Understanding and Measuring the Impact of Distance on Health

Evidence from Two Studies

Mahesh Karra
Pardee School of Global Studies
Boston University
October 3, 2017

Background and Motivation

Background

- Despite progress to reducing child mortality, nearly 18,000 children under 5 die every day
- Many of these deaths could be avoidable with increased utilization of health services
- But health service utilization by women around the world remains low

Motivation

- A large theoretical and empirical literature on geographical determinants for health care seeking and MCH outcomes
- Role of physical access (travel distance) to services
- Evidence of association between distance to facility and utilization has been generally consistent
- Empirical evidence on association between distance to facility and health outcomes (e.g. child mortality) is limited and mixed
- Methodological concerns around how distance is measured
  - Travel distance (Euclidean, road), travel time
  - Issues around measurement error and bias in distance

Objectives

- To understand how distance is related to utilization and health
- To explore measurement problems with distance data
- To propose a methodological solution to these problems

Objectives

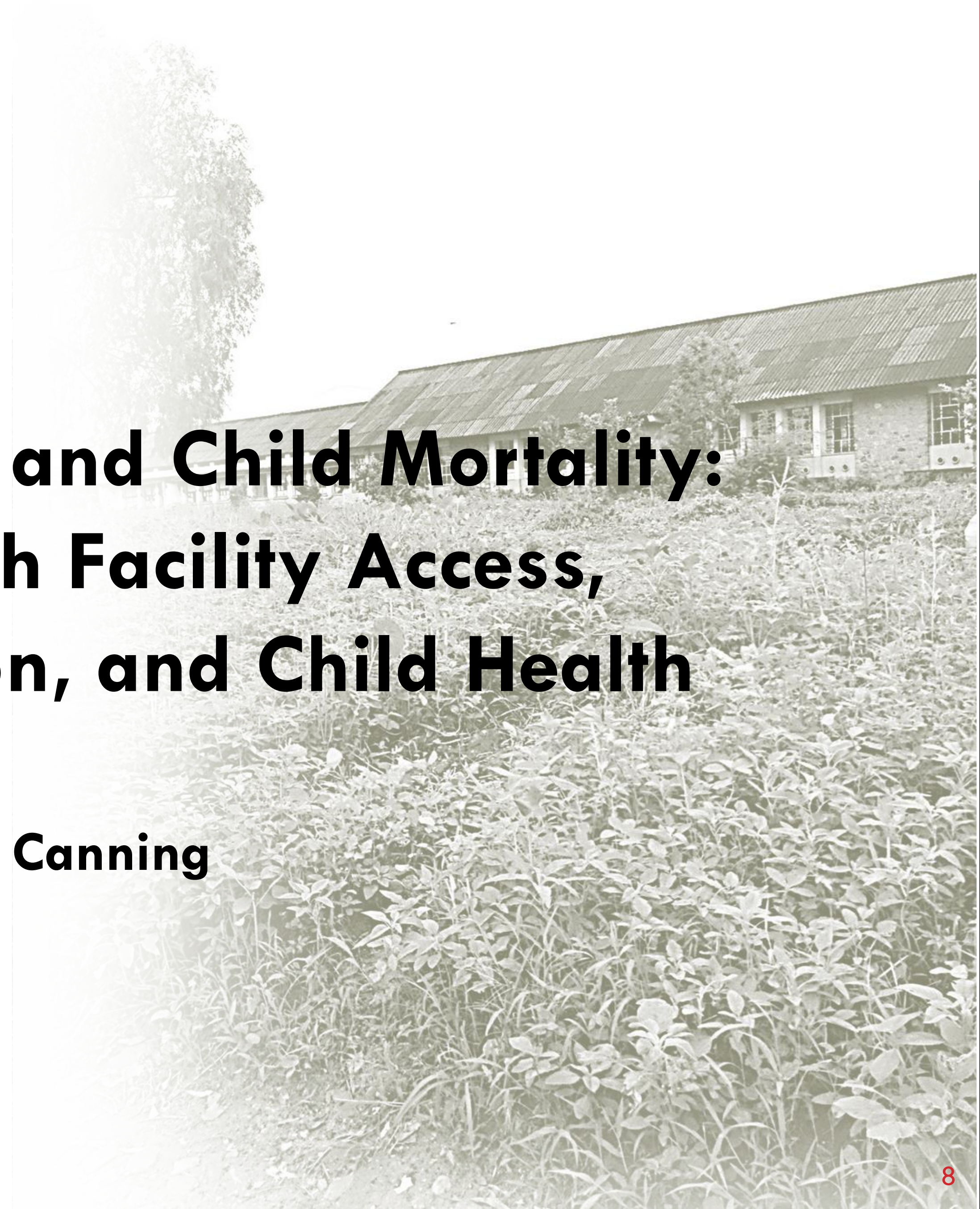
Study 1 Objectives

- To empirically examine the relationships between
  - Travel distance to facility and health care utilization
    - Receipt of antenatal care
    - Delivery in a health facility
  - Travel distance to facility and health
    - Child mortality

Objectives

Study 2 Objectives

- To develop a theory that allows for unbiased and consistent estimation when we have deliberately induced measurement error in our distance data
  - And mismeasured explanatory variables, more generally



Facility Distance and Child Mortality: A Study of Health Facility Access, Service Utilization, and Child Health

