Major Statistical Challenges in Count Data Analysis

Hui Zhang, PhD

Associate Member, Department of Biostatistics

St. Jude Children's Research Hospital

Outline

- **Significance of modeling count data**
- **Over-dispersion in cross-sectional counts**
- **Over-dispersion in longitudinal counts**
  Comparison of two popular methods
  Detection over-dispersion in longitudinal counts
  Address missing data
- **Zero-inflation in cross sectional and longitudinal counts**
- **An example of future research projects**

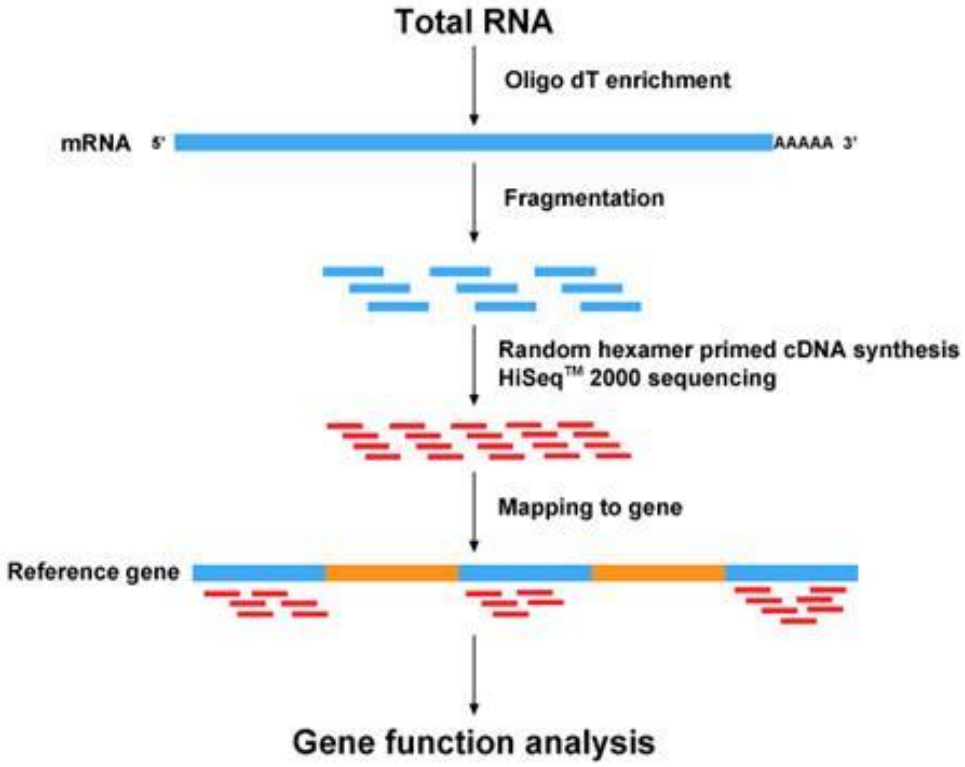
What are count data?

In statistics, count data represent a type of data, in which the observations can take only the non-negative integer values {0, 1, 2, 3, ...}, and where these integers arise from counting rather than ranking.

- Relation to binomial/binary data

Why count data?

- Common in biomedical and clinical research, for example, the number of hospitalizations in a given time period
- Next Generation Sequencing, such as RNA-Seq, generates count data.



Common RNA-seq workflow (from bgisequence.com)

Parametric distributions to model counts

- Poisson
- Negative Binomial
- Separate semi-parametric methods from parametric methods: quasi-likelihood Poisson

Poisson distribution

- The probability mass function of Poisson distribution:

$$\Pr(Y = y \mid \lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad \text{for } y = 0,1,2,\ldots$$

- Poisson is a one parameter distribution (λ)

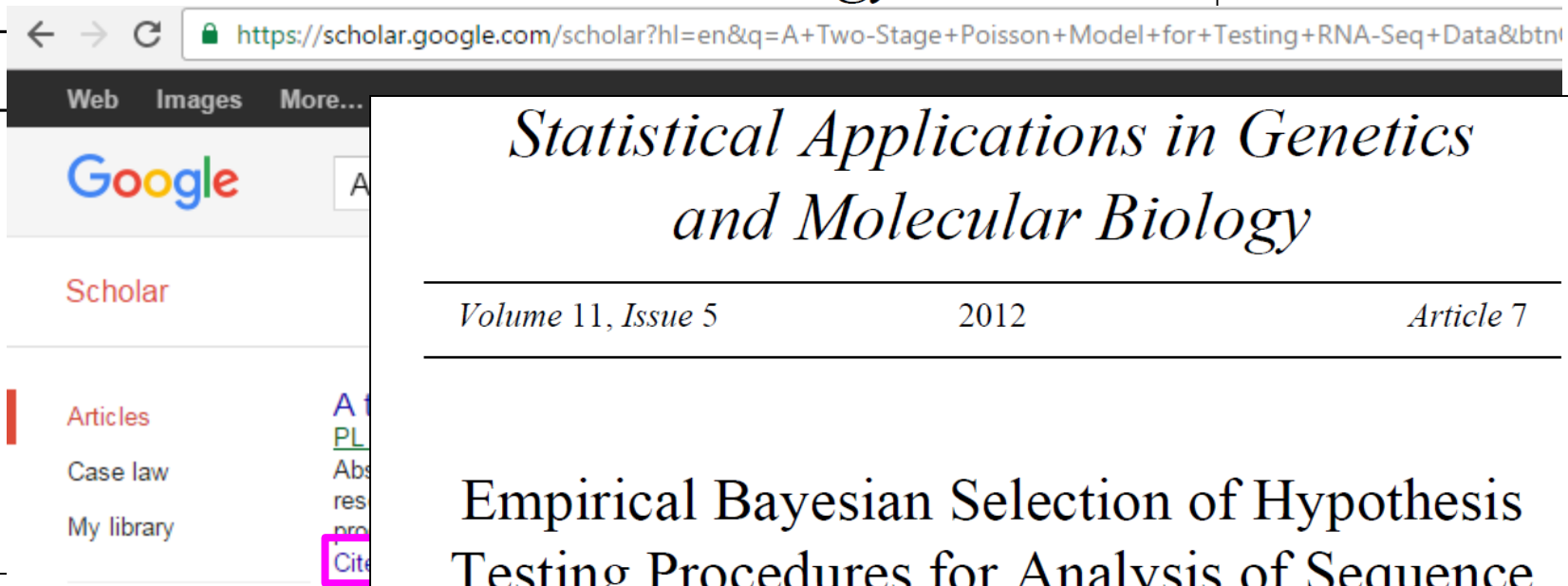
- λ is the mean or expected value of a Poisson distribution

- λ is also the variance of a Poisson distribution

- In real count data, it is very common that variance>>mean, called as over-dispersion, and we should use alterative method, such as negative binomial or quasi-likelihood Poisson

Over-dispersion

Statistical Applications in Genetics and Molecular Biology



Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 5 2012 *Article* 7

Empirical Bayesian Selection of Hypothesis Testing Procedures for Analysis of Sequence Count Expression Data

Stanley B. Pounds, *St. Jude Children's Research Hospital*
Cuilan L. Gao, *University of Tennessee at Chattanooga*
Hui Zhang, *St. Jude Children's Research Hospital*

Detect over-dispersion

- Parametric method:

  Goodness-of-fit

- Semi-parametric method:

  Quasi-likelihood

Detect over-dispersion

34 *The Open Bioinformatics Journal*, 2013, *7*, (Suppl 1: M3) 34-40

Open Access

Statistical Methods for Overdispersion in mRNA-Seq Count Data

Hui Zhang*, Stanley B. Pounds and Li Tang

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Abstract: Recent developments in Next-Generation Sequencing (NGS) technologies have opened doors for ultra high throughput sequencing mRNA (mRNA-seq) of the whole transcriptome. mRNA-seq has enabled researchers to comprehensively search for underlying biological determinants of diseases and ultimately discover novel preventive and therapeutic solutions. Unfortunately, given the complexity of mRNA-seq data, data generation has outgrown current analytical capacity, hindering the pace of research in this area. Thus, there is an urgent need to develop novel statistical methodology that addresses problems related to mRNA-seq data. This review addresses the common challenge of the presence of overdispersion in mRNA count data. We review current methods for modeling overdispersion, such as negative binomial, quasi-likelihood Poisson method, and the two-stage adaptive method; introduce related statistical theories; and discuss their applications to mRNA-seq count data.

Keywords: Count response, mRNA-seq, negative binomial theory, over-dispersion, Poisson, quasi-likelihood.

Outline

- **Significance of modeling count data**
- **Over-dispersion in cross-sectional counts**
- **Over-dispersion in longitudinal counts**
  - Comparison of two popular methods
  - Detection over-dispersion in longitudinal counts
  - Address missing data
- **Zero-inflation in cross sectional and longitudinal counts**
- **An example of future research projects**

Longitudinal data in clinical trials

- Modern clinical trials usually last for a long time, and even for decades.
- Repeated measures on the same patients
- Missing data are very common

A real data project example

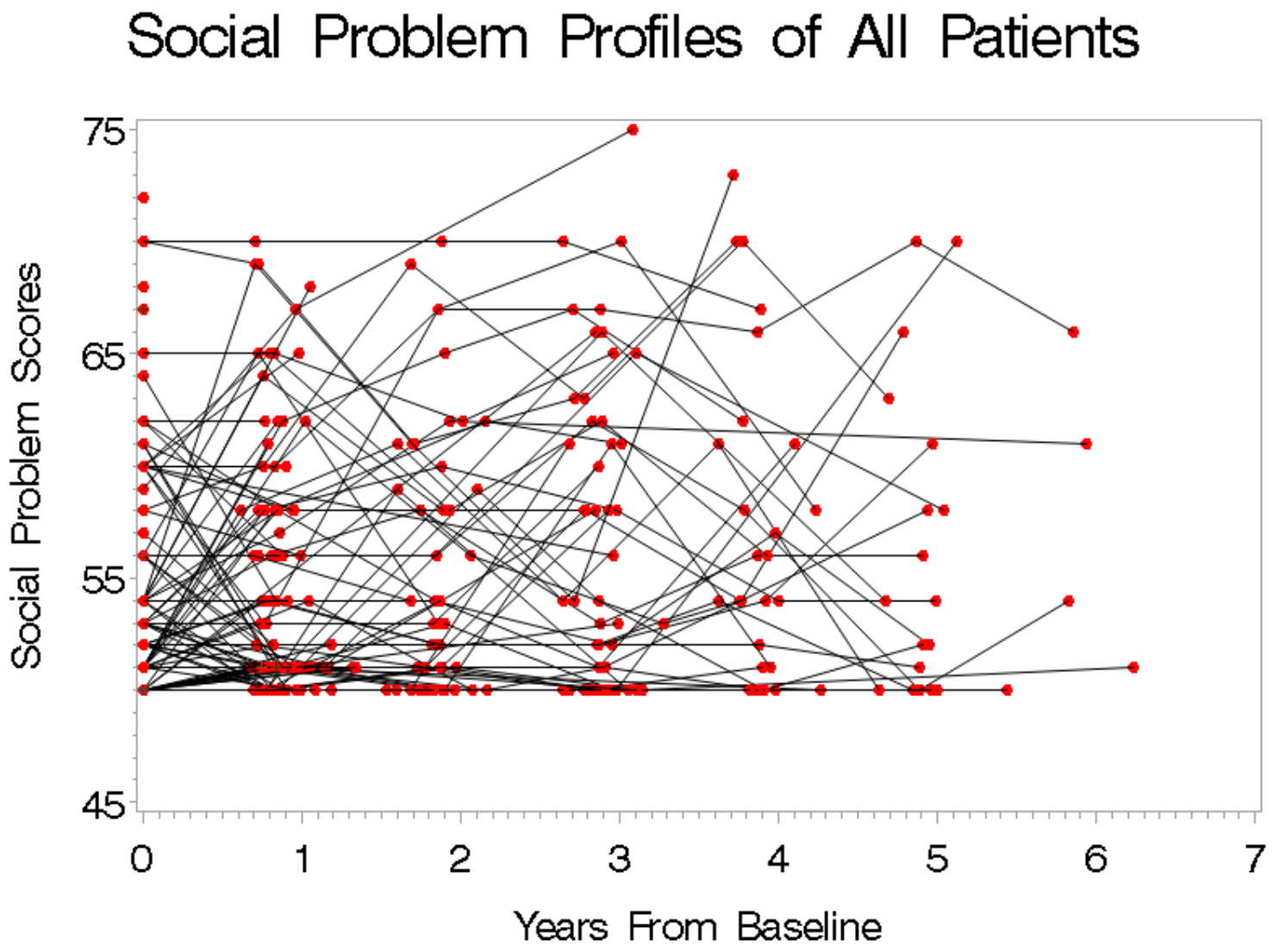
St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

A B S T R A C T

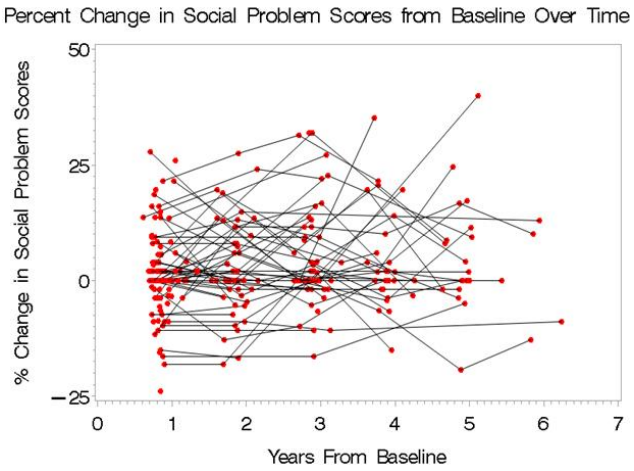
Purpose
To examine longitudinal parent-reported social outcomes for children treated for pediatric embryonal brain tumors.

