ and $CD8^+$ T cell activation levels than uninfected participants ($P < .001$, for both) and higher $CD8^+$ T cell activation levels than treated participants with undetectable plasma HIV RNA levels ($P = .048$).

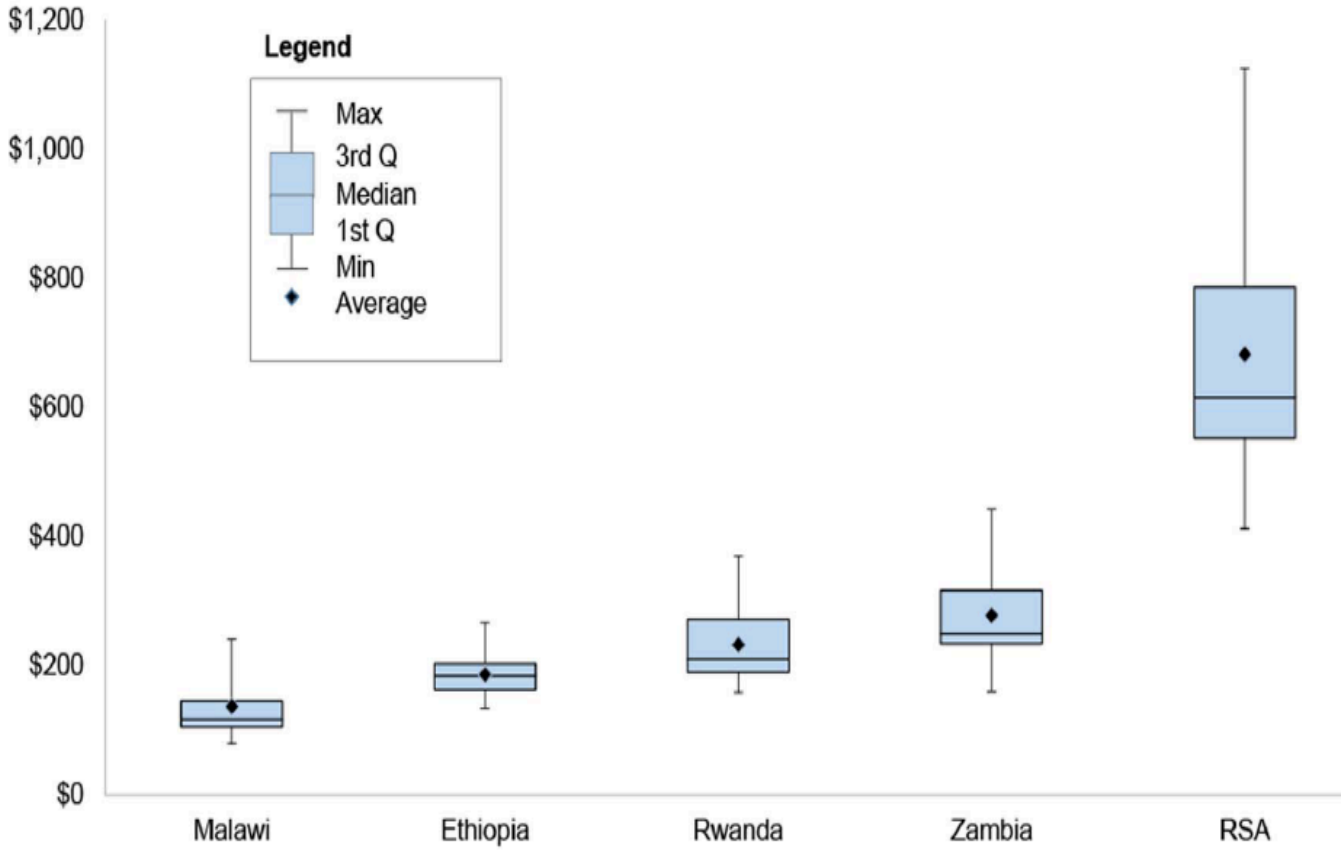
Hunt et al., *JID*, 2008



Hunt et al., *PLOS ONE*, 2011

Box Plot

Figure 1. Cost per ART Patient-Year by Country (USD)
doi:10.1371/journal.pone.0108304.g001