M. Karra, G. Fink, and D. Canning

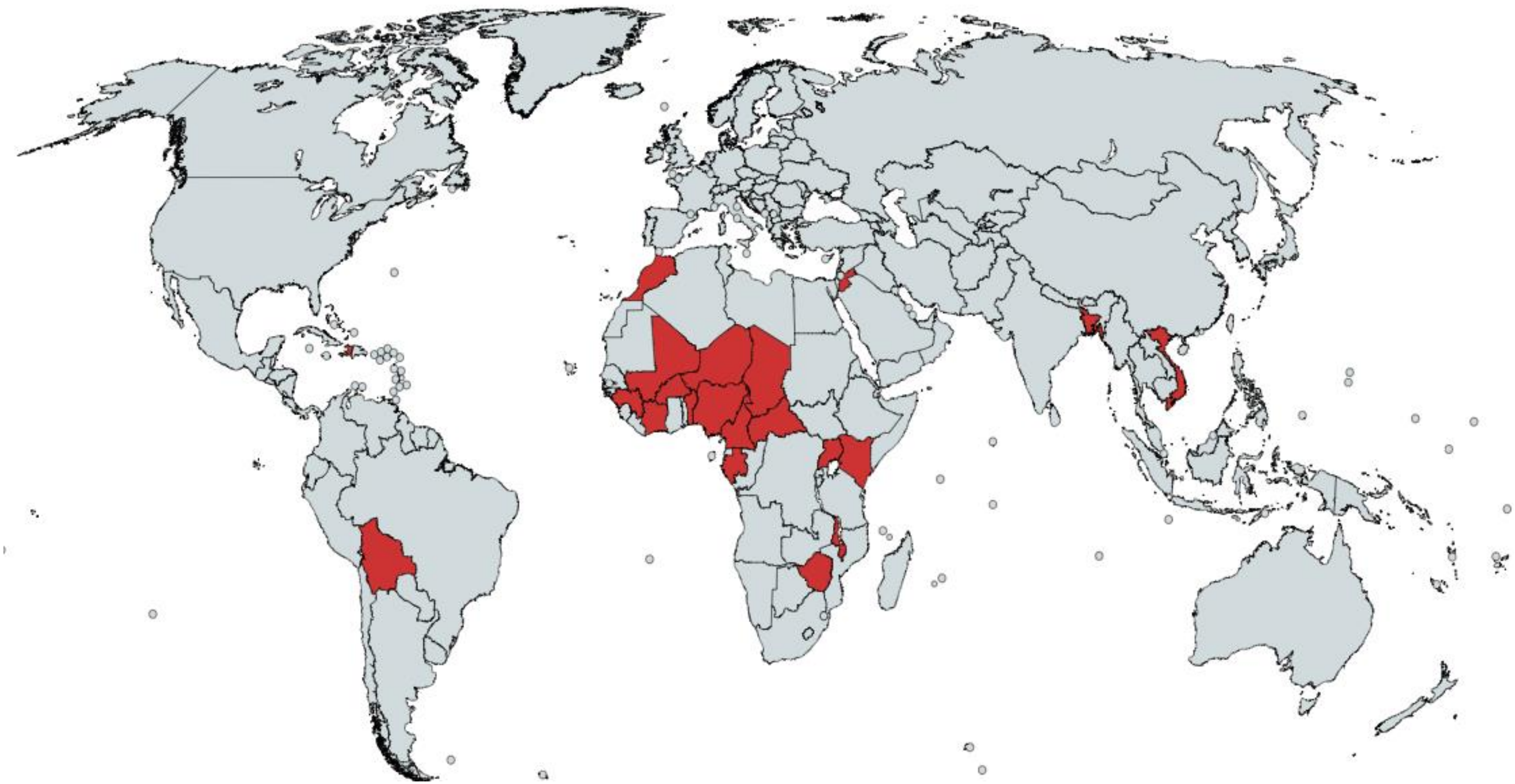
Objectives

- To examine the relationships between
  - Travel distance to facility and maternal health care utilization
    - Receipt of antenatal care (WHO-recommended 4 visits)
    - Delivery in a health facility
  - Travel distance to facility and child mortality
    - Disaggregated into neonatal, post-neonatal infant, and post-infant child

Data and Methods

- Pool data from Demographic and Health Surveys
  - 126,835 births to 124,719 mothers across 7,901 DHS clusters in 21 countries across 29 DHS surveys between 1990 and 2011
- Travel distance from DHS Service Availability Questionnaire (SAQ)
  - Administered at DHS cluster level
  - Group interview with 3-4 key informants in cluster
  - Informants identify nearest facility of each type from cluster
    - Hospital, health center, clinic, pharmacy, others

Countries



Distance Data – The SAQ

For each facility type:

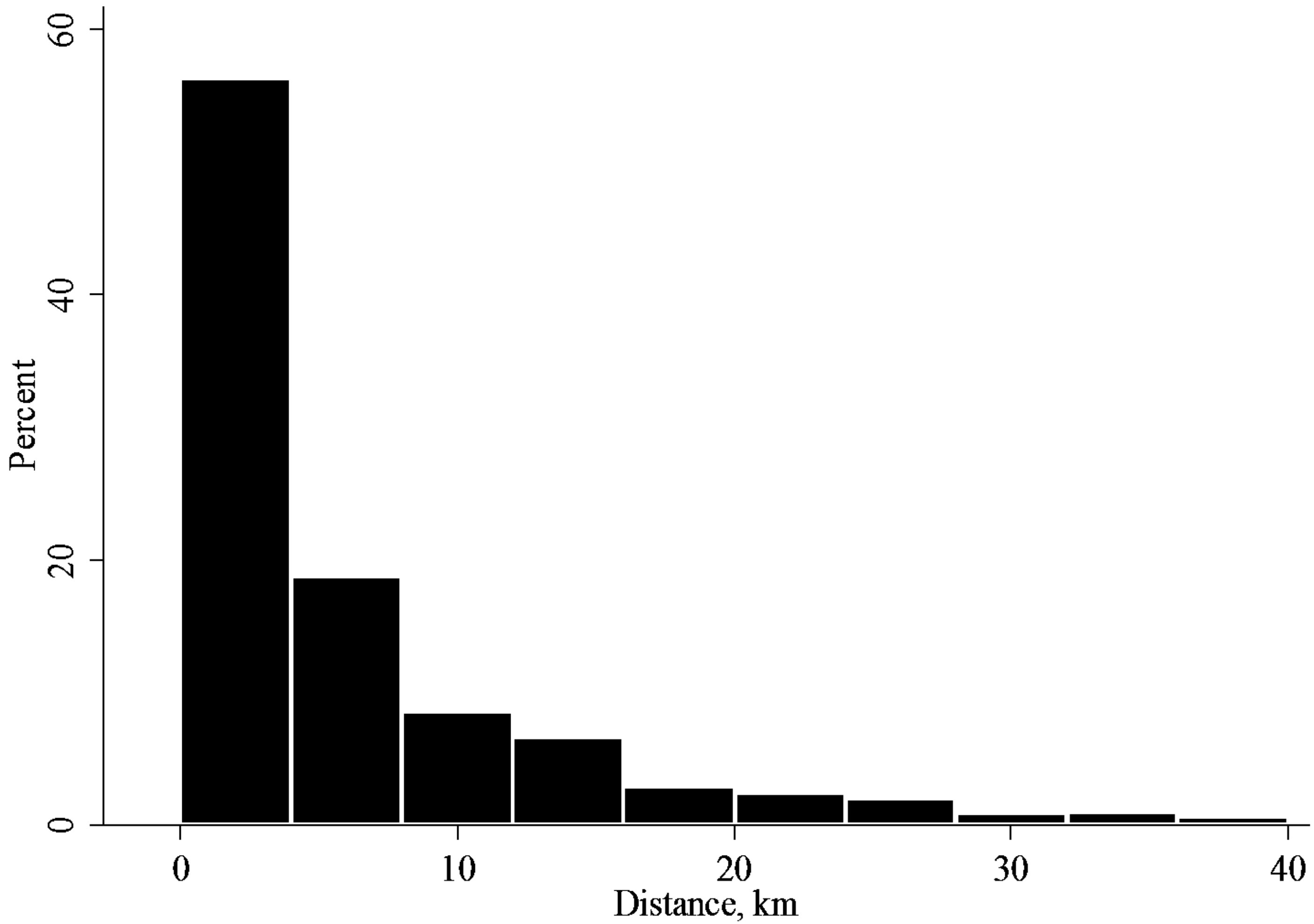
1. How far in miles/km is the facility located from the cluster center?
2. Most common mode of transportation that is used to go to this facility?
3. How long (minutes/hours) does it take to go to the facility using the most common type of transportation?