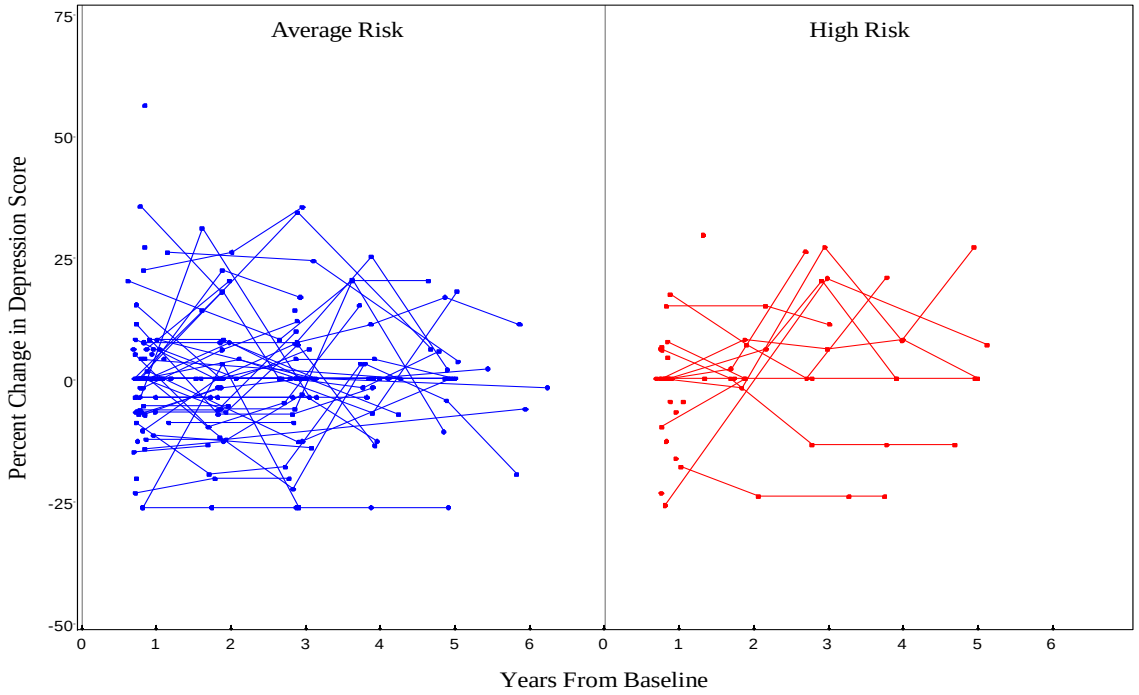
Patients and Methods
Patients (N = 220) were enrolled onto a multisite clinical treatment protocol. Parents completed the Child Behavior Checklist/6-18 at the time of their child's diagnosis and yearly thereafter. A generalized linear mixed effects model regression approach was used to examine longitudinal changes in parent ratings of social competence, social problems, and withdrawn/depressed behaviors with demographic and treatment factors as covariates.