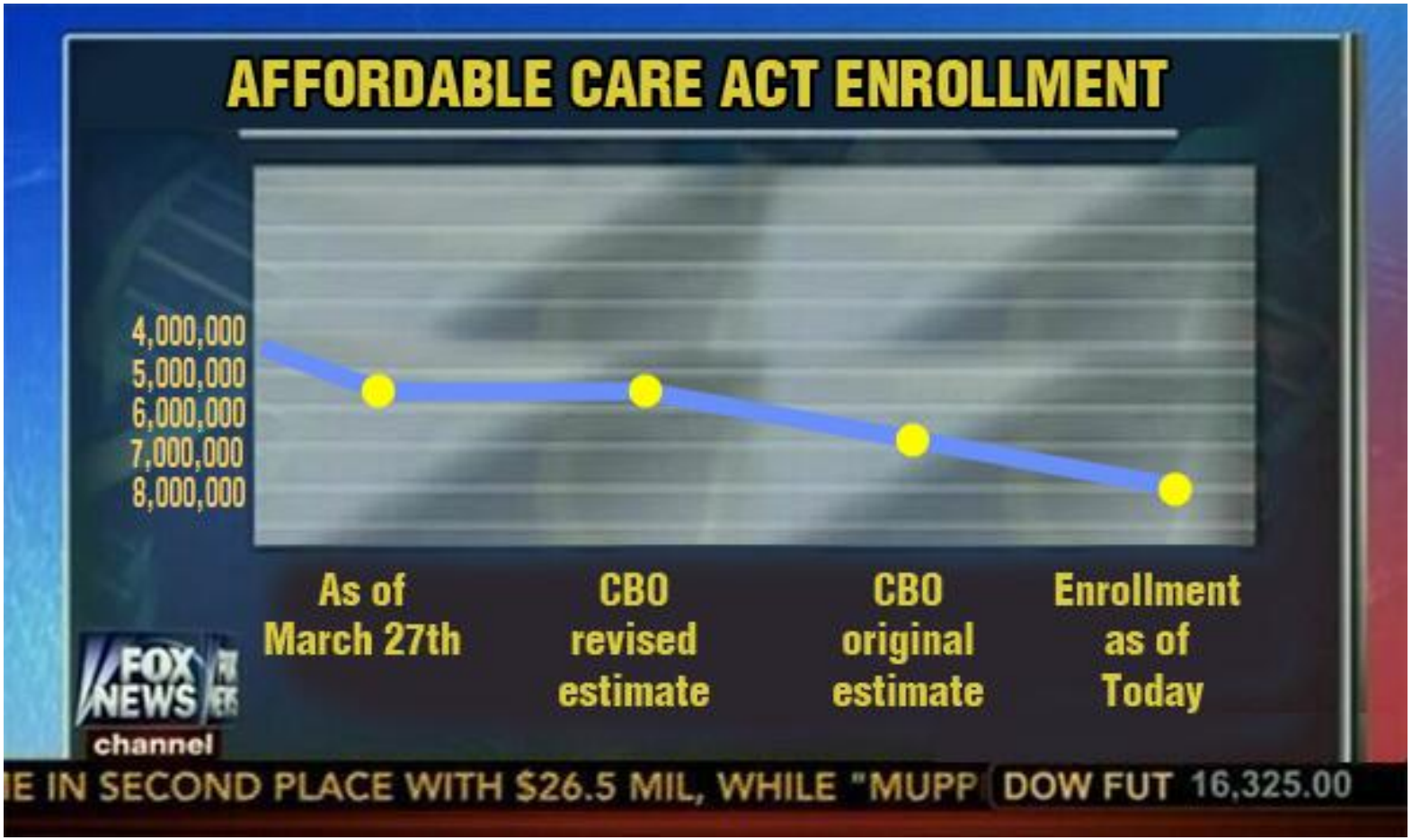


Tagar et al. *PLOS ONE*, 2014

Graphics: Responsibility

- Graphics are powerful – with great power comes great responsibility

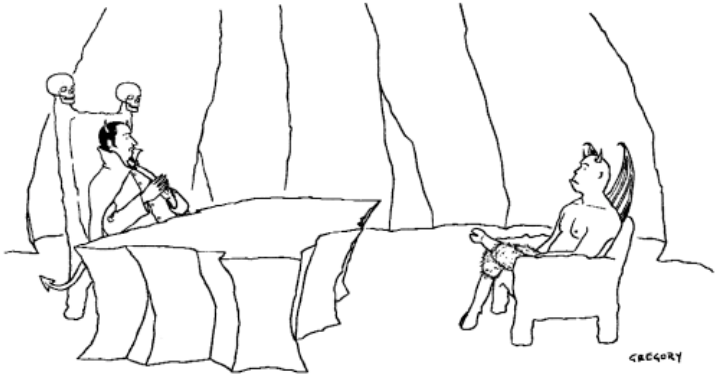
Courtesy:
Fox News,
5/3/2014



Summary

- Examine data and the hoped for message it will convey
- Choose graphical display based on this
- Graphical excellence
- Data-to-ink ratio
- Eliminate unnecessary embellishments, lines, boxes, etc.
- Don't be afraid to label in order to explain:
  - "To clarify, add detail"
- Develop preferred formats and use them repeatedly

Thanks!



"I need someone well versed in the art of torture—do you know PowerPoint?"

JUST DO THE PRESENTATION WILLIAMS AND LET THE NUMBERS SPEAK FOR THEMSELF.

Happy to take any questions!