Yale SCHOOL *of* MANAGEMENT

YALE UNIVERSITY
School of Public Health

YALE UNIVERSITY
School of Engineering and
Applied Science

Estimating the Size of Hidden Populations

Edward H. Kaplan

How Many Homeless?

- Three cities conducted systematic sweeps of the homeless population when people were out on the streets
- Homeless shelter census was also known in each city
- For every 100 persons found in homeless shelters, there were:
  - 129 persons on the street in Boston
  - 273 persons on the street in Phoenix
  - 130 on the street in Pittsburgh
- National estimate for total number in homeless shelters was 69,000. Can you estimate total size of the homeless population?

Ratio Estimate for Number of Homeless

- Idea: let p = Pr{Shelter | Homeless}, x = # homeless in shelter
- Estimate $N = x / p$
- Told that nationwide x = 69,000
- From surveys estimate that p falls between 0.27 and 0.44
- So estimate that N falls between 69K / 0.44 and 69K / 0.27 or between 157K and 256K
- If add in confidence interval variability, widen to 155K – 258K

The New York Times

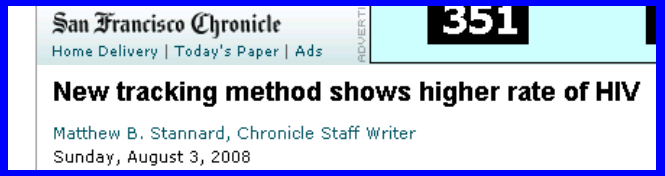
Health

H.I.V. Study Finds Rate 40% Higher Than Estimated

By LAWRENCE K. ALTMAN
Published: August 3, 2008

SIGN IN TO E-MAIL
OR SAVE THIS
PRINT
SINGLE PAGE
REPRINTS
SHARE

MEXICO CITY — The United States has significantly underreported the number of new H.I.V. infections occurring nationally each year, with a study released here on Saturday showing that the annual infection rate is 40 percent higher than previously estimated.



San Francisco Chronicle
Home Delivery | Today's Paper | Ads

351

New tracking method shows higher rate of HIV

Matthew B. Stannard, Chronicle Staff Writer
Sunday, August 3, 2008

THE WALL STREET JOURNAL. Digital Network | WSJ.com | MarketWatch | BARRON'S | D | All Things Digital. | More | Enter Symbol(s)

THE WALL STREET JOURNAL.

The Online Journal GET 2 WEEKS FREE ← SUBSCRIBE NOW

The Print Journal GET 2 WEEKS FREE SUBSCRIBE NOW →

HIV Infections Were Undercounted In U.S. for Nearly 10 Years, CDC Says Wall Street Journal

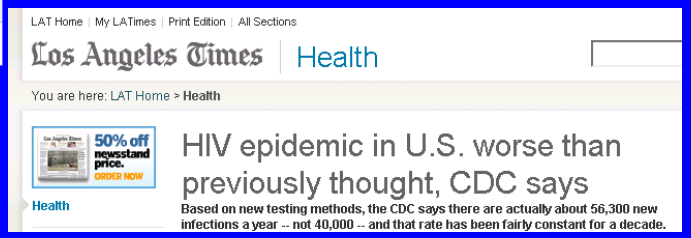
washingtonpost.com Sign In | Register Now

The Washington Post
Print Edition | Subscribe | PostPoints

AIDS Infection Rate in U.S. Higher Than Previously Estimated

By David Brown
Washington Post Staff Writer
Saturday, August 2, 2008; 2:24 PM

TOOLBOX
Resize Print E-mail



LAT Home | My LATimes | Print Edition | All Sections

Los Angeles Times | Health

You are here: LAT Home > Health

50% off newsstand price. ORDER NOW

Health

HIV epidemic in U.S. worse than previously thought, CDC says

Based on new testing methods, the CDC says there are actually about 56,300 new infections a year -- not 40,000 -- and that rate has been fairly constant for a decade.

How Many New HIV Infections Occur in the US Each Year?

- Want to know HIV incidence (stratified by important covariates) for monitoring, evaluation of prevention programs, and resource allocation
- For many years, CDC suggested 40,000/yr
- Where did the 40,000 come from?
  - Back-of-the-envelope calculation