First outcome: Social Problem Profile

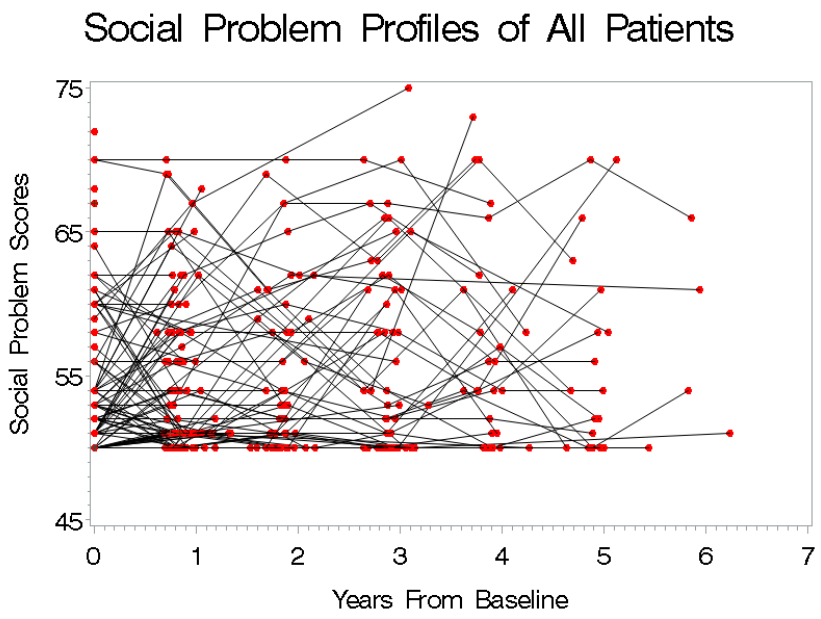


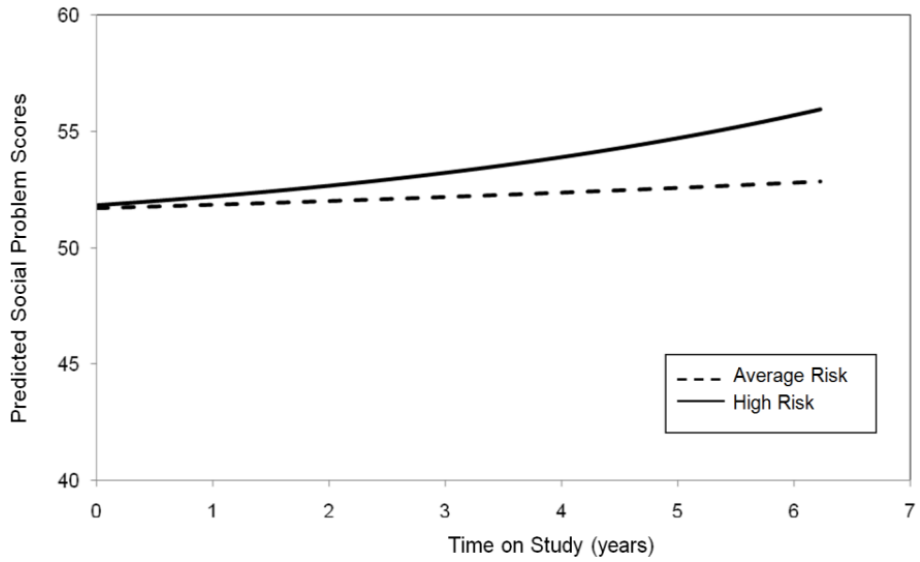
First outcome: Social Problem Profile



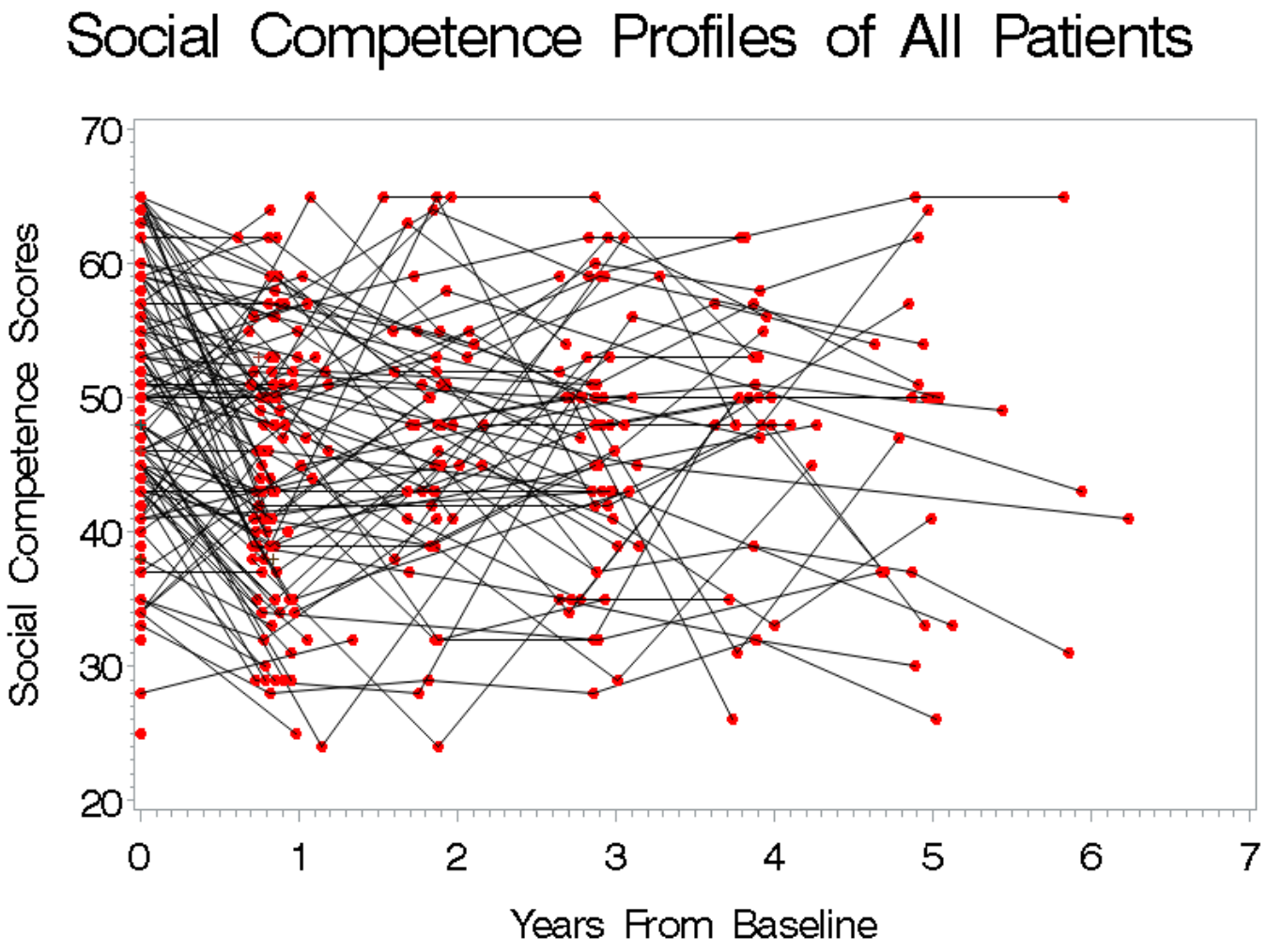


First outcome: Social Problem Profile

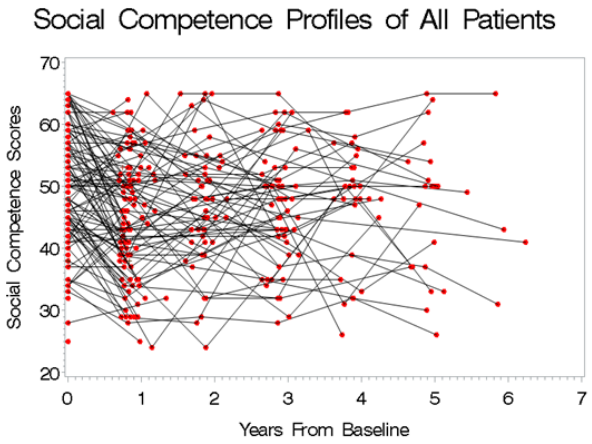


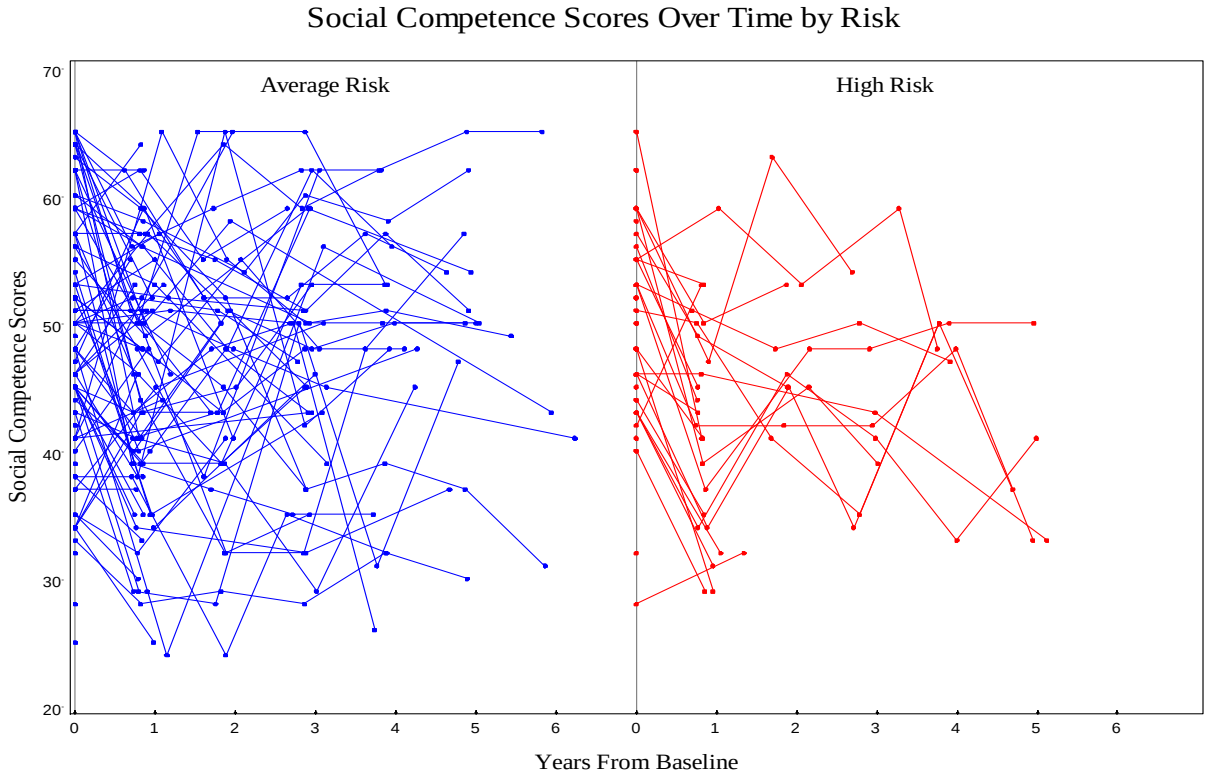


Second outcome: Social Competence Profile

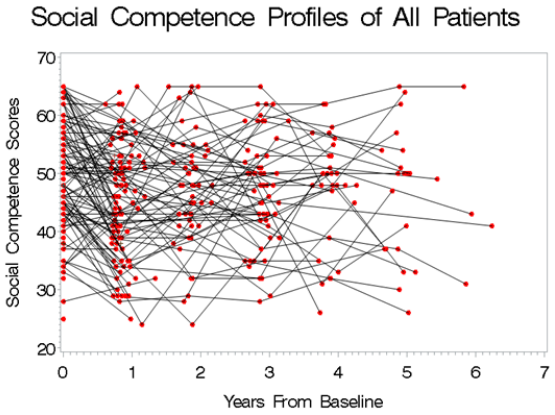


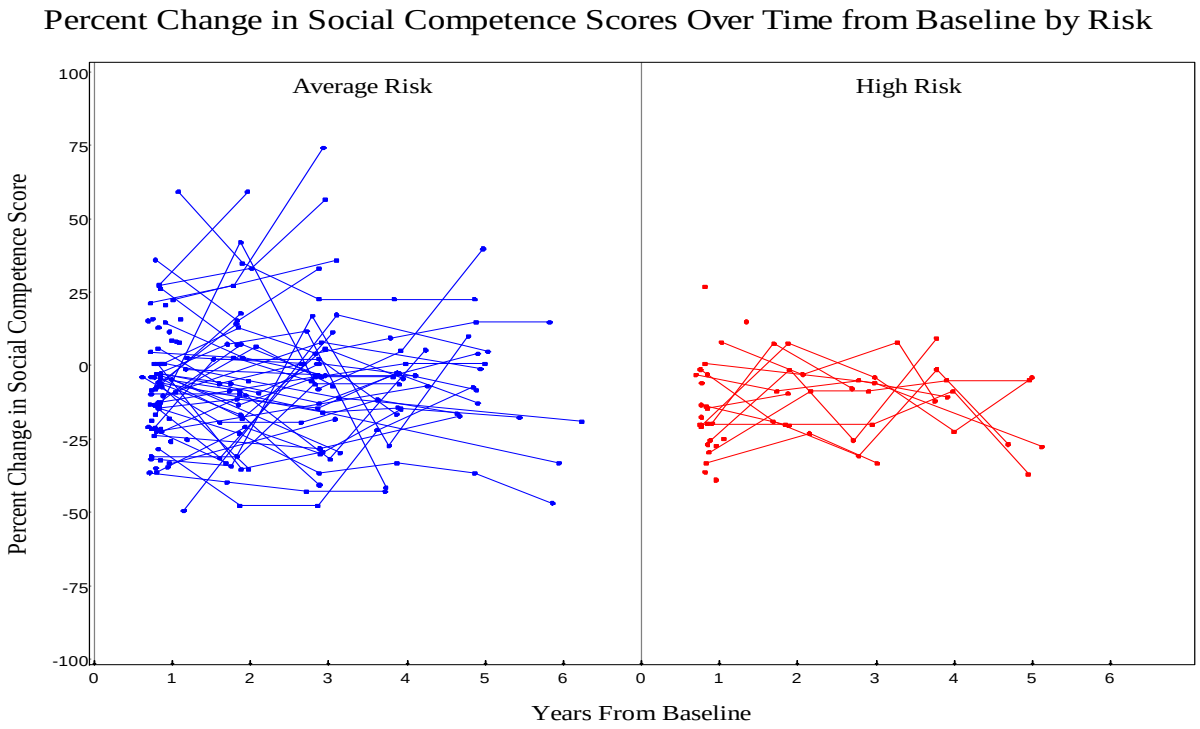
Second outcome: Social Competence Profile



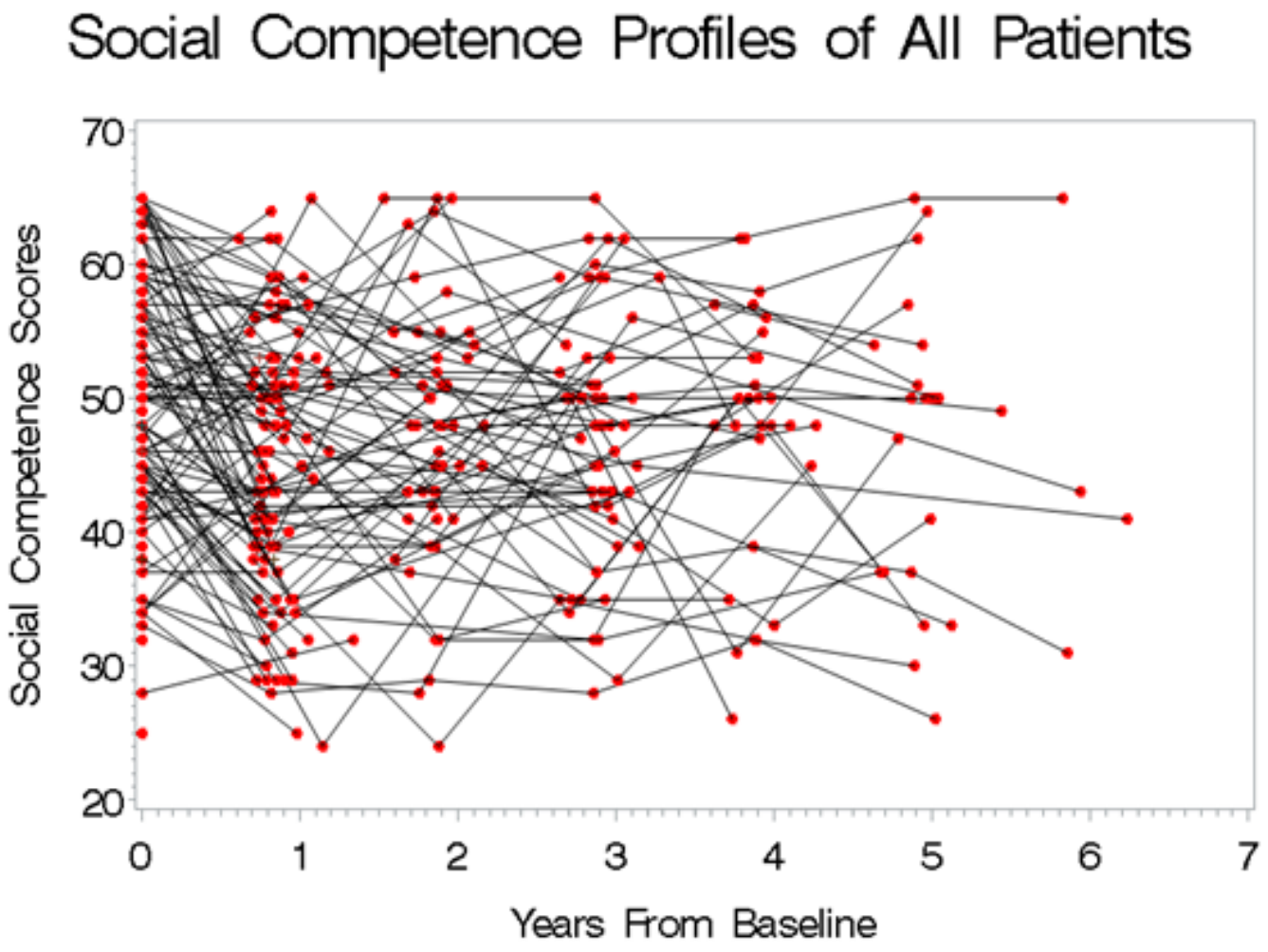


Second outcome: Social Competence Profile

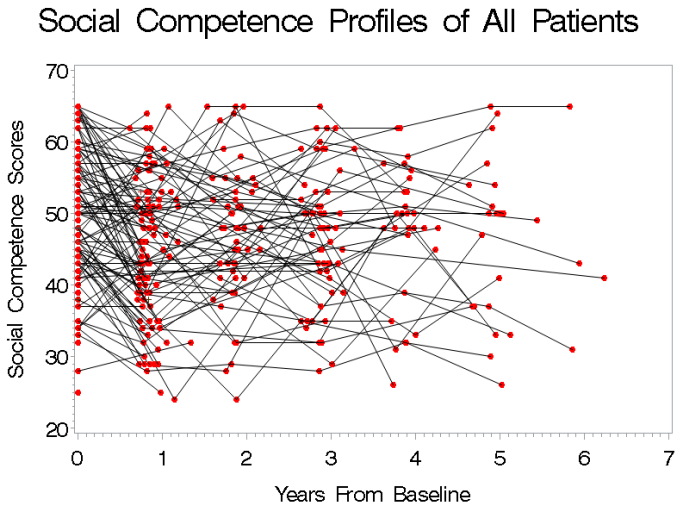


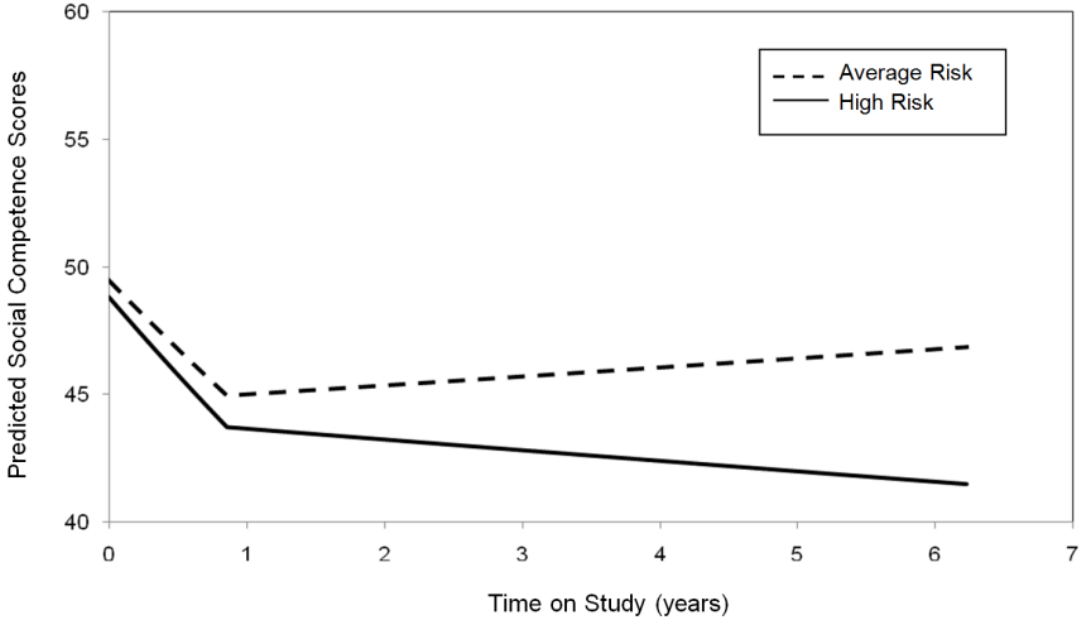


Second outcome: Social Competence Profile



Second outcome: Social Competence Profile





Related publication

VOLUME 30 · NUMBER 33 · NOVEMBER 20 2012

JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

Parent-Reported Social Outcomes After Treatment for Pediatric Embryonal Tumors: A Prospective Longitudinal Study

Tara M. Brinkman, Shawna L. Palmer, Si Chen, Hui Zhang, Karen Evankovich, Michelle A. Swain, Melanie J. Bonner, Laura Janzen, Sarah Knight, Carol L. Armstrong, Robyn Boyle, and Amar Gajjar

Tara M. Brinkman, Shawna L. Palmer, Si Chen, Hui Zhang, Amar Gajjar, St Jude Children's Research Hospital, Memphis, TN; Karen Evankovich, Texas Children's Hospital, Houston, TX; Melanie J. Bonner, Duke University Medical Center, Durham, NC; Carol L. Armstrong, Children's Hospital of Philadelphia, Philadelphia, PA; Laura Janzen, The Hospital for Sick Children, Toronto, Canada; Michelle A. Swain, Royal Children's Hospital of Brisbane, Herston; Sarah Knight, Royal Children's Hospital of Melbourne, Parkville; Robyn Boyle, Sydney Children's Hospital, Randwick, Australia.

Submitted November 15, 2011; accepted August 15, 2012; published online ahead of print at www.jco.org on October 15, 2012.

Supported in part by Cancer Center Support Grant No. P30-CA21765 from the National Cancer Institute, the Noyes Brain Tumor Foundation, Musicians Against Childhood Cancer, and by the American Lebanese Syrian Associated Charities.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Tara M. Brinkman, PhD, St Jude Children's Research Hospital, Department of Epidemiology and Cancer Control, 262 Danny Thomas Place, MS 735, Memphis, TN 38105, e-mail: tara.brinkman@stjude.org.

0732-183X/12/3033-4134/$20.00

DOI: 10.1200/JCO.2011.40.6702

A B S T R A C T

Purpose
To examine longitudinal parent-reported social outcomes for children treated for pediatric embryonal brain tumors.

Patients and Methods
Patients (N = 220) were enrolled onto a multisite clinical treatment protocol. Parents completed the Child Behavior Checklist/6-18 at the time of their child's diagnosis and yearly thereafter. A generalized linear mixed effects model regression approach was used to examine longitudinal changes in parent ratings of social competence, social problems, and withdrawn/depressed behaviors with demographic and treatment factors as covariates.

Results
During the 5-year period following diagnosis and treatment, few patients were reported to have clinically elevated scores on measures of social functioning. Mean scores differed significantly from population norms, yet remained within the average range. Several factors associated with unfavorable patterns of change in social functioning were identified. Patients with high-risk treatment status had a greater increase in parent-reported social problems ($P = .001$) and withdrawn/depressed behaviors ($P = .01$) over time compared with average-risk patients. Patients with posterior fossa syndrome had greater parent-reported social problems over time ($P = .03$). Female patients showed higher withdrawn/depressed scores over time compared with male patients ($P < .001$). Patient intelligence, age at diagnosis, and parent education level also contributed to parent report of social functioning.

Conclusion
Results of this study largely suggest positive social adjustment several years after diagnosis and treatment of a pediatric embryonal tumor. However, several factors, including treatment risk status and posterior fossa syndrome, may be important precursors of long-term social outcomes. Future research is needed to elucidate the trajectory of social functioning as these patients transition into adulthood.

J Clin Oncol 30:4134-4140. © 2012 by American Society of Clinical Oncology

INTRODUCTION

Survivors of pediatric brain tumors are at particularly high risk for experiencing adverse effects related to their disease and treatments.[1,2] Although substantial effort has been directed at characterizing medical and neurocognitive outcomes,[1,3-6] considerably less attention has focused on behavioral and social consequences of treatment for childhood brain tumors. Although evidence suggests that deficits in social functioning represent a significant part of the morbidity experienced by these survivors,[7] the nature and time course of these difficulties remain poorly understood.

Previous cross-sectional studies, using heterogeneous samples of brain tumor survivors, have reported that survivors have fewer close friendships[8,9] and are socially isolated compared with peers.[9] Survivors also demonstrate greater social problems[10,11] and diminished social competence[12,13] relative to normative samples. Compared with siblings, adolescent survivors are reported to have increased depression/anxiety and antisocial behaviors, as well as reduced social competence.[14] In a rare longitudinal study of 53 patients treated with cranial radiation therapy for posterior fossa tumors, Mabbott et al[15] reported a progressive decline in social functioning with increasing time from diagnosis.