- Following interview, facilities that were mentioned are visited by enumerator
- Advantages over using DHS GPS locations to match clusters to facilities
    - **Avoids the bias induced by spatial displacement of clusters**
    - Arguably more meaningful than straight-line distances

Distance Variable

- We consider reported distances to one of 4 facility types:
  - Nearest hospital
  - Nearest doctor or low-tiered clinic
  - Nearest mid-level health center
  - Nearest MCH center
- Calculate minimum distance to any of these 4 facility types
- Divide the distance variable into interval categorical variable
  - < 1 km to nearest facility, 1-2 km, 2-3 km, 3-5 km, 5-10 km, > 10 km

Distances to the Nearest Facility



Main Analysis

- Dependent variables for health care utilization:
    - Receipt of WHO-recommended 4 or more ANC visits
    - Whether or not the birth was delivered in a health facility
- Dependent variables for child mortality:
    - Child mortality (neonatal, post-neonatal infant, post-infant child)
- Main independent variable:
    - Interval categorical distance to nearest facility
- Analysis:
    - Multivariate logistic regression, reported odds ratios

Main Results: Utilization

Distance is strongly, inversely associated with service utilization

- Compared to living < 1 km from a facility, living > 10 km from a facility:
    - 38.8 percent lower odds of receiving 4 ANC visits
    - 55.3 percent lower odds of delivering in a facility

- Very similar findings when using time to facility
- Robust to alternative specifications
    - In-patient facilities only, non-migrating mothers, urban/rural, controlling for distance to other locations (school, market)

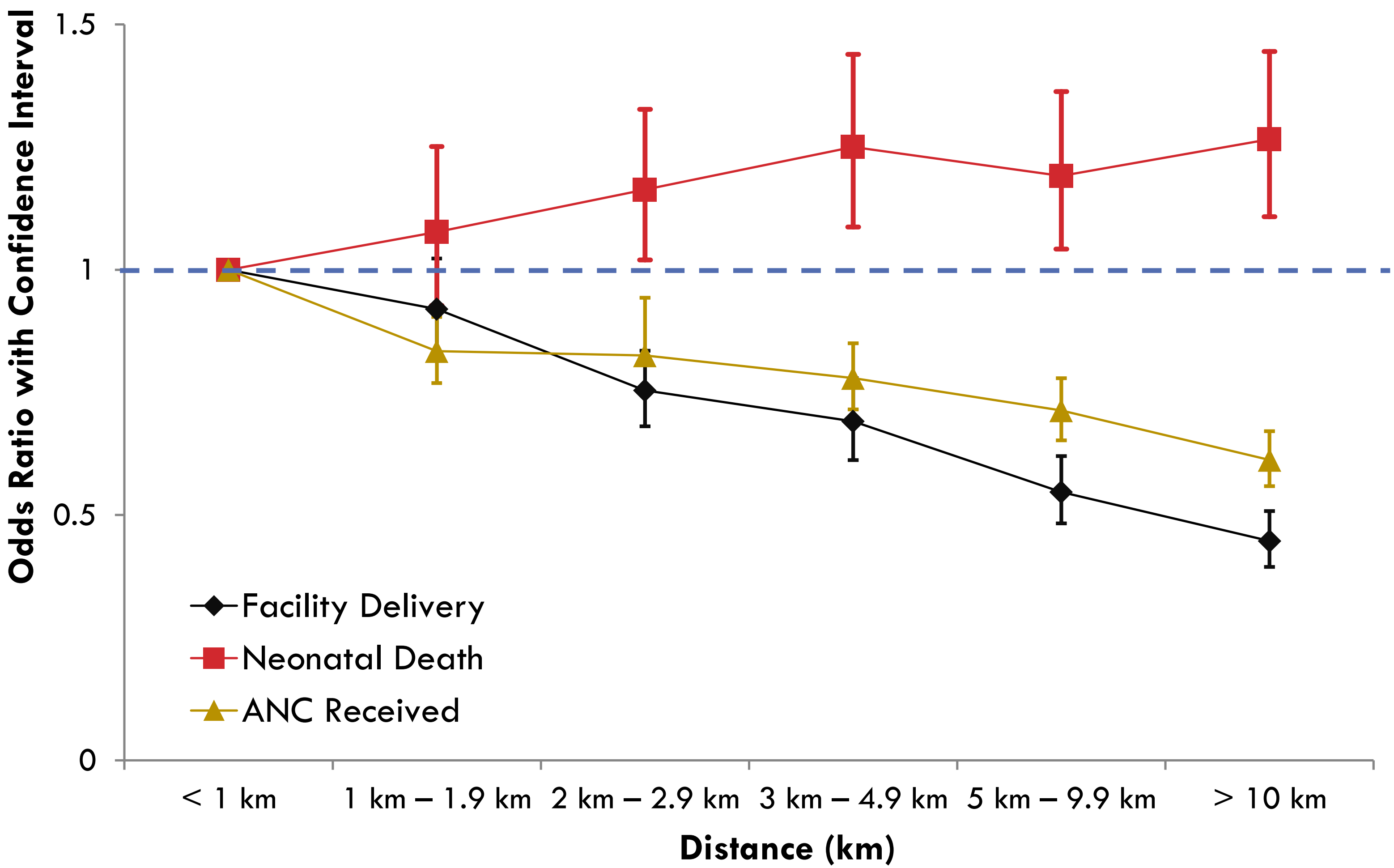
Main Results: Mortality

Distance is positively associated with child mortality (specifically in young children)