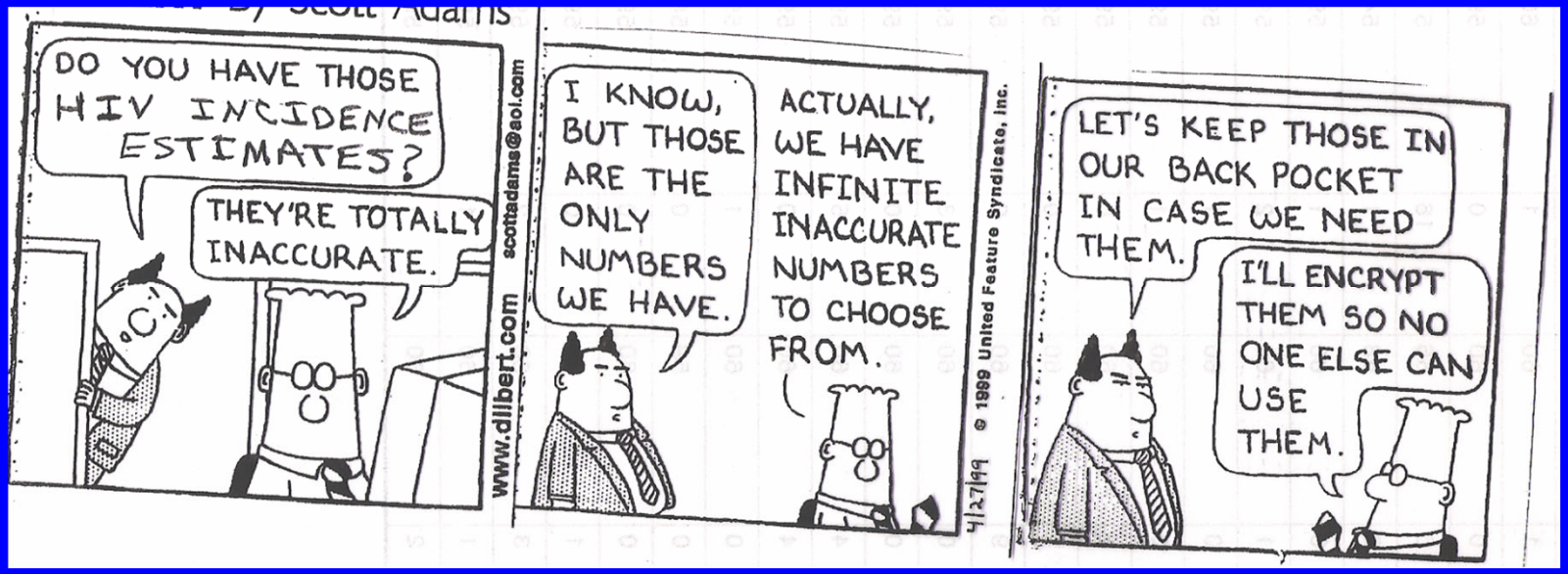
Estimated HIV incidence

1. Based on women

AIDS-OI incidence, mid-1990s, ~10,000/year
stable seroprevalence among child-bearing women
30% of incidence in women
total incidence ~ 33,000/year

2. Based on MSM

60.5 million men ages 15-44, 1997 (70.6, ages 15-49)
2% active MSM — 1.2 million MSM
~0.4 million with HIV — 0.8 million at risk
2% incidence/year — 16,000 infections/year
60% of infections in MSM — 27,000/year in men
70% of all infxs in men — 38,000 infections/year


DO YOU HAVE THOSE HIV INCIDENCE ESTIMATES?
THEY'RE TOTALLY INACCURATE.
I KNOW, BUT THOSE ARE THE ONLY NUMBERS WE HAVE.
ACTUALLY, WE HAVE INFINITE INACCURATE NUMBERS TO CHOOSE FROM.
LET'S KEEP THOSE IN OUR BACK POCKET IN CASE WE NEED THEM.
I'LL ENCRYPT THEM SO NO ONE ELSE CAN USE THEM.
www.dilbert.com scottadams@aol.com
© 1999 United Feature Syndicate, Inc.
4/27/99

Using The BED Assay

ESTIMATING HIV INCIDENCE IN THE UNITED STATES

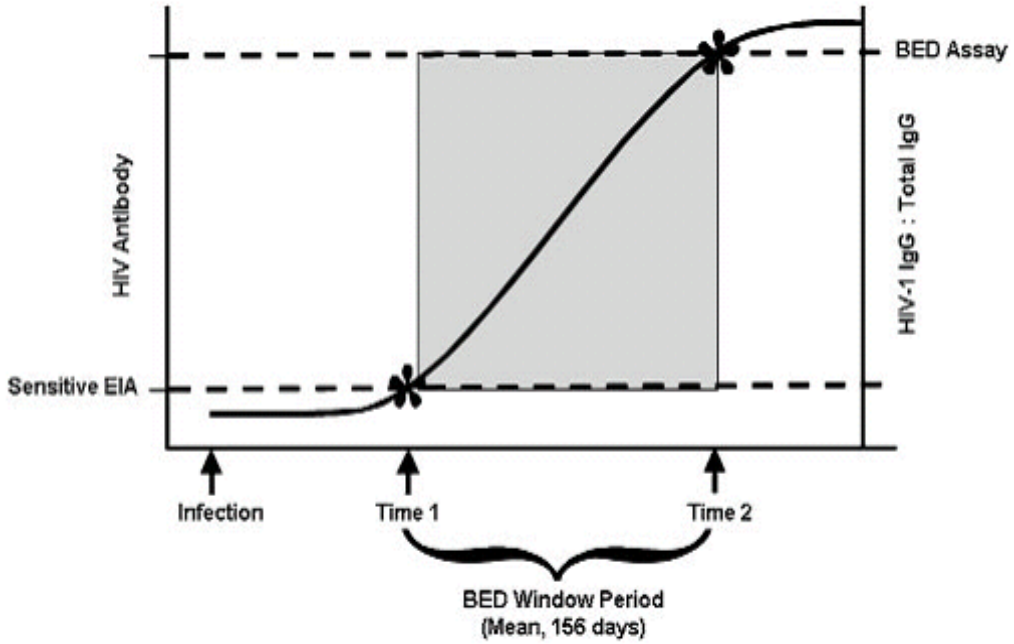