Table 3. Parent-Reported Social Outcomes by Time on Study

	Social Competence*							Social Problems†							Withdrawn/Depressed†						
Year	No.	Mean	SD	P‡	No. of Patients§	% §	P‖	No.	Mean	SD	P‡	No. of Patients	%	P‖	No.	Mean	SD	P‡	No. of Patients	%	P‖
Baseline	168	49.9	9.1	.94	3	1.8	1.0	169	53.5	4.7	**< .001**	3	1.8	1.0	169	56.0	7.3	**< .001**	10	5.9	**.002**
1	135	44.8	9.0	**< .001**	9	6.7	**.002**	140	54.8	5.7	**< .001**	4	2.9	.37	140	57.2	8.2	**< .001**	15	10.7	**< .001**
2	63	46.5	9.0	**.003**	3	4.8	.13	62	55.5	6.4	**< .001**	2	3.2	.35	62	56.5	6.9	**< .001**	3	4.8	.13
3	75	45.5	9.2	**< .001**	3	4.0	.19	76	56.4	7.2	**< .001**	5	6.6	**.02**	76	57.1	7.8	**< .001**	6	7.9	**.004**
4	41	47.3	9.1	.07	1	2.4	.56	41	56.0	6.6	**< .001**	3	7.3	.05	41	55.4	7.3	**< .001**	3	7.3	.05
5	33	45.9	10.3	**.03**	3	9.1	**.03**	33	57.4	8.0	**< .001**	4	12.1	**.004**	33	57.0	7.1	**< .001**	1	3.0	.49

NOTE. Bold font indicates significance.
Abbreviation: SD, standard deviation.
*Average range defined as T scores ranging from 36-50. Clinically significant scores are defined as T scores ≤ 30.
†Average range defined as T scores ranging from 50-64. Clinically significant scores defined as T scores ≥ 70.
‡*t* test for equality of means, with expected mean of 50.
§No. of patients and corresponding % refer to those whose scores exceeded clinical significance.
‖Exact binomial test, with expected clinical proportion of 2%.

years), analysis of this trend using a discontinuous-slope GLMM revealed that the change in slope was significant ($P < .001$), with a significant negative slope between diagnosis and initial follow-up ($P = .001$) and a negative but nonsignificant slope after 1 year postdiagnosis. Figure 3 shows change in social competence over time by patient risk status using the discontinuous-slope GLMM.

Impact of Long-Term Observations

Because the study remained open to accrual, a larger number of patients contributed data to earlier study time points than later. To determine the impact of having a lower number of evaluations at 4 years postdiagnosis and beyond, the models were examined using only observations up to and including 3 years postdiagnosis. For social problems and social competence, the results remained identical to the models using all time points. A similar pattern of results was found for withdrawn/depressed behaviors. Though the sex-by-time and PFS-by-time interactions were not retained in the best-fitting 3-year model, single covariate models including the interactions of sex and PFS with time since diagnosis were significant at 3 and 5 years. Therefore, it was concluded that including observations at later time points, although fewer in number, did not significantly alter the interpretation of study results.

DISCUSSION

To our knowledge, this is the largest longitudinal study of parent-reported social outcomes for pediatric brain tumor survivors. Importantly, our sample was relatively homogeneous with respect to diagnosis and treatment, factors that have been difficult to disentangle in previous research on social outcomes. We found that few patients were reported to have clinically elevated scores on measures of social competence, social problems, or withdrawn/depressed behavior; however, the proportion of survivors with clinically elevated scores often exceeded the expected proportion based on population data.

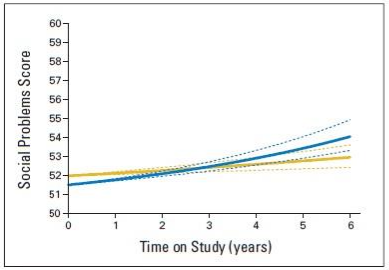


Fig 2. Patient risk status and parent-reported social problems over time (T scores: mean, 50; standard deviation, 10). Lower social problems scores reflect better functioning. Solid gold line indicates average risk; dashed gold lines indicate 95% CI. Solid blue line indicates high risk; dashed blue lines indicate 95% CI.

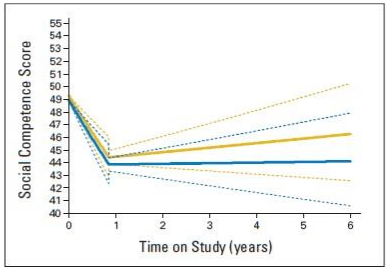


Fig 3. Patient risk status and parent-reported social competence before and after mean time until first follow-up (0.86 years) using a discontinuous-slope generalized linear mixed effects model (T scores: mean, 50; standard deviation, 10). Higher social competence scores reflect better functioning. Solid blue line indicates high risk; dashed blue lines indicate 95% CI. Solid gold line indicates average risk; dashed gold lines indicate 95% CI.

A Clinical Publication

St. Jude Children's Research Hospital
ALSAC · Danny Thomas, Founder

Published Ahead of Print on October 12, 2015 as 10.1200/JCO.2015.61.6672
The latest version is at http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2015.61.6672

JOURNAL OF CLINICAL ONCOLOGY | ORIGINAL REPORT

Computerized Cognitive Training for Amelioration of Cognitive Late Effects Among Childhood Cancer Survivors: A Randomized Controlled Trial

Heather M. Conklin, Robert J. Ogg, Jason M. Ashford, Matthew A. Scoggins, Ping Zou, Kellie N. Clark, Karen Martin-Elbahesh, Kristina K. Hardy, Thomas E. Merchant, Sima Jeha, Lu Huang, and Hui Zhang

Heather M. Conklin, Robert J. Ogg, Jason M. Ashford, Matthew A. Scoggins, Ping Zou, Kellie N. Clark, Karen Martin-Elbahesh, Thomas E. Merchant, Sima Jeha, Lu Huang, and Hui Zhang, St Jude Children's Research Hospital, Memphis, TN; and Kristina K. Hardy, Children's National Medical Center and George Washington University School of Medicine, Washington, DC.

Published online ahead of print at www.jco.org on October 12, 2015.

Supported in part by the National Cancer Institute (Core Grant No. P30 CA21765 to St Jude Cancer Center Support), American Cancer Society (Grant No. RSGPB-11-009-01-CPPB to H.M.C.), and American Lebanese Syrian Associated Charities. Cogmed software was provided by Pearson Education (New York, NY) for research purposes.

Presented in part at the 49th Annual Meeting of the American Society of Clinical Oncology, Chicago, IL, May 31-June 4, 2013; the Annual Meeting of the Society for Neuro-Oncology, San Francisco, CA, November 21-24, 2013; and the Annual Meeting of the International Neuropsychological Society, Seattle, WA, February 12-15, 2014.

Pearson Education did not play a role in the design or conduct of the study; analysis or interpretation of the data; or preparation, review, or approval of the manuscript.

Authors' disclosures of potential conflicts of interest are found in the

A B S T R A C T

Purpose
Children receiving CNS-directed therapy for cancer are at risk for cognitive problems, with few available empirically supported interventions. Cognitive problems indicate neurodevelopmental disruption that may be modifiable with intervention. This study evaluated short-term efficacy of a computerized cognitive training program and neural correlates of cognitive change.

Patient and Methods
A total of 68 survivors of childhood acute lymphoblastic leukemia (ALL) or brain tumor (BT) with identified cognitive deficits were randomly assigned to computerized cognitive intervention (male, n = 18; female, n = 16; ALL, n = 23; BT, n = 11; mean age ± standard deviation, 12.21 ± 2.47 years) or waitlist (male, n = 18; female, n = 16; ALL, n = 24; BT, n = 10; median age ± standard deviation, 11.82 ± 2.42 years). Intervention participants were asked to complete 25 training sessions at home with weekly, telephone-based coaching. Cognitive assessments and functional magnetic resonance imaging scans (intervention group) were completed pre- and postintervention, with immediate change in spatial span backward as the primary outcome.

Results
Survivors completing the intervention (n = 30; 88%) demonstrated greater improvement than controls on measures of working memory (mean ± SEM; eg, Wechsler Intelligence Scale for Children [fourth edition; WISC-IV] spatial span backward, 3.13 ± 0.58 *v* 0.75 ± 0.43; $P = .002$; effect size [ES], 0.84), attention (eg, WISC-IV spatial span forward, 3.30 ± 0.71 *v* 1.25 ± 0.39; $P = .01$; ES, 0.65), and processing speed (eg, Conners' Continuous Performance Test hit reaction time, −2.10 ± 1.47 *v* 2.54 ± 1.25; $P = .02$; ES, .61) and showed greater reductions in reported executive dysfunction (eg, Conners' Parent Rating Scale III, −6.73 ± 1.51 *v* 0.41 ± 1.53; $P = .002$; ES, 0.84). Functional magnetic resonance imaging revealed significant pre- to post-training reduction in activation of left lateral prefrontal and bilateral medial frontal areas.

Conclusion
Study findings show computerized cognitive training is feasible and efficacious for childhood cancer survivors, with evidence for training-related neuroplasticity.

J Clin Oncol 33.

Media Coverage

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

Computerized cognitive training improves childhood cancer survivors' at... http://www.sciencedaily.com/releases/2015/10/151012174502.htm

ScienceDa
Your source for the latest research

Science News

Computerized cognit
childhood cancer sur
memory

Date: October 12, 2015
Source: St. Jude Children's R
Summary: Computer-based cog
medication for improv
childhood cancer sur
new study.

Share: 6 38

FULL STORY

Intensive, adaptive com
training presented as a
prove working memory
childhood cancer survi
olutionizing managemer
cer treatment. St. Jude
tal investigators led the
Journal of Clinical Onco

Working memory improved significa
also improved for childhood cancer
and 30 computer-based training ses
speed at which the brain processes
45 minutes and included verbal and
games but designed to improve wor

1 of 5

Oklahoma's Own
News Weather Sports Vide

Computerized cognitive t
childhood cancer survivc
memory

Information contained on this page is provided by ar
provider. WorldNow and this Station make no warra
therewith. If you have any questions or comments a
pressreleases@worldnow.com.

SOURCE St. Jude Children's Research Hospital

St. Jude Children's Research Hospital study sho
training is as effective as medication for improvi
in childhood cancer survivors with cognitive def

MEMPHIS, Tenn., Oct. 12, 2015 /PRNewswire-USN
computer-based cognitive training presented as a vi
memory and other cognitive skills of childhood canc
revolutionizing management of the late effects of ca
Research Hospital investigators led the study, which
Clinical Oncology.

Working memory improved significantly, and attentio
improved for childhood cancer survivors who comple
computer-based training sessions. Processing spee
brain processes information. The sessions lasted 30
and visual-spatial exercises presented as games bu
memory.

The benefits to working memory and attention from
reported in previous studies of stimulant medications
moved performance of the 30 survivors who comple
Caregivers also reported significant improvement in

FOX NEWS | FOX BUSINESS | FOX NEWS Latino | FOX NEWS GO | FOX NEWS RADIO | FOX NATION | FOX NEWS INSIDER | LOGIN

FOX NEWS

Search foxnews.com

Health Home | Men's Health | Women's Health | Children's Health | Alternative Medicine | Diabetes | Heart Health | Allergies

NERVOUS SYSTEM HEALTH

Computer training may improve memory for childhood cancer survivors

Published October 13, 2015 · Reuters 13 0


Alexandra Munoz, 5, who lost her hair due to chemotherapy to treat a malignant brain tumor, undergoes a session of treatment with the help of a nurse in the cancer ward of the Luis Calvo Mackenna Hospital in Santiago, October 20, 2014. REUTERS/Rodrigo Garrido

Children who receive cancer treatments may suffer thinking problems later, but using an at-home computer training program can help reduce these deficits, according to a new study.

"This is the only computerized training so far in childhood cancer survivors," said lead author Heather M. Conklin of St. Jude Children's Research Hospital in Memphis, Tennessee.

The study included 68 survivors of acute lymphoblastic leukemia (ALL), a blood cancer, or brain tumors, who had all survived at least one year after their cancer

More from Fox News


Experts caution on study citing method to predict sexual orientation


Age-associated financial vulnerability often overlooked

Longitudinal data modeling

- Generalized Linear Mixed-effect Model (GLMM)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} / \boldsymbol{l}(\boldsymbol{E}(\boldsymbol{y})) = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

$\boldsymbol{y}$ is a known vector of observations, with mean $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X\beta}$;

- Generalized Estimating Equations (GEE)

$$\sum_{i=1}^{N} \frac{\partial \mu_{ij}}{\partial \beta_k} V_i \quad ^{-1} \{Yi - \mu_i(\beta)\} = 0$$

- Non-parametric methods

Longitudinal counts: GLMM v.s. GEE

Journal of Applied Statistics
Vol. 39, No. 9, September 2012, 2067–2079

Taylor & Francis
Taylor & Francis Group

A new look at the difference between the GEE and the GLMM when modeling longitudinal count responses

H. Zhang[a]*, Q. Yu[b], C. Feng[b], D. Gunzler[c], P. Wu[b] and X.M. Tu[b,d,e]

[a]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [b]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA; [c]School of Medicine, Case Western Reserve University, MetroHealth Medical Center, Cleveland, OH 44109-1998, USA; [d]Department of Psychiatry, University of Rochester, Rochester, NY 14642, USA; [e]VA Center of Excellence at Canandaigua, Canandaigua VAMC, Canandaigua, NY 14424, USA

Longitudinal counts: GLMM v.s. GEE

Real date analysis using COMBINE, a multi-site clinical trial conducted from 2001 to 2004 on 1383 individuals with alcohol dependence. Two primary outcomes of the study were (1) days of no heavy drinking (2) days of no drinking, which were collected at baseline, weeks 8 (visit 1), 16 (visit 2) and 26 (visit 3). We are interested to know the primary outcomes' changes since baseline after adjusted by some demographic variables (not shown).

Comparison of estimates (standard errors $\times 10^{-2}$) between GEE and GLMM

Models fit	Visit 1 (β_1 or $\tilde{\beta}_1$)	Visit 2 (β_2 or $\tilde{\beta}_2$)	Visit 3 (β_3 or $\tilde{\beta}_3$)
Days of no heavy drinking			
GEE ($\boldsymbol{\beta}$)	1.908(3.6)	0.144(2.2)	0.144(2.8)
GLMM ($\tilde{\boldsymbol{\beta}}$)	0.694(6.2)	0.144(1.5)	0.144(1.5)
Days of no drinking			
GEE ($\boldsymbol{\beta}$)	1.307(4.7)	0.224(3.3)	0.216(4.1)
GLMM ($\tilde{\boldsymbol{\beta}}$)	−0.30(7.3)	0.224(1.9)	0.216(1.9)

Reader's Comments

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

Zhang, Hui

From: Gregoire, Timothy <timothy.gregoire@yale.edu>
Sent: Thursday, October 04, 2012 7:47 PM
To: Zhang, Hui
Cc: David Affleck
Subject: cudos on JAS article. .

Dear Hui,

A late night thank you for your informative article (JAS, 2012, v39) on GEE versus GLMM for count data. Excellently and clearly written, and very insightful.