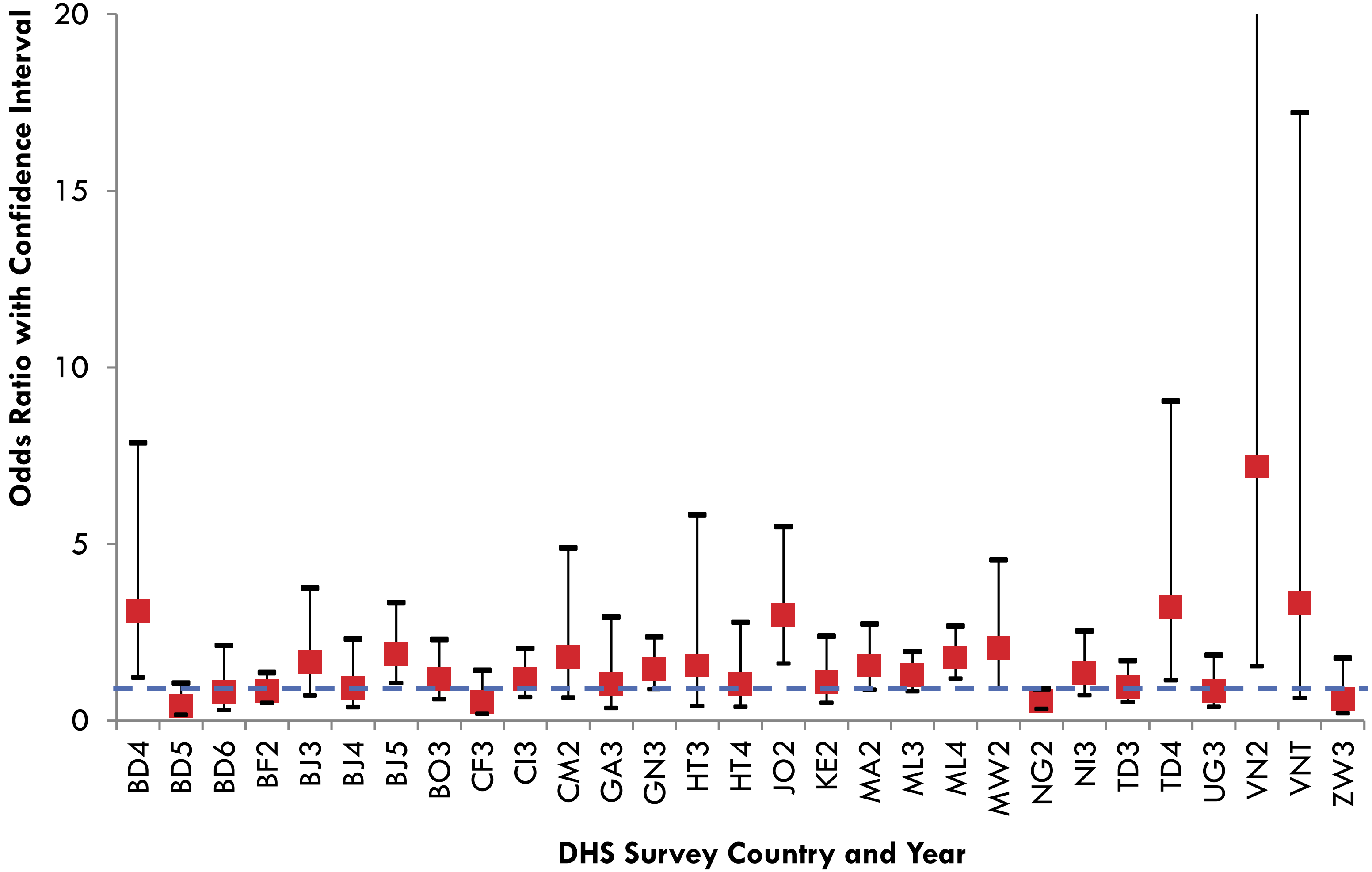
- Compared to living < 1 km from a facility, living > 10 km from a facility:
  - 17.9 percent higher odds of dying before 5th birthday
- Disaggregation suggests that the results driven by neonatal mortality
  - 26.6 percent higher odds of dying within the first 28 days

Distance not significantly associated with mortality in older age groups (post-neonatal infants and post-infant children)

Main Travel Distance Results



Neonatal Death by Survey



Conclusions

- People live relatively close to facilities
    - Literature is focused on the most remote areas (> 5 km or > 10 km), but such distances are rare
    - 50-60 percent of households are within 3 km
- Distance to facilities does not only matter when facilities are far, but also within relatively narrow radiuses
    - Suggests that relatively minor factors are likely to have substantial effects on health behaviors
- Reducing distance to facilities may increase health care utilization and, more importantly, improve neonatal survival

Estimation with Induced Measurement Error in Explanatory Variables: A Numerical Integration Approach

M. Karra and D. Canning

The Measurement Error Problem

- Measurement error in an explanatory variable in a regression yields biased (attenuated) and inconsistent estimates
- Typically, structure of measurement error is unknown
- Sometimes, however, measurement error is often added to data to protect respondent confidentiality
- The structure of this induced measurement error may be known

The Measurement Error Problem

- Examples include:
  - Coarsening of the variable into bands (age, income, location)
  - Building error into the data collection (randomized response)
  - Deliberately adding noise / scrambling data (geographic locations)

- Naïve regressions with perturbed data can seriously bias results
- Previous methods to adjust for the error (e.g. regression calibration) assume normality in the variable and in the error

The Measurement Error Problem

- Want to estimate:

$$y_i = \alpha + \beta g(x_i) + \gamma z_i + \varepsilon_i$$

- In the data, x_i not observed but we do get m_i, which is x_i measured with error
- Running the regression with m_i, i.e.

$$y_i = \alpha + \beta g(m_i) + \gamma z_i + \varepsilon_i$$

will yield biased estimates of β

Objective

- To develop a theory that allows for unbiased and consistent estimation of a linear regression where measurement error in the explanatory variable is known

Approach

- Calculate the expected value of the true explanatory variable, given mismeasured variable and error generating process
  - Integrate over all possible actual values of the true data, weighted by conditional probability of data values given the observed perturbed data
- Replace the perturbed variable with this expectation
- This approach is related to regression calibration
  - Regression calibration is a special case where the true variable and error are independent and normally distributed

Data Requirement

- Our approach typically will require an independent source of the underlying true distribution of data, $p(x)$
  - To link individuals to exposures at the zip code level when the data reports only at the state level, we need independent information on the population distribution in each zip code
- One possible exception: if the distribution of the perturbed data can be inverted (see Appendix for technical explanation)

Applications of the Method

- Special cases include:
    - Normally distributed additive error (regression calibration)
- Applications include:
    - Coarsened location variables (state-county-zip, etc.)
    - Continuous variables in intervals (income levels, age bands)
    - Randomized responses in data (throwing a die to tell the truth)
    - Perturbed spatial data (geoscrambling)

Application to Perturbed Spatial Data: A Simulation Exercise

Geoscrambling in the DHS

- In the Demographic and Health Surveys (DHS), GPS coordinates of surveyed household (HH) clusters are collected
- These coordinates are then scrambled using a random angle, random radius displacement algorithm
    - Urban HH clusters: displaced up to 2 km
    - Rural HH clusters: displaced up to 5 km, with every 100th cluster displaced up to 10 km