Tim

Timothy G. Gregoire
J. P. Weyerhaeuser Professor of Forest Management
School of Forestry & Environmental Studies, Yale University
360 Prospect Street, New Haven, CT 06511-2104 U.S.A.

office: 1.203.432.9398 mobile: 1.203.508.4014, fax: 1.203.432.3809
timothy.gregoire@yale.edu
G&V sampling text: http://crcpress.com/product/isbn/9781584883708

Longitudinal counts: GLMM v.s. GEE

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

- GEE is more robust to distribution mis-specification while GLMM is more sensitive to distribution assumption.

- For Poisson, the most common violation of distribution assumption is over-dispersion.

- For most available SAS procedures and R packages to model GLMM, one could address the over-dispersion by changing the setting accordingly.

- How reliable are they?

Over-dispersion in longitudinal counts

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2015
http://dx.doi.org/10.1080/00949655.2015.1111376

Taylor & Francis
Taylor & Francis Group

Comparison of different computational implementations on fitting generalized linear mixed-effects models for repeated count measures

Lu Huang[a], Li Tang[a], Bo Zhang[b], Zhiwei Zhang[b] and Hui Zhang[a]

[a]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA; [b]Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD, USA

$$y_{it} \mid x_i, \mathbf{b}_i \sim \text{Poisson}\,(\mu_{it}) \quad \log(\mu_{it}) = \beta_0 + b_{i0} + x_i(\beta_1 + b_{i1}),$$

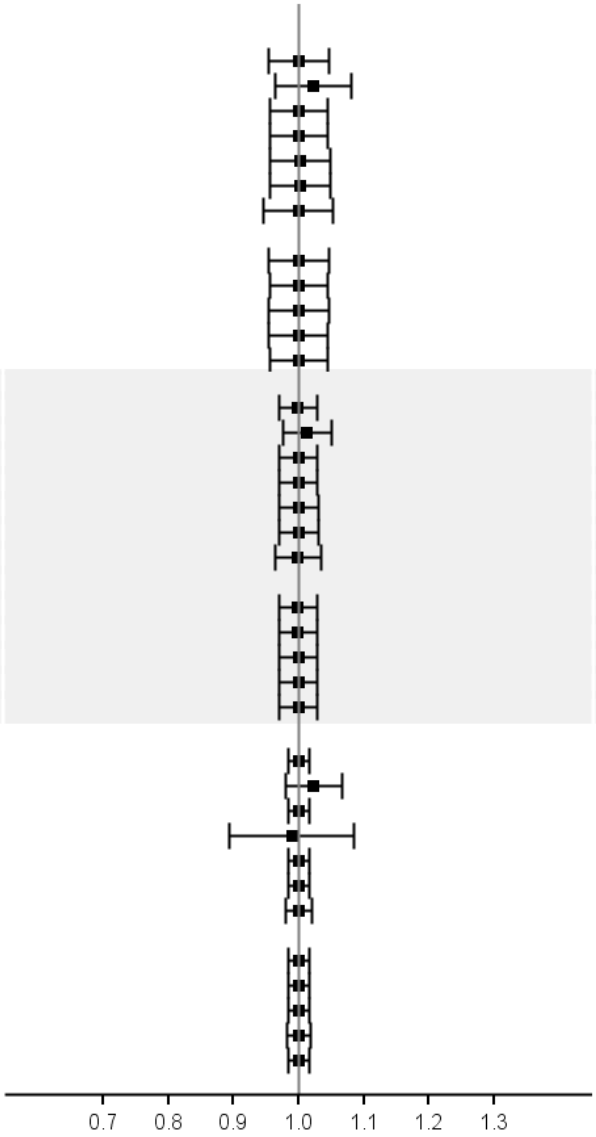
$$\mathbf{b}_i = \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right), \quad t = 1, 2, 3,$$

$$y_{it} \mid x_i, \mathbf{b}_i \sim \text{Negative Binomial}\,(\mu_{it}, \iota_{it}) \quad \log(\mu_{it}) = \beta_0 + b_{i0} + x_i(\beta_1 + b_{i1}),$$

$$\mathbf{b}_i = \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right), \quad t = 1, 2, 3,$$

Over-dispersion in longitudinal counts

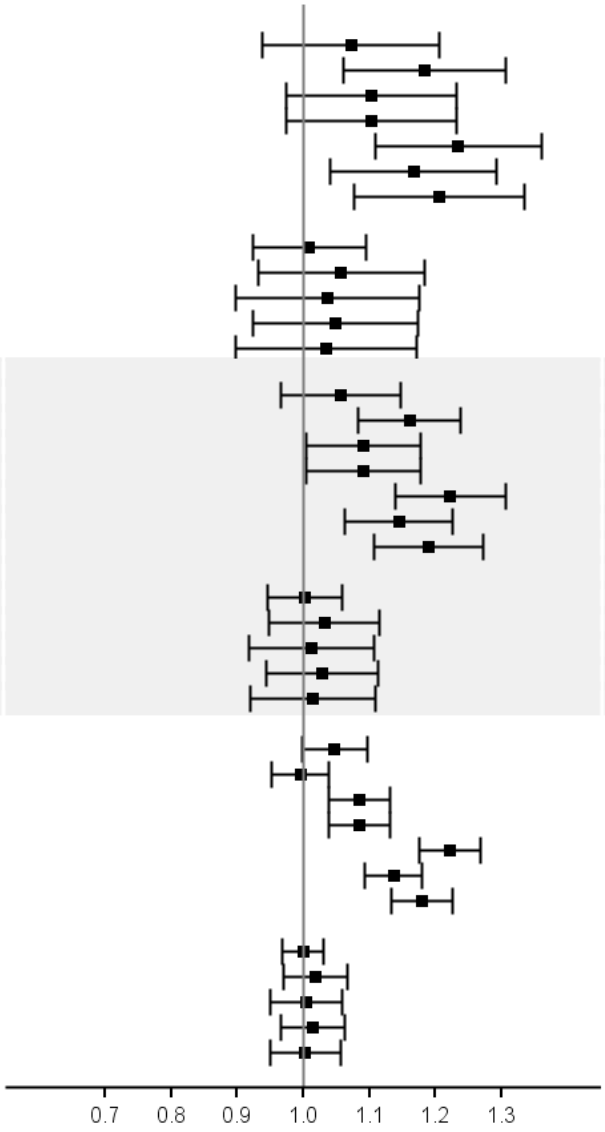
	Package/Procedure	Type I error	Slope Estimate	Slope SD
n=50	R			
	glmmADMB	0.048	1.000	0.046
	glmmAK	0.004	1.022	0.058
	glmmML_GHQ	0.055	1.000	0.045
	glmmML_Laplace	0.055	1.000	0.045
	lme4_Laplace	0.050	1.002	0.047
	lme4_AGQ	0.052	1.002	0.047
	MCMCglmm	0.230	1.000	0.053
	SAS			
	NLMIXED	0.052	1.000	0.046
	GLIMMIX RSPL	0.045	1.000	0.045
	GLIMMIX RMPL	0.051	1.000	0.046
	GLIMMIX MSPL	0.047	0.999	0.045
	GLIMMIX MMPL	0.047	1.000	0.045
n=100	R			
	glmmADMB	0.033	0.999	0.029
	glmmAK	0.012	1.013	0.038
	glmmML_GHQ	0.034	0.999	0.029
	glmmML_Laplace	0.034	0.999	0.029
	lme4_Laplace	0.035	1.000	0.030
	lme4_AGQ	0.038	1.000	0.030
	MCMCglmm	0.263	0.999	0.036
	SAS			
	NLMIXED	0.037	0.999	0.029
	GLIMMIX RSPL	0.036	0.999	0.030
	GLIMMIX RMPL	0.034	0.999	0.030
	GLIMMIX MSPL	0.035	0.999	0.030
	GLIMMIX MMPL	0.034	0.999	0.029
n=300	R			
	glmmADMB	0.038	0.999	0.016
	glmmAK	0.420	1.022	0.043
	glmmML_GHQ	0.042	0.999	0.016
	glmmML_Laplace	0.045	0.989	0.096
	lme4_Laplace	0.041	1.000	0.017
	lme4_AGQ	0.041	1.000	0.017
	MCMCglmm	0.254	1.000	0.019
	SAS			
	NLMIXED	0.042	0.999	0.016
	GLIMMIX RSPL	0.043	1.000	0.016
	GLIMMIX RMPL	0.044	1.000	0.016
	GLIMMIX MSPL	0.044	1.000	0.017
	GLIMMIX MMPL	0.040	1.000	0.016



Small within subject correlation (ρ<0.05), No over-dispersion

Over-dispersion in longitudinal counts

	Package/Procedure	Type I error	Slope Estimate	Slope SD
n=50	R			
	glmmADMB	0.083	1.072	0.134
	glmmAK	0.221	1.184	0.123
	glmmML_GHQ	0.118	1.102	0.129
	glmmML_Laplace	0.119	1.102	0.129
	lme4_Laplace	0.438	1.234	0.126
	lme4_AGQ	0.434	1.167	0.126
	MCMCglmm	0.396	1.206	0.129
	SAS			
	NLMIXED	0.036	1.009	0.085
	GLIMMIX RSPL	0.075	1.057	0.127
	GLIMMIX RMPL	0.064	1.037	0.139
	GLIMMIX MSPL	0.068	1.049	0.125
	GLIMMIX MMPL	0.063	1.034	0.137
n=100	R			
	glmmADMB	0.093	1.057	0.091
	glmmAK	0.419	1.161	0.078
	glmmML_GHQ	0.177	1.091	0.086
	glmmML_Laplace	0.180	1.090	0.086
	lme4_Laplace	0.736	1.223	0.085
	lme4_AGQ	0.604	1.145	0.082
	MCMCglmm	0.641	1.190	0.084
	SAS			
	NLMIXED	0.049	1.001	0.056
	GLIMMIX RSPL	0.059	1.032	0.083
	GLIMMIX RMPL	0.058	1.012	0.095
	GLIMMIX MSPL	0.062	1.029	0.085
	GLIMMIX MMPL	0.063	1.013	0.095
n=300	R			
	glmmADMB	0.124	1.047	0.050
	glmmAK	0.353	0.996	0.043
	glmmML_GHQ	0.375	1.085	0.047
	glmmML_Laplace	0.376	1.085	0.047
	lme4_Laplace	0.999	1.222	0.047
	lme4_AGQ	0.952	1.137	0.044
	MCMCglmm	0.978	1.180	0.046
	SAS			
	NLMIXED	0.043	0.999	0.031
	GLIMMIX RSPL	0.060	1.019	0.048
	GLIMMIX RMPL	0.050	1.004	0.055
	GLIMMIX MSPL	0.060	1.014	0.049
	GLIMMIX MMPL	0.044	1.002	0.053



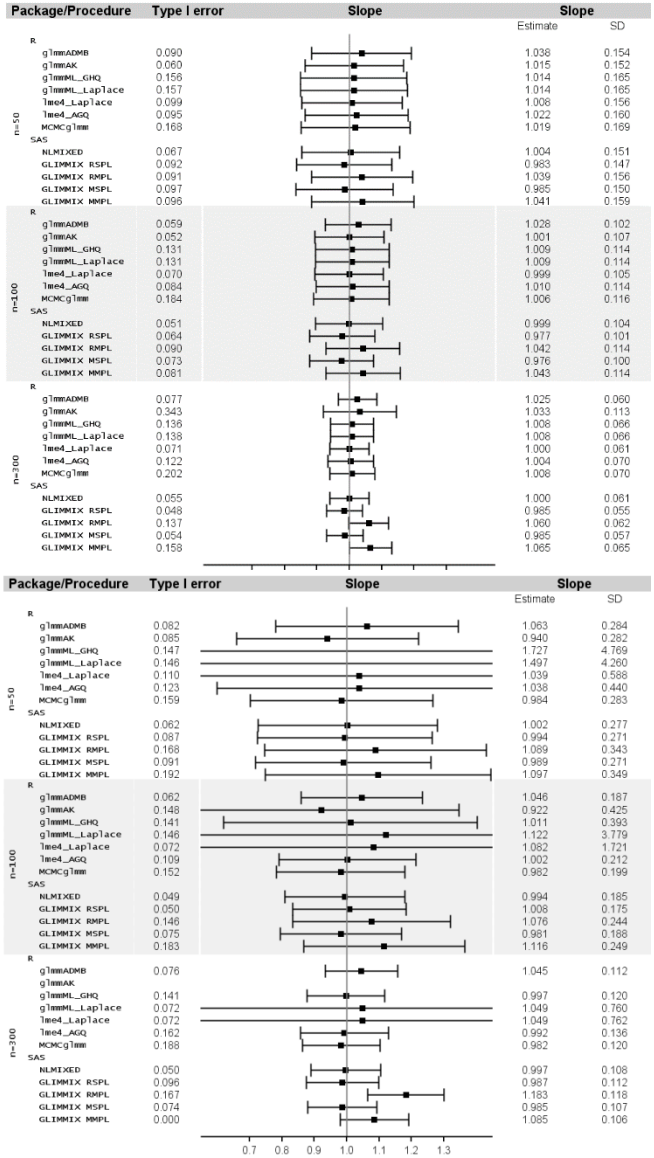
Small within subject correlation (ρ<0.05), over-dispersion = 5

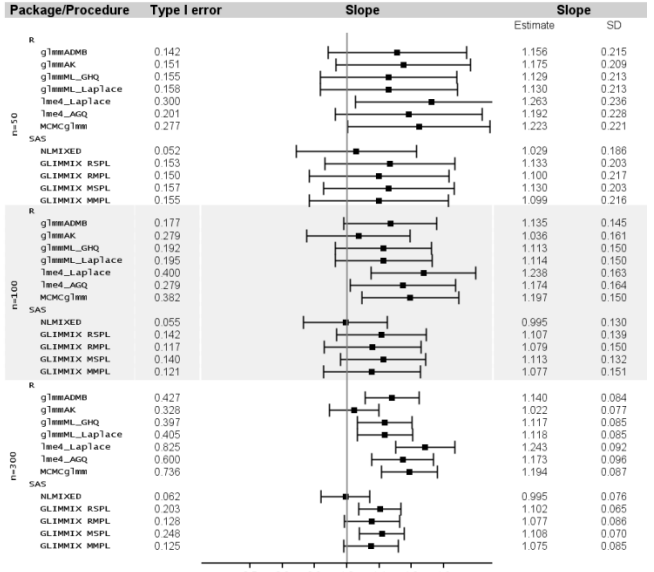
Over-dispersion in longitudinal counts

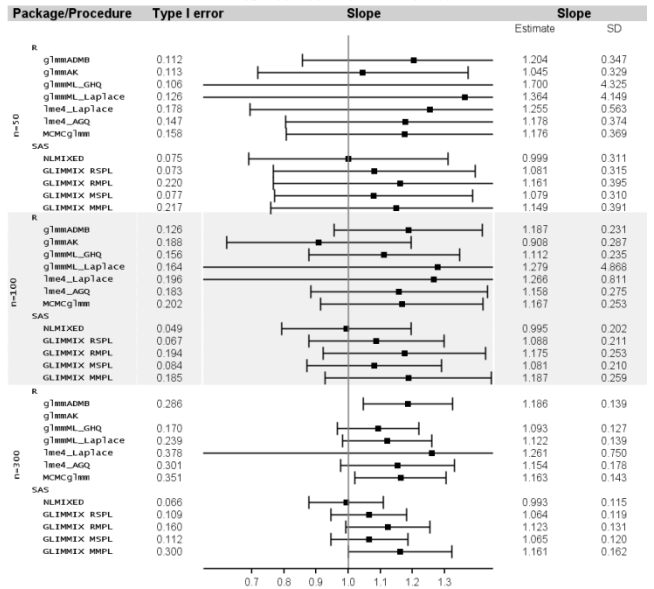
No over-dispersion *over-dispersion = 5*

Medium within subject correlation (ρ=0.5)

large within subject correlation (ρ=0.9)







What are other challenges in overdispersed longitudinal counts?

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

- Detect over-dispersion in longitudinal counts.
- A robust non-parametric method not relying on distribution assumption?
- How to address missing data?

Detection of over-dispersion for longitudinal counts

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280215583397
smm.sagepub.com
SAGE

A non-parametric model to address overdispersed count response in a longitudinal data setting with missingness

Hui Zhang,[1] Hua He,[2] Naiji Lu,[2] Liang Zhu,[1]
Bo Zhang,[3] Zhiwei Zhang[3] and Li Tang[1]

Functional response model (FRM)

Consider distribution-free regression model:

$$\begin{aligned} E\left[\mathbf{f}\left(\mathbf{y}_{j_1},\ldots,\mathbf{y}_{j_q}\right) \mid \mathbf{x}_{j_1},\ldots,\mathbf{x}_{j_q}\right] &= \mathbf{h}\left(\mathbf{x}_{j_1},\ldots,\mathbf{x}_{j_q};\theta\right), \\ \left(j_1,\ldots,j_q\right) &\in C_q^n, \quad 1 \le q \le n, \end{aligned}$$

- $\mathbf{y}_i = \left(y_{i1},\ldots,y_{im}\right)^\top$: the vector of response from the ith subject
- $\mathbf{f}$: vector-valued function
- $\mathbf{h}\left(\theta\right)$: vector-valued smooth function (with continuous derivatives up to the second order)
- θ : vector of parameters of interest
- q : a positive integer, and C_q^n the set of $\binom{n}{q}$ combinations of q distinct elements $\left(j_1,\ldots,j_q\right)$ from the integer set $\left\{1,\ldots,n\right\}$
- We call this model as *functional response model* as it generalizes the single-subject response to a general function of responses from multiple subjects

Functional response model definition

$$
\begin{aligned}
\mathbf{f}_{k\mathbf{i}} &= \mathbf{f}\left(y_{ki}, y_{kj}\right) = \left(f_1\left(y_{ki}, y_{kj}\right), f_2\left(y_{ki}, y_{kj}\right)\right)^{\top}, \\
f_{k1\mathbf{i}} &= f_1\left(y_{ki}, y_{kj}\right) = \frac{1}{2}\left(y_{ki} + y_{kj}\right), \\
f_{k2\mathbf{i}} &= f_2\left(y_{ki}, y_{kj}\right) = \frac{1}{2}\left(y_{ki} - y_{kj}\right)^2, \\
\mathbf{h}_k &= \mathbf{h}\left(\boldsymbol{\theta}_k\right) = \left(\theta_{k1}, \theta_{k2}\right)^{\top} = \left(\mu_k, \sigma_k^2\right)^{\top}, \\
\mathbf{i} &= (i,j) \in C_2^{n_k}, \quad 1 \le k \le K,
\end{aligned}
$$

$$
E\left(\mathbf{f}_{k\mathbf{i}}\right) = E\left(\mathbf{f}\left(y_{ki}, y_{kj}\right)\right) = \mathbf{h}_k = \mathbf{h}\left(\boldsymbol{\theta}_k\right), \mathbf{i} = (i,j) \in C_2^{n_k}, 1 \le k \le K
$$

$$
\widehat{\boldsymbol{\theta}}_k = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^{n_k}} \mathbf{f}\left(y_{ki}, y_{kj}\right) \quad 1 \le k \le K
$$

Extension to longitudinal data

$$
\begin{aligned}
\mathbf{y}_{ki} &= (y_{ki1}, \ldots, y_{kiM})^\top, \quad \boldsymbol{\theta}_{km} = (\theta_{k1m}, \theta_{k2m})^\top = \left(\mu_{km}, \sigma^2_{km}\right)^\top, \\
\boldsymbol{\theta}_k &= \left(\boldsymbol{\theta}^\top_{k1}, \ldots, \boldsymbol{\theta}^\top_{kM}\right)^\top, \\
\mathbf{f}_{k\mathbf{i}} &= \mathbf{f}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}\right) = \left(\mathbf{f}^\top_{k\mathbf{i}1}, \ldots, \mathbf{f}^\top_{k\mathbf{i}M}\right)^\top, \quad \mathbf{f}_{k\mathbf{i}m} = (f_{k1\mathbf{i}m}, f_{k2\mathbf{i}m})^\top, \\
f_{k1\mathbf{i}m} &= f_{1m}\left(y_{kim}, y_{kjm}\right) = \frac{1}{2}\left(y_{kim} + y_{kjm}\right), \\
f_{k2\mathbf{i}m} &= f_{2m}\left(y_{kim}, y_{kjm}\right) = \frac{1}{2}\left(y_{kim} - y_{kjm}\right)^2, \\
\mathbf{h}_k &= \mathbf{h}\left(\boldsymbol{\theta}_k\right) = \boldsymbol{\theta}_k, \quad \mathbf{i} = (i,j) \in C_2^{n_k}, \quad 1 \le k < K, \quad 1 \le m < M, \\
\Longrightarrow E\left(\mathbf{f}_{k\mathbf{i}}\right) &= E\left(\mathbf{f}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}\right)\right) = \boldsymbol{\theta}_k, \quad \mathbf{i} = (i,j) \in C_2^{n_k}, \quad 1 \le k \le K.
\end{aligned}
$$

$$
\widehat{\boldsymbol{\theta}}_k = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^{n_k}} \mathbf{f}\left(y_{ki}, y_{kj}\right)
$$

Theorem 1

For $1 \leq k \leq K$, let

$$\begin{aligned} \mathbf{v}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}\right) &= \mathbf{f}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}\right) - \boldsymbol{\theta}_k, \quad \widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}\right) = E\left(\mathbf{v}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}\right) \mid \mathbf{y}_{ki}\right), \\ \Phi_k &= Var\left(\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}\right)\right) = E\left(\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}\right)\widetilde{\mathbf{v}}^{\top}\left(\mathbf{y}_{ki}\right)\right). \end{aligned}$$

Then, under mild regularity conditions,

$$\widehat{\boldsymbol{\theta}}_k \rightarrow_p \boldsymbol{\theta}_k, \quad \sqrt{n_k}\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\right) \rightarrow_d AN\left(\mathbf{0}, \Sigma_k = 4\Phi_k\right).$$

A consistent estimate of Σ_k is $\widehat{\Sigma}_k = 4\widehat{\Phi}_k$, with $\widehat{\Phi}_k$ given by:

$$\begin{aligned} \widehat{\Phi}_k &= \frac{1}{4}\frac{1}{n_k - 1}\sum_{i=1}^{n_k}\left(\widehat{\mathbf{u}}_{ki} - \widehat{\boldsymbol{\theta}}_k\right)\left(\widehat{\mathbf{u}}_{ki} - \widehat{\boldsymbol{\theta}}_k\right)^{\top}, \\ \widehat{\mathbf{u}}_{ki} &= \left(\widehat{\mathbf{u}}_{ki1}^{\top}, \ldots, \widehat{\mathbf{u}}_{kiM}^{\top}\right)^{\top}, \quad \widehat{\mathbf{u}}_{kim} = \left(y_{kim}, \left(y_{kim} - \widehat{\mu}_{km}\right)^2\right)^{\top}. \end{aligned}$$

Application of theorem 1: detect over-dispersion

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

- For example, $\boldsymbol{\theta} = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mu_3, \sigma_3^2\right)^\top$, test $H_0 : \mu_1 = \sigma_1^2, \mu_2 = \sigma_2^2, \mu_3 = \sigma_3^2$
- Write as, $H_0 : K\boldsymbol{\theta} = 0$, with $K = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 \end{pmatrix}$
- By Theorem 1, $K\widehat{\boldsymbol{\theta}}$ has an asymptotic normal distribution: $\sqrt{n}K\widehat{\boldsymbol{\theta}} \overset{H_0}{\rightarrow}_d AN\left(0, K\Sigma K^\top\right)$
- Therefore, under H_0, $W = n\widehat{\boldsymbol{\theta}}^\top K^\top \left(K\widehat{\Sigma}K^\top\right)^{-1} K\widehat{\boldsymbol{\theta}} \backsim \chi_3^2$

By specifying the matrix K appropriately, we could test various hypothesis.

Application of theorem 1 to testing other hypotheses

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

- For example, $\boldsymbol{\theta} = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mu_3, \sigma_3^2\right)^\top$, test $H_0 : \mu_1 = \mu_2 = \mu_3, \sigma_1^2 = \sigma_2^2 = \sigma_3^2,$
- Write as, $H_0 : K\boldsymbol{\theta} = 0$, with $K = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}$
- By Theorem 1, $K\widehat{\boldsymbol{\theta}}$ has an asymptotic normal distribution: $\sqrt{n}K\widehat{\boldsymbol{\theta}} \overset{H_0}{\to}_d AN\left(0, K\Sigma K^\top\right)$
- Therefore, under H_0, $W = n\widehat{\boldsymbol{\theta}}^\top K^\top \left(K\widehat{\Sigma}K^\top\right)^{-1} K\widehat{\boldsymbol{\theta}} \backsim \chi_4^2$

Simulation result: Type I error

$$\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})^\top \sim Poisson\left(\boldsymbol{\lambda} = (\mu_1, \mu_2, \mu_3)^\top\right)$$

Correlation between time points: ρ=0.2

$\mu_t \mid \sigma_t^2$				$H_0 : \mu_t = \sigma_t^2,\ t = 1, 2, 3$
Sample Size	time 1	time 2	time 3	Type I error
50	5.01 | 4.97	4.99 | 4.97	4.99 | 4.99	0.102
100	5.00 | 4.98	5.01 | 4.98	4.01 | 4.00	0.068
200	5.00 | 5.00	5.00 | 5.00	4.99 | 5.00	0.053
300	5.00 | 5.01	5.00 | 5.01	5.00 | 5.00	0.055

Simulation result: power

Estimates of μ_m and σ_m^2 over time and power estimates from tests of no overdispersion under NB (true O.D=2)						
$\widehat{\mu}_m \vert \widehat{\sigma}_m^2$ (true=5|10)						$H_0 : \mu_k = \sigma_k^2, k = 1, 2, 3$
time 1		time 2		time 3		Estimated Power
Sample size = 50						
4.98	9.99	4.95	9.84	4.94	9.90	0.90
Sample size = 100						
4.98	9.94	4.98	9.94	5.01	10.1	1.00
Sample size = 200						
4.98	9.95	5.00	10.1	5.00	10.0	1
Sample size = 300						
4.99	10.0	5.00	10.0	5.00	10.0	1

Outline

- **Significance of modeling count data**
- **Over-dispersion in cross-sectional counts**
- **Over-dispersion in longitudinal counts**
  Comparison of two popular methods
  Detection over-dispersion in longitudinal counts
  Address missing data
- **Zero-inflation in cross sectional and longitudinal counts**

Introduction to data missing

Classification of DOM Based on Rubin (1976), Little and Rubin (1987), and Little (1995)

- **Missing Completely At Random (MCAR)**: DOM does not depend on covariates or outcomes $P(R_i|x_i, y_i; \theta) = P(R_i|\theta)$
- **Missing At Random (MAR)**: DOM may depend on covariates and observed outcomes $P(R_i|X_i, Y_i; \theta) = P(R_i|X_i, Y_{i(obs)}; \theta)$ Note that $MCAR \subset MAR$.
- **Missing Not At Random (MNAR)**: Any violation of MAR; DOM still depends on $Y_{i(mis)}$ even after any dependence on X_i and $Y_{i(obs)}$.