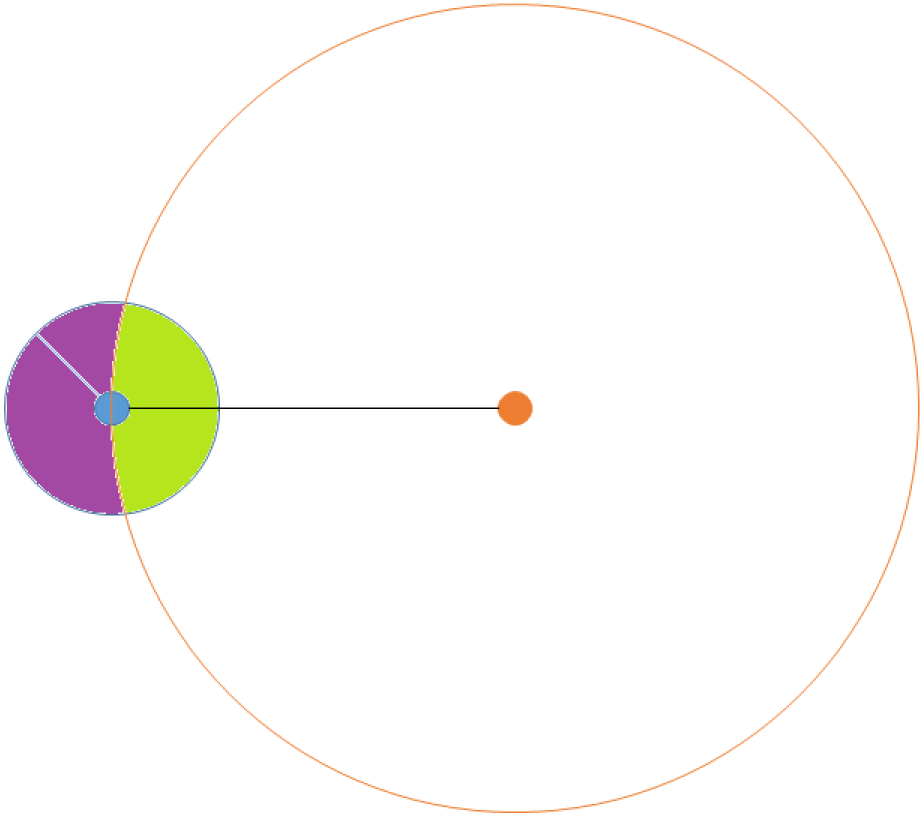
Geoscrambling in the DHS

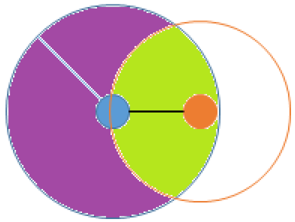
- A graphic example of having one facility (orange dot) and one HH cluster (blue dot)
- HH cluster is displaced by a distance at a random radius
- Calculating distance measures to this facility will be measured with error, and this error will bias estimates

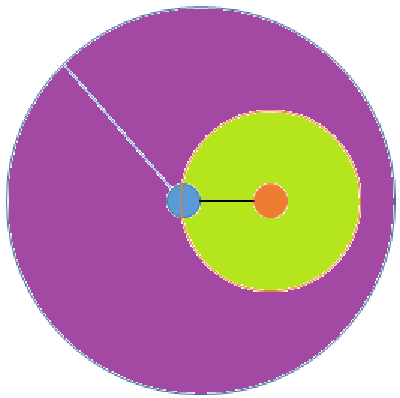
Example: One Facility, One Cluster

One Facility, One Cluster

- Start with simple example of having one facility (orange dot) and one cluster (blue dot)
- Blue dot is displaced by various distances







One Facility, One Cluster

- Measurement of distance more likely to be biased upwards
- Displaced distances are more likely to be larger than original distances

Two Facilities, One Cluster

Two Facilities, One Cluster

- Extend the example of one facility-one cluster that is displaced to two facilities-one cluster
- This implies that the cluster can potentially be mismeasured (distance is wrong) and mismatched (facility is wrong)

Simulation Setup

- Generate a 100 x 100 grid space
- Place 100 facilities uniformly across this grid at locations $r = \left(r_{z_1}, r_{z_2}\right)$ for $z_1, z_2 = 1, \dots, 100$
- Place 1,000 HH clusters uniformly across this grid at locations $x = (x_1, x_2)$. Cluster i is denoted $x_i = (x_{i1}, x_{i2})$
- Since the placement of clusters is uniform, we know that $p(x) = p(x_1, x_2)$ is uniform

Simulation Setup

- We want to run the regression of the association between distance from the cluster to the nearest facility, $g(x_i)$ on an outcome of interest, y_i
- In the equation $y_i = \alpha + \beta g(x_i) + \gamma z_i + \varepsilon_i$, the component $g(x_i)$ is the function that specifies the facility that is nearest to a household cluster, i.e.

$$g(x_i) = \min_{z_1, z_2} \sqrt{\left(x_{i1} - r_{z_1}\right)^2 + \left(x_{i2} - r_{z_2}\right)^2}$$

- We calculate the distance to the nearest facility $g(x_i)$ for each cluster x_i

Simulation Setup

- For simulation purposes, we generate the outcome of interest y_i in accordance to relationship:

$$y_i = 1 + 1 \cdot g(x_i) + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$

- Here, the true parameter values are $\alpha, \beta = 1$ and $\gamma = 0$
- To validate, we can estimate this equation

$$y_i = \alpha_x + \beta_x g(x_i) + \varepsilon_i$$

and show that $\widehat{\beta_x}$ is unbiased.

Simulation Setup

- We now assume that we are given displaced cluster coordinates $m = (m_1, m_2)$ instead of (x_1, x_2)
- The displacement of the cluster is given by:
  - Random angle uniformly selected between $[0,2\pi]$
  - Random distance uniformly selected between $[0,5]$
- We run the regression

$$y_i = \alpha_m + \beta_m g(m_i) + \varepsilon_i$$

to show the bias in the $\widehat{\beta_m}$ estimate

Simulation Setup

- Under these conditions, we know that the mechanism to induce the displacement error is:

$$p((m_1, m_2)|(x_1, x_2)) = \begin{cases} 0, & \sqrt{(m_1 - x_1)^2 + (m_2 - x_2)^2} > 5 \\ \dfrac{1}{5 \cdot 2\pi\sqrt{(m_1 - x_1)^2 + (m_2 - x_2)^2}}, & \sqrt{(m_1 - x_1)^2 + (m_2 - x_2)^2} \leq 5 \end{cases}$$

- We now have all of the components to do our simulation

Simulation Setup

- Run numerical integration over entire grid to get expectation
- Run the regression

$$y_i = \alpha_C + \beta_C E[g(x_i)|m_i] + \varepsilon_i$$

- Compare estimated $\widehat{\beta_C}$ with $\widehat{\beta_m}$ and true value of $\beta = 1$, and show that $\widehat{\beta_C}$ is unbiased

Simulation Steps

1. Generate fixed set of 100 facilities and 1,000 clusters
2. Calculate real minimum distances for each cluster

Iterate over following 4 steps:

3. Draw random error ε_i and generate outcome y_i
4. Run the true regression and get $\widehat{\beta_x}$ estimate (unbiased)
5. Perturb each cluster x_i to m_i, run naïve regression with m_i and get $\widehat{\beta_m}$ (biased)
6. Estimate expectation of the true distance by numerical integration, run adjusted regression, and get $\widehat{\beta_C}$ (unbiased)

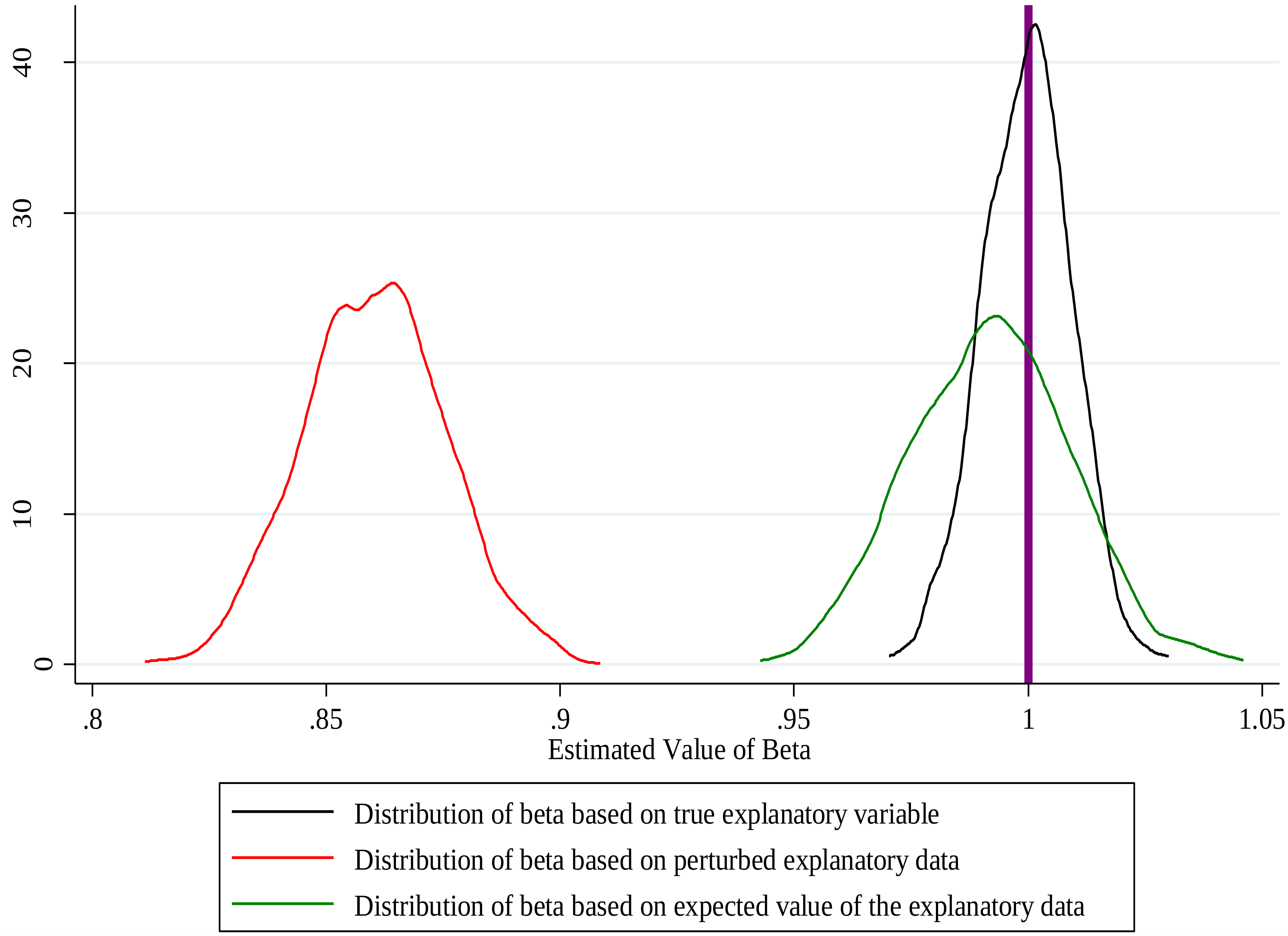
Iterate 1,000 times to get empirical distributions of $\widehat{\beta_x}$, $\widehat{\beta_m}$, $\widehat{\beta_C}$

Simulation Results

Empirical Distributions of $\widehat{\beta_x}$, $\widehat{\beta_m}$, $\widehat{\beta_C}$ under 1,000 iterations, mesh length $h = 1$ (100 x 100 mesh)

	Mean	SD	Minimum	Maximum
$\widehat{\boldsymbol{\beta}_x}$	0.9997	0.0094	0.9703	1.0301
$\widehat{\boldsymbol{\alpha}_x}$	1.0004	0.0587	0.8193	1.1965
$\widehat{\boldsymbol{\beta}_m}$	0.8604	0.0151	0.8112	0.9085
$\widehat{\boldsymbol{\alpha}_m}$	1.7238	0.0951	1.4458	2.0546
$\widehat{\boldsymbol{\beta}_c}$	0.9920	0.0170	0.9427	1.0460
$\widehat{\boldsymbol{\alpha}_c}$	1.0524	0.0945	0.7785	1.3634
N	**1,000**			

Simulation Results



Discussion and Conclusions

Conclusions

This Study:

- Proposes a general method for consistent inference when an independent variable is deliberately measured with error
- Shows how we can use numerical integration to calculate the expected value of the true variable
- Shows an example of how the method can be used through a simulation exercise

Future Work:

- Apply this method to real datasets (e.g. DHS)

Thank You!

For additional information: mvkarra@bu.edu

Appendices

Previous Work

- Association between distance and MCH service utilization: well-established
  - Literature review by Gabrysch and Campbell (2009)
    - Found overall negative relationship between distance and utilization
  - Subsequent studies in Zambia, Bangladesh, Malawi have confirmed this inverse relationship

Previous Work

- Association between distance and child mortality remains unclear
  - Literature review by Rutherford, Mulholland, and Hill (2010)
    - Inconclusive evidence to demonstrate an association
  - Some studies found positive effects (Vietnam, Burkina Faso, Ethiopia)
  - Some studies found no effects (Malawi, Zambia, Kenya)
  - Literature review by Okwaraji and Edmond (2012)
    - Selection bias towards significant results, cannot pool results well
    - **Issues around how distance is measured**

Measures of Distance

- Key measure for analysis: travel distance to the nearest facility
  - Generate four distance indicators
    - Distance to the nearest hospital
    - Distance to the nearest low-tiered clinic (HC3)
    - Distance to the nearest mid-level health center (HC2)
    - Distance to the nearest MCH center or PHC (HC1)
  - Take the minimum of the four distance indicators
  - For main analysis, divide into interval categories:
    - < 1 km (ref.), 1 km – 1.9 km, 2 km – 2.9 km, 3 km – 4.9 km, 5 km – 9.9 km, > 10 km
- Similar measure created for time to nearest facility
    - < 10 min (ref.), 10 – 19.9 min, 20 – 29.9 min, 30 – 59.9 min, > 60 min

Specification

$$\ln\left(\frac{Pr\left[Y_{ihcj} = 1 | \boldsymbol{X}_{ih}, \boldsymbol{Z}_C, \zeta_j\right]}{1 - Pr\left[Y_{ihcj} = 1 | \boldsymbol{X}_{ih}, \boldsymbol{Z}_C, \zeta_j\right]}\right) = \beta_0 + \beta_D D_c + \boldsymbol{X}_{ih}\gamma + \boldsymbol{Z}_C\delta + \zeta_j + \varepsilon_{ihcj}$$

- Y_{ih} is the binary dependent variable for birth i in household h in cluster c in survey j
- D_c is the travel distance to nearest facility variable for cluster c
- X_{ih} is the vector of individual-level and HH-level controls
- Z_C is the vector of cluster-level controls
- ζ_j are survey-level fixed effects

- Regression standard errors are clustered at the DHS cluster level

DHS Countries, Years

Country	Year	Country	Year
Bangladesh	2004	Haiti	1994-95
Bangladesh	2007	Haiti	2000
Bangladesh	2011	Jordan	1990
Benin	1996	Kenya	1993
Benin	2001	Malawi	1992
Benin	2006	Mali	1995-96
Bolivia	1994	Mali	2001
Burkina Faso	1993	Morocco	1992
Cameroon	1991	Niger	1998
CAR	1994-95	Nigeria	1990
Chad	1996-97	Uganda	1995
Chad	2004	Vietnam	1997
Cote d’Ivoire	1994	Vietnam	2002
Gabon	2000	Zimbabwe	1994
Guinea	1999		

Control Variables

- Birth- and HH-level controls:
  - Birth order, mother's education (categorical), HH wealth (quintiles), age of mother (categorical), place of residence (urban/rural)
  - For mortality regressions, hypothetical age of the child and the age of the child squared are added
- Cluster-level controls
  - Average wealth (quintiles), average schooling for mothers

Descriptive Statistics: Distances

BIRTHS **Minimum Travel Distance, categorical**	**Mean**	**No.**	**Urban Mean**	**Rural Mean**
Minimum distance to facility, < 1 km	0.279	35,387	0.534	0.177
Minimum distance to facility, 1 – 1.9 km	0.091	11,542	0.160	0.064
Minimum distance to facility, 2 – 2.9 km	0.152	19,279	0.158	0.150
Minimum distance to facility, 3 – 4.9 km	0.121	15,347	0.066	0.143
Minimum distance to facility, 5 – 9.9 km	0.153	19,406	0.050	0.194
Minimum distance to facility, > 10 km	0.204	25,874	0.031	0.272
N		**126,835**	**42,746**	**84,089**

Descriptive Statistics: Outcomes

Outcome Variables	**Mean**	**No.**
WHO Recommended ANC Visits (1 = yes)	0.394	49,186
Delivery in a health facility (1 = yes)	0.426	53,152
Child death	0.082	10,427
Neonatal death	0.030	3,806
Post-neonatal infant death	0.034	4,427
Post-infant child death	0.017	2,189
N	**126,835**	

Descriptive Statistics: Distances

CLUSTERS **Minimum Travel Distance, categorical**	**Mean**	**No.**	**Urban Mean**	**Rural Mean**
Minimum distance to facility, < 1 km	0.318	2,514	0.538	0.186
Minimum distance to facility, 1 – 1.9 km	0.111	869	0.169	0.074
Minimum distance to facility, 2 – 2.9 km	0.170	1,340	0.160	0.175
Minimum distance to facility, 3 – 4.9 km	0.116	915	0.058	0.150
Minimum distance to facility, 5 – 9.9 km	0.133	1,052	0.048	0.185
Minimum distance to facility, > 10 km	0.153	1,211	0.027	0.229
N		**7,901**	**3,346**	**4,555**

Descriptive Statistics: Covariates

Mother-Level Covariates	**Mean**	**SD**	**No.**
Wealth, quintiles	2.893	1.392	
Education, none (1 = yes)	0.532		66,323
Education, primary (1 = yes)	0.271		33,777
Education, secondary (1 = yes)	0.176		21,890
Education, higher (1 = yes)	0.022		2,727
Maternal age, years	28.214	7.041	
Marital status (1 = married)	0.865		107,875
Urban (1 = yes)	0.284		35,399
Cluster-Level Covariates			
Average wealth, quintiles	2.889	1.066	
Average education, highest level	0.682	0.616	
Distance to primary school, km	1.724	4.822	
N	**124,719**		

Descriptive Statistics: Covariates

Birth-Level Covariates	**Mean**	**SD**	**No.**
Birth order	3.876	2.651	
Multiple birth (1 = yes)	0.027		3,383
Child sex (1= female)	0.494		62,705
Time from birth to survey date, months	24.311	16.115	
N	**126,835**		

Main Travel Distance Results

	(1) Neonatal	(2) ANC Visits	(3) Delivery
Reference : < 1 km			
1 km – 1.9 km	1.077	0.834***	0.920
	(0.927 - 1.251)	(0.769 - 0.904)	(0.828 - 1.023)
2 km – 2.9 km	1.163**	0.825***	0.754***
	(1.020 - 1.327)	(0.767 - 0.887)	(0.681 - 0.835)
3 km – 4.9 km	1.250***	0.779***	0.691***
	(1.087 - 1.439)	(0.715 - 0.850)	(0.612 - 0.779)
5 km – 9.9 km	1.191**	0.713***	0.547***
	(1.042 - 1.363)	(0.652 - 0.779)	(0.483 - 0.620)
> 10 km	1.266***	0.612***	0.447***
	(1.108 - 1.445)	(0.559 - 0.671)	(0.394 - 0.508)
N	**125,167**	**124,719**	**124,719**

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Main Travel Time Results

	(1) Neonatal	(2) ANC Visits	(3) Delivery
Reference : < 10 min			
Time: 10 min – 19.9 min	1.074	0.872***	0.794***
	(0.952 - 1.212)	(0.814 - 0.933)	(0.722 - 0.873)
Time: 20 min – 29.9 min	1.157**	0.807***	0.732***
	(1.015 - 1.319)	(0.745 - 0.874)	(0.659 - 0.814)
Time: 30 min – 59.9 min	1.223***	0.748***	0.602***
	(1.078 - 1.389)	(0.692 - 0.809)	(0.538 - 0.674)
Time: > 60 min	1.256***	0.688***	0.477***
	(1.105 - 1.429)	(0.627 - 0.753)	(0.419 - 0.543)
N	**125,167**	**124,719**	**124,719**

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Check: In-Patient Facilities Only

	(1) ANC	(2) Delivery	(3) Neonatal	(4) Post-Neonatal	(5) Child 1-5
Reference : < 1 km					
1 km – 1.9 km	0.825***	0.904*	1.044	1.034	1.049
	(0.760 - 0.896)	(0.808 - 1.012)	(0.896 - 1.217)	(0.879 - 1.218)	(0.860 - 1.279)
2 km – 2.9 km	0.801***	0.711***	1.211***	1.113	1.094
	(0.742 - 0.865)	(0.638 - 0.793)	(1.054 - 1.392)	(0.964 - 1.285)	(0.913 - 1.310)
3 km – 4.9 km	0.736***	0.619***	1.314***	1.048	1.193*
	(0.673 - 0.805)	(0.546 - 0.701)	(1.134 - 1.523)	(0.901 - 1.220)	(0.988 - 1.441)
5 km – 9.9 km	0.699***	0.543***	1.175**	0.931	1.013
	(0.640 - 0.763)	(0.479 - 0.616)	(1.022 - 1.351)	(0.809 - 1.072)	(0.847 - 1.212)
> 10 km	0.587***	0.435***	1.295***	1.108	1.108
	(0.538 - 0.640)	(0.385 - 0.492)	(1.132 - 1.481)	(0.972 - 1.262)	(0.941 - 1.305)
N	**124,719**	**124,719**	**125,167**	**87,289**	**83,176**

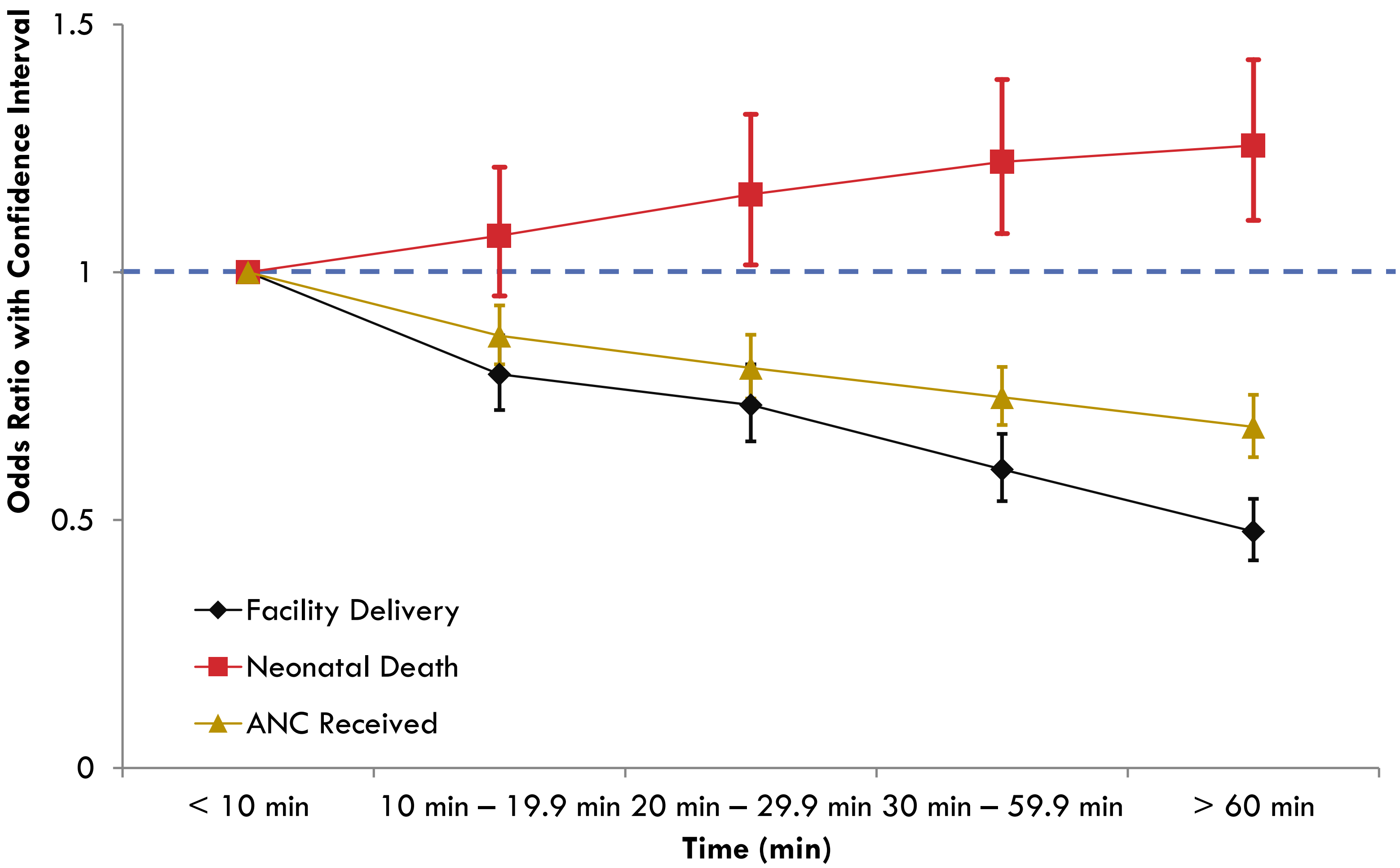
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Check: Control School Distance

	(1) ANC	(2) Delivery	(3) Neonatal	(4) Post-Neonatal	(5) Child 1-5
Reference : < 1 km					
1 km – 1.9 km	0.855***	0.856***	1.021	1.058	1.010
	(0.782 - 0.935)	(0.762 - 0.961)	(0.866 - 1.203)	(0.881 - 1.271)	(0.811 - 1.260)
2 km – 2.9 km	0.845***	0.707***	1.163**	1.079	1.150
	(0.776 - 0.920)	(0.630 - 0.794)	(1.000 - 1.353)	(0.911 - 1.278)	(0.938 - 1.409)
3 km – 4.9 km	0.774***	0.603***	1.273***	1.043	1.191
	(0.694 - 0.864)	(0.521 - 0.698)	(1.079 - 1.501)	(0.874 - 1.243)	(0.953 - 1.489)
5 km – 9.9 km	0.739***	0.529***	1.200**	0.993	1.034
	(0.661 - 0.826)	(0.456 - 0.614)	(1.029 - 1.399)	(0.846 - 1.166)	(0.844 - 1.266)
> 10 km	0.571***	0.416***	1.240***	1.091	1.108
	(0.506 - 0.644)	(0.356 - 0.485)	(1.062 - 1.447)	(0.942 - 1.265)	(0.914 - 1.343)
N	**95,108**	**95,108**	**95,300**	**66,071**	**62,972**

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Main Travel Time Results



Interpretation of Results

- Stronger association for in-facility delivery than for ANC coverage
  - Women can better plan ANC visits compared to when going to deliver
  - ANC is repeated, but delivery is one-shot
- Reasons for null, insignificant findings in older children
  - Seeking neonatal care not as easily anticipated as seeking care for older child, who is less susceptible
- Composition effects – which type of women use facilities?
  - Women who plan ahead vs. women who do not plan
  - But we see no differences for non-migrating mothers
- No qualitative differences between spatial and temporal distance

Approach

- Calculate the expected value of the true explanatory variable:

$$E[g(x_i)|m_i] = \int_X g(x)p(x|m_i)dx$$

- Set $g(x_i) = E[g(x_i)|m_i] + u_i$, where u_i is an error term with mean 0 and is independent of x_i and z_i
- Rewrite the estimating equation as:

$$y_i = \alpha + \beta E[g(x_i)|m_i] + \gamma z_i + \nu_i$$

where $\nu_i = \beta u_i + \varepsilon_i$

- This yields unbiased estimates of α, β, γ

Calculating $E[g(x_i)|m_i]$

- Calculate the expected value of the true explanatory variable using Bayes' Rule:

$$E[g(x_i)|m_i] = \int_X g(x)p(x|m_i)dx$$

$$= \int_X g(x)\frac{p(m_i|x)p(x)}{\int_X p(m_i|x)p(x)dx}dx$$

where $p(m_i|x)$ is the PDF of the error generation process and $p(x)$ is the PDF of the true values of the data, x

Calculating $E[g(x_i)|m_i]$

- In some cases, the integration needed to calculate the expectation is straightforward
- In some cases, there may not be an analytic solution
- Use numerical integration methods (sum over grid with interval $s = 0, \ldots, S$ and mesh h) to approximate the expectation

$$\sum_{s=0}^{S-1} g(x_s) \frac{p(m_i|x_s)p(x_s)h}{\sum_{s=0}^{S-1} p(m_i|x_s)p(x_s)h}$$

$$\approx \int_X g(x) \frac{p(m_i|x)p(x)}{\int_X p(m_i|x)p(x)dx} dx$$

A Possible Exception: Inversion

- Since we know the form of the measurement error, it may be possible to invert the distribution of perturbed data to generate the underlying distribution of the true data
  - Distributions of the true and perturbed variables are linked by a non-homogenous Fredholm integral equation of the first kind
  - Solution of this equation is well-studied
- But the inverse problem is generally not well posed
  - Cannot guarantee the existence or uniqueness of a solution
  - So then we require data on the underlying distribution