Extension to missing data

- $r_{kim} = \begin{cases} 1 & \text{if } y_{kim} \text{ is observed} \\ 0 & \text{if } y_{kim} \text{ is missing} \end{cases}$, $\mathbf{r}_{ki} = (r_{ki1}, \ldots, r_{kiM})^\top$
- Let $\pi_{kim} = \Pr(r_{kim} = 1 \mid \mathbf{y}_{ki})$, the monotone missing data pattern (MMDP) assumption:

$$\begin{aligned} \pi_{kim} &= \Pr(r_{kim} = 1 \mid \mathbf{y}_{ki}) = \Pr(r_{kim} = 1 \mid \tilde{\mathbf{y}}_{kim}), \\ \tilde{\mathbf{y}}_{kim} &= \{y_{kis}; 1 \le s \le m-1\}, \qquad 1 \le k \le K, 2 \le m \le M \end{aligned}$$

- We model π_{kim} using a logistic regression as follows:

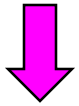
$$\begin{aligned} \text{logit}\ (p_{kim}) &= \text{logit}\left(\Pr\left(r_{kim} = 1 \mid r_{ki(m-1)} = 1, \tilde{\mathbf{y}}_{kim}\right)\right) \\ &= \alpha_{km} + \boldsymbol{\beta}_{km}^\top \tilde{\mathbf{y}}_{kim}, \\ \pi_{kim}(\gamma_k) &= \prod_{s=2}^{m} p_{kis}(\gamma_{ks}), \quad 2 \le m \le M, \quad 1 \le k \le K, \end{aligned}$$

where $\gamma_{km} = \left(\alpha_{km}, \boldsymbol{\beta}_{km}^\top\right)^\top$ and $\gamma_k = \left(\gamma_{k2}^\top, \ldots, \gamma_{kM}^\top\right)^\top$

Extension to missing data

$$\text{logit}\ (\pi_{kit}) = \text{logit}\left(\Pr\left(r_{kit}=1 \mid \widetilde{\mathbf{y}}_{kit}\right)\right) = \alpha_{kt} + \boldsymbol{\beta}_{kt}^{\top}\widetilde{\mathbf{y}}_{kit}, \quad 2 \leqslant t \leqslant T$$

$$\widehat{\alpha}_{kt}, \widehat{\boldsymbol{\beta}}_{kt} \Longrightarrow \widehat{\pi}_{kit}$$



$$\widehat{\boldsymbol{\theta}}_k = \binom{n_k}{2}^{-1} \sum_{\mathbf{i} \in C_2^{n_k}} \mathbf{g}_{k\mathbf{i}} = \binom{n_k}{2}^{-1} \sum_{(i,j) \in C_2^{n_k}} \mathbf{g}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}; \widehat{\mathbf{r}}_{ki}, \widehat{\mathbf{r}}_{kj}\right)$$

$$\mathbf{g}_{ki\mathbf{t}} = \mathbf{g}\left(y_{kit}, y_{kjt}; r_{kit}, r_{kjt}\right) = \frac{r_{kit} r_{kjt}}{\pi_{kit}\pi_{kjt}} \mathbf{h}_t\left(y_{kit}, y_{kjt}\right),$$

$$\mathbf{g}_{k\mathbf{i}} = \mathbf{g}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}; \mathbf{r}_{ki}, \mathbf{r}_{kj}\right) = \left(\mathbf{g}_{k\mathbf{i}1}^{\top}, \ldots, \mathbf{g}_{k\mathbf{i}T}^{\top}\right)^{\top}, \quad 1 \leqslant k \leqslant K.$$

$$\widehat{\boldsymbol{\theta}}_k \to_p \boldsymbol{\theta}_k, \quad \sqrt{n_k}\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\right) \to_d N\left(\mathbf{0}, \Sigma_k = 4\left(\Phi_k + \boldsymbol{\Psi}_k\right)\right)$$

Extension to missing data

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

For $1 \leq k \leq K$, let

$$
\begin{aligned}
v\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}, \mathbf{r}_{ki}, \mathbf{r}_{kj}\right) &= g_{k\mathbf{i}} - \theta_k, \\
\widetilde{\mathbf{g}}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right) &= E\left(\mathbf{g}\left(\mathbf{y}_{ki}, \mathbf{y}_{kj}, \mathbf{r}_{ki}, \mathbf{r}_{kj}\right) \mid \mathbf{y}_{ki}, \mathbf{r}_{ki}\right), \\
\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right) &= \widetilde{\mathbf{g}}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right) - \theta_k, \\
\Phi_k &= Var\left(\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right)\right) = E\left(\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right) \widetilde{\mathbf{v}}^{\top}\left(\mathbf{y}_{ki}, \mathbf{r}_{ki}\right)\right) \\
C_k &= E\left(\frac{\partial^{\top}}{\partial \gamma_k} \widetilde{\mathbf{g}}\left(\mathbf{y}_{ki}; \mathbf{r}_{ki}, \gamma_k\right)\right), \quad H_k = E\left(\frac{\partial^{\top}}{\partial \gamma_k} \mathbf{w}_{ki}\left(\gamma_k\right)\right) \\
F_k &= E\left(\widetilde{\mathbf{v}}\left(\mathbf{y}_{ki}; \mathbf{r}_{ki}, \gamma_k\right) \mathbf{w}_i^{\top} H_k^{-1} C_k^{\top}\right), \\
\Psi_k &= -\left(C_k H_k^{-1} C_k^{\top} + F_k + F_k^{\top}\right).
\end{aligned}
$$

Simulation result with missing: power

Estimates of μ_m / σ^2_m over time and power estimates from tests of null hypothesis of no overdispersion under NB(O.D=1.5)					
n	α	Mean | Variance			$H_0 : \mu_k = \sigma^2_k$
		time 1	time 2	time 3	Power
50	0.1	4.98 | 7.49	4.94 | 7.40	4.94 | 7.41	0.381
100	0.1	4.98 |7.47	4.98 | 7.47	5.00 | 7.53	0.786
200	0.1	4.98 | 7.46	5.00 | 7.54	5.00 | 7.54	0.991
300	0.1	4.99 | 7.51	5.00 | 7.51	5.00 | 7.53	1

Two major problems using Poisson

- Over-dispersion

- Zero-inflation

Why to discuss zero-inflation?

A small piece of real RNA-Seq data (18 of 434480 rows, 5 of 22 subjects) from Acute megakaryoblastic leukemia (AML-M7) patients are shown:

Chr	Start	End	GC	01D	01G	02D	02G	03D	03G	04D	04G	05D	05G
3	183528262	183528361	38	53	65	95	103	92	80	56	65	71	64
3	183528362	183528371	6	20	43	52	58	54	51	39	39	41	31
3	183535143	183535224	47	17	6	12	2	25	3	7	10	3	12
3	183535781	183535880	59	258	253	344	296	427	315	230	292	233	250
3	183535881	183535899	10	123	164	238	177	211	140	126	174	131	152
3	183542934	183543033	76	19	26	11	11	25	7	14	9	4	10
3	183543034	183543133	42	6	8	3	3	10	1	4	2	1	5
3	183543134	183543233	30	0	0	3	0	1	0	0	0	0	1
3	183543234	183543333	72	2	0	4	0	0	0	0	0	0	0
3	183543334	183543335	0	2	0	2	0	0	0	0	0	0	0
10	103534886	103534973	66	7	0	0	9	4	1	3	1	1	1
10	103535496	103535533	24	0	4	0	1	4	2	0	0	0	1
10	103535625	103535657	25	0	0	0	0	1	0	0	0	0	0
10	111674768	111674857	38	0	0	0	0	0	1	0	0	0	0
10	111683159	111683191	20	0	0	0	0	0	0	0	0	0	4
18	48345950	48346049	65	0	0	0	2	1	0	0	0	0	0
18	48346050	48346072	9	0	0	0	0	0	0	0	0	0	0
18	48346266	48346285	12	0	0	0	0	0	1	0	0	0	0

Zero-inflated Poisson (ZIP)

$$\text{Mixed distribution} \begin{cases} \text{constant } 0 & prob = \rho \\ \text{Poisson}\,(\lambda) & prob = 1-\rho \end{cases}$$

$$PMF = \begin{cases} \rho + (1-\rho)\, e^{-\lambda} & , y = 0 \\ (1-\rho)\, \frac{\lambda^y e^{-\lambda}}{y!} & , y \geqslant 1 \end{cases}$$

$$\text{Moments} \begin{cases} \mu = (1-\rho)\,\lambda \\ \sigma^2 = \lambda\,(1-\rho)\,(1-\lambda\rho) \end{cases}$$

A new approach to address zero-inflation

The Canadian Journal of Statistics

Vol. xx, No. yy, 20??, Pages 1–29

La revue canadienne de statistique

Distribution-free Models for Latent Mixed Population Responses in a Longitudinal Setting with Missing Data

BlindedA[1]* and BlindedB[2]

Illustration of the example

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

- $\begin{cases} y_{i1} - \text{RNA-Seq counts for a single gene from multiple patients} \\ y_{i2} - \#\text{ of days showing a specific symptom during last month} \end{cases}$

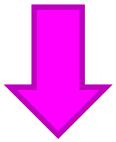
- The models for the count and primary outcomes under a parametric setup are

$$\begin{aligned} y_{i1} \quad &| \quad \mu, \rho \sim ZIP\left(\mu, \rho\right), \\ y_{i2} \quad &| \quad m_i, p_r, r_i \sim \text{Bin}\left(m_i, \left(1 - r_i\right) p_0 + r_i p_1\right), \\ r_i \quad &= \quad \begin{cases} 0 & \text{if no transcription} \\ 1 & \text{if transcription} \end{cases} . \end{aligned}$$

New definition of FRM

The mean and variance of the marginal ZIP and the mean of the Binomial outcome are

$$
\begin{aligned}
E\left(y_{i1}\right) &= \left(1-\rho\right)\mu \\
Var\left(y_{i1}\right) &= \left(1-\rho\right)\left(1+\rho\mu\right)\mu \\
E\left(y_{i2}\right) &= \rho m_i p_0 + \left(1-\rho\right) m_i p_1 \\
E\left(y_{i2} I_{\{y_{i1}>0\}}\right) &= \left(1-\rho\right)\left(1-e^{-\mu}\right) m_i p_1
\end{aligned}
$$



$$
\begin{aligned}
E\left(\mathbf{f}_i\right) &= \mathbf{h}_i, \quad \mathbf{f}_i = \left(\mathbf{f}_{i1}^\top, \mathbf{f}_{i2}^\top, \mathbf{f}_{i3}^\top, \mathbf{f}_{i4}^\top\right)^\top, \quad \mathbf{h}_i = \left(\mathbf{h}_{i1}^\top, \mathbf{h}_{i2}^\top, \mathbf{h}_{i3}^\top, \mathbf{h}_{i4}^\top\right)^\top \\
\mathbf{f}_i &= \left(f_{1i}, f_{2i}, f_{3i}, f_{4i}\right)^\top = \left(y_{i1}, y_{i1}^2, y_{i2}, y_{i2} I_{\{y_{i1}>0\}}\right)^\top, \\
\mathbf{h}_i &= \left(h_{1i}, h_{2i}, h_{3i}, h_{4i}\right)^\top, \quad 1 \le i \le n, \quad 1 \le t \le m. \\
h_{1i} &= \left(1-\rho\right)\mu, \quad h_{3i} = \rho m_i p_0 + \left(1-\rho\right) m_i p_1, \\
h_{2i} &= \mu\left(1-\rho\right)\left(1+\mu\right), \quad h_{4i} = \left(1-\rho\right)\left(1-e^{-\mu}\right) m_i p_1
\end{aligned}
$$

Extension to regression

- Let $\mathbf{u}_i$, $\mathbf{v}_i$, and $\mathbf{w}_i$ denote the covariates for the ZIP and Binomial models. Then, the marginal models are

$$\begin{aligned}
y_{i1} \quad &| \quad \mu_i, \rho_i \sim ZIP\left(\mu_i, \rho_i\right), \\
y_{i2} \quad &| \quad m_i, p_{r_i i}, r_i \sim \text{Bin}\left(m_i, \left(1 - r_i\right) p_{0i} + r_i p_{1i}\right) \\
r_i \quad &= \quad \begin{cases} 0 & \text{if no transcription for subject } i \\ 1 & \text{if transcription for subject } i \end{cases}. \\
\text{logit}\left(\rho_i\right) \quad &= \quad \mathbf{u}_i^\top \boldsymbol{\beta}_u, \quad \log\left(\mu_i\right) = \mathbf{v}_i^\top \boldsymbol{\beta}_v, \\
\text{logit}\left(p_{0i}\right) \quad &= \quad \mathbf{w}_i^\top \boldsymbol{\beta}_{0w}, \quad \text{logit}\left(p_{1i}\right) = \mathbf{w}_i^\top \boldsymbol{\beta}_{1w}.
\end{aligned}$$

- Define $\mathbf{f}_i$ and $\mathbf{h}_i$ the same as in the homogeneous case, but expand the model to include the above link functions to link ρ_i, μ_i, p_{0i} and p_{1i} to the respective covariates. The FRM is

$$\begin{aligned}
E\left(\mathbf{f}_i \mid \mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i\right) \quad &= \quad \mathbf{h}_i\left(\boldsymbol{\beta}\right), \quad \boldsymbol{\beta} = \left(\boldsymbol{\beta}_u^\top, \boldsymbol{\beta}_v^\top, \boldsymbol{\beta}_{0w}^\top, \boldsymbol{\beta}_{1w}^\top\right)^\top, \\
1 \quad &\leq \quad i \leq n, 1 \leq t \leq m.
\end{aligned}$$

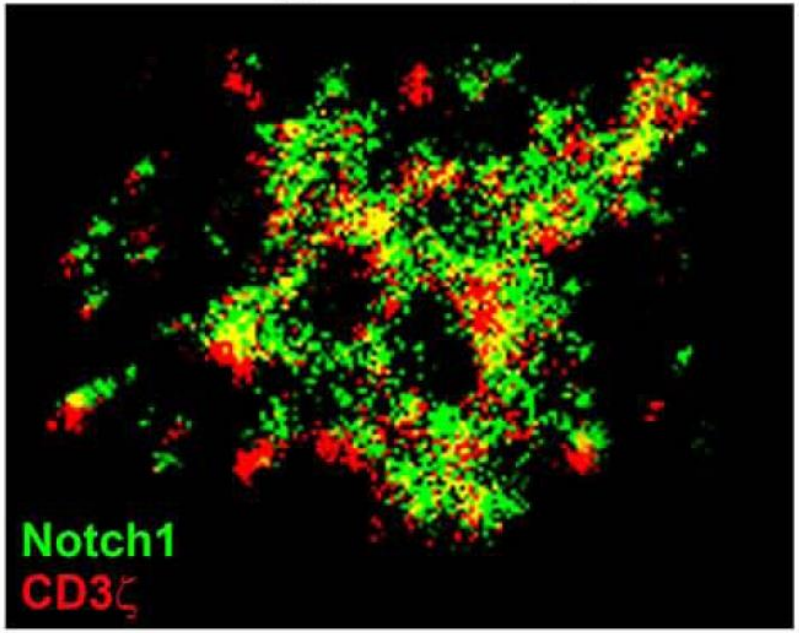
Outline

- **Significance of modeling count data**
- **Over-dispersion in cross-sectional counts**
- **Over-dispersion in longitudinal counts**
  Comparison of two popular methods
  Detection over-dispersion in longitudinal counts
  Address missing data
- **Zero-inflation in cross sectional and longitudinal counts**
- **An example of future research projects**

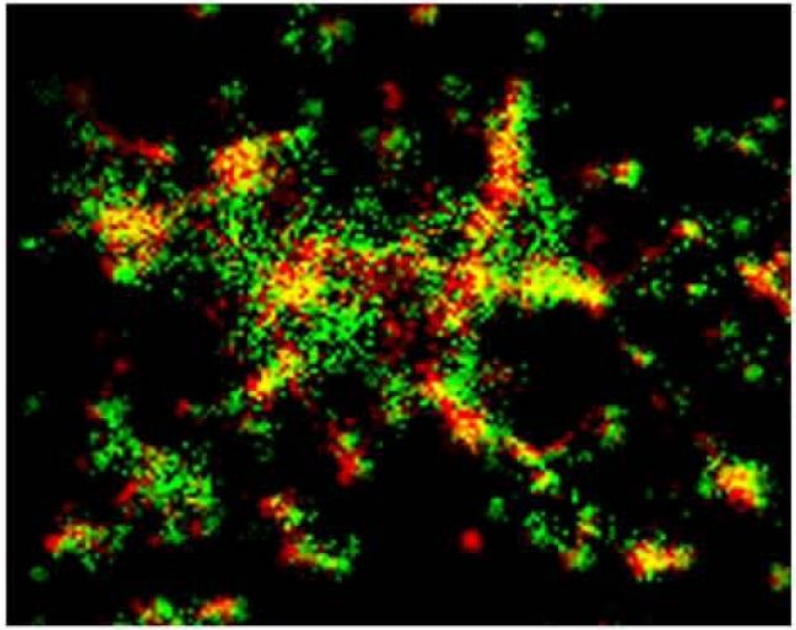
One of future research projects

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

Wild-type CD3 complex



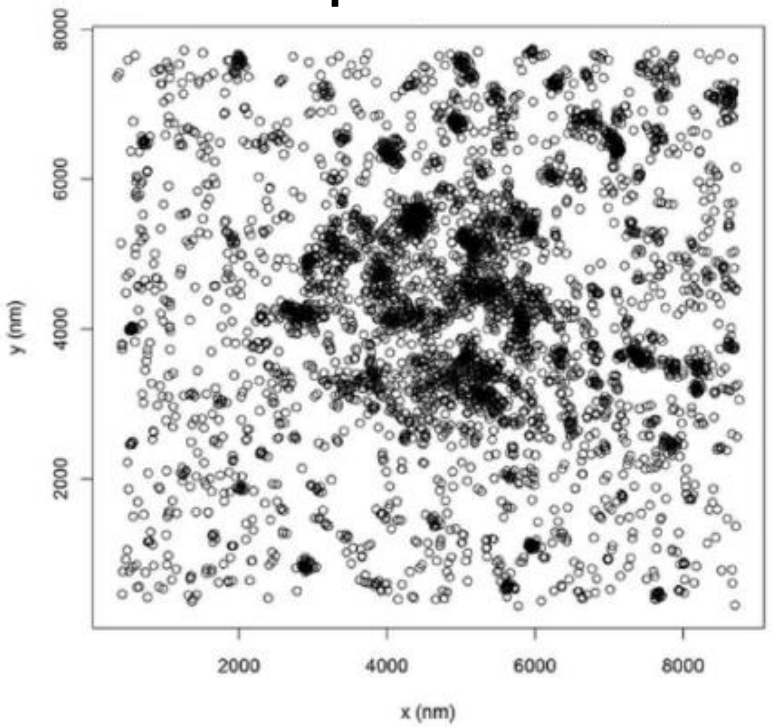
CD3εPRS



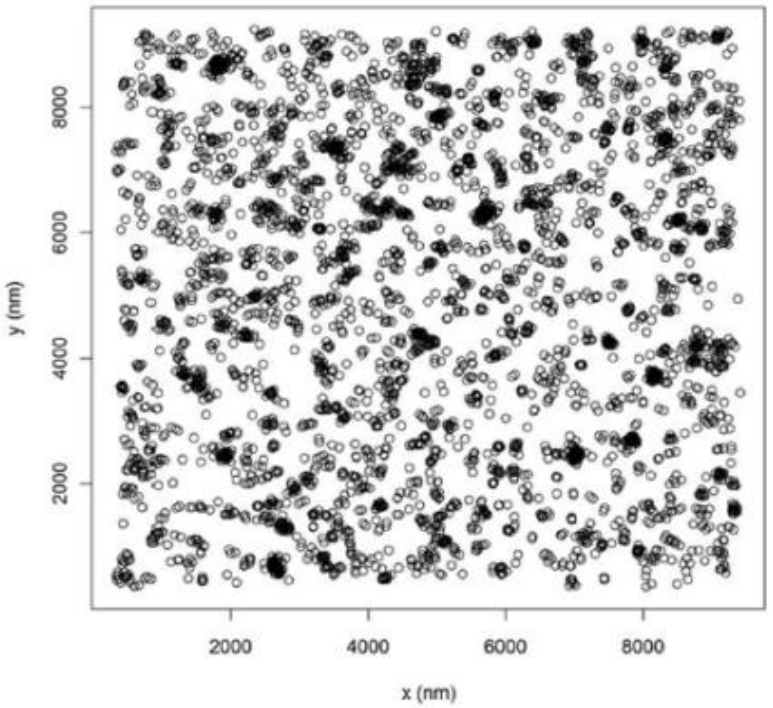
One of future research projects

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

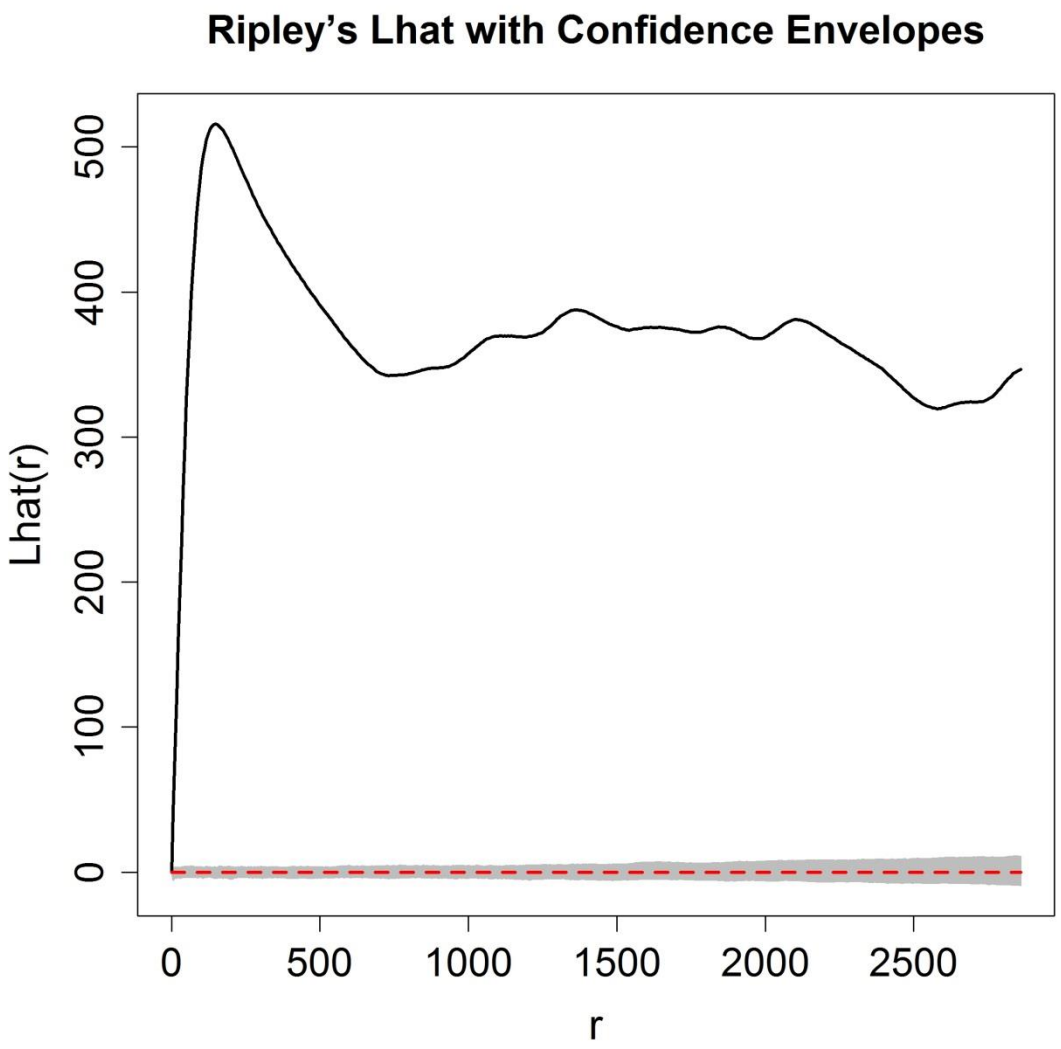
Experimental



Control



One of future research projects



One of future research projects

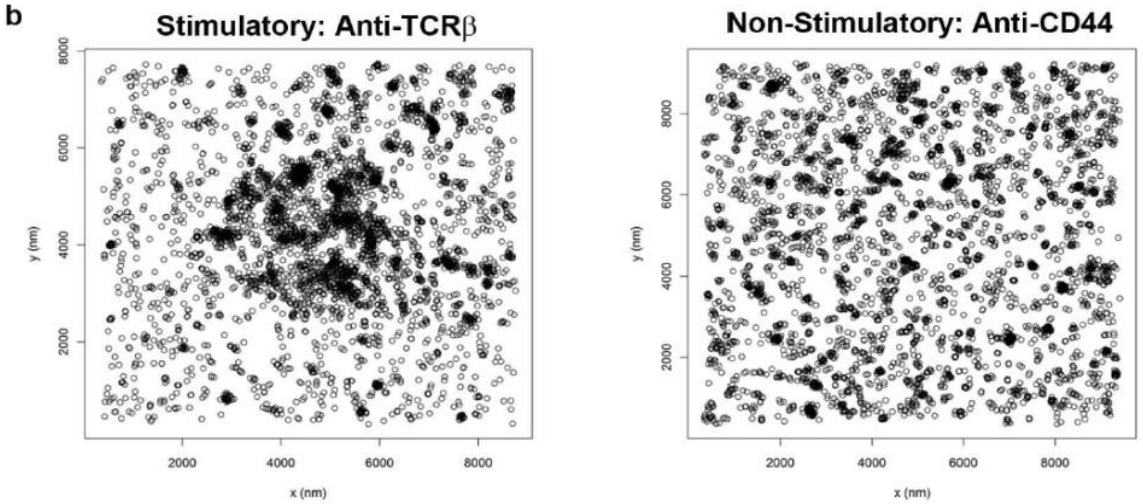
St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

Motivating current collaborating project:

VOLUME 14 NUMBER 3 MARCH 2013 NATURE IMMUNOLOGY

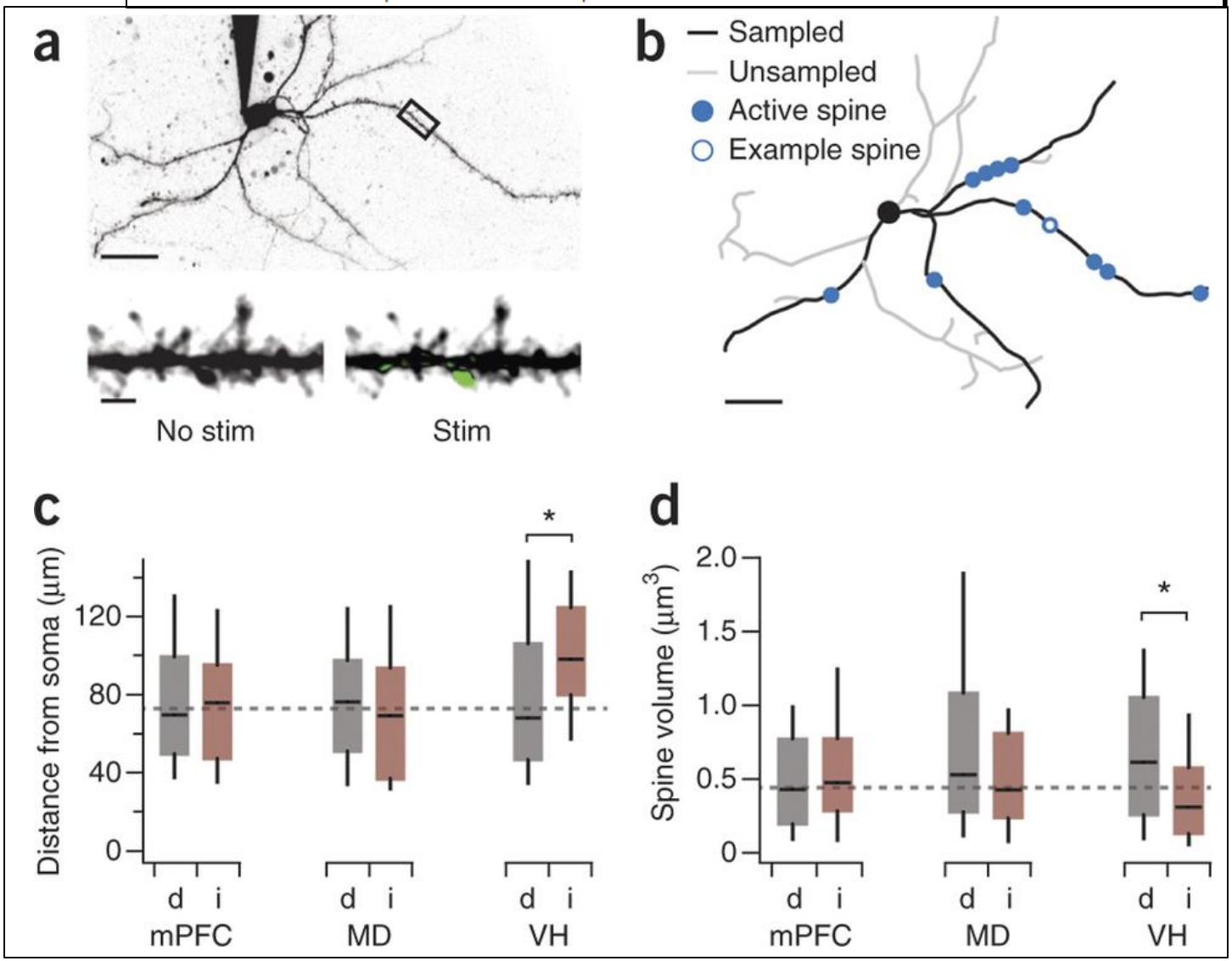
Distinct TCR signaling pathways drive proliferation and cytokine production in T cells

Clifford S Guy[1], Kate M Vignali[1], Jamshid Temirov[2], Matthew L Bettini[1], Abigail E Overacre[1], Matthew Smeltzer[3], Hui Zhang[3], Johannes B Huppa[4], Yu-Hwai Tsai[5], Camille Lobry[6], Jianming Xie[7], Peter J Dempsey[5], Howard C Crawford[8], Iannis Aifantis[6], Mark M Davis[7] & Dario A A Vignali[1]



One of future research projects

VOLUME 15 | NUMBER 12 | DECEMBER 2012 NATURE NEUROSCIENCE



Summary

- **Significance of modeling count data**
- **Over-dispersion in cross-sectional counts**
- **Over-dispersion in longitudinal counts**
  Comparison of two popular methods
  Detection over-dispersion in longitudinal counts
  Address missing data
- **Zero-inflation in cross sectional and longitudinal counts**
- **An example of future research projects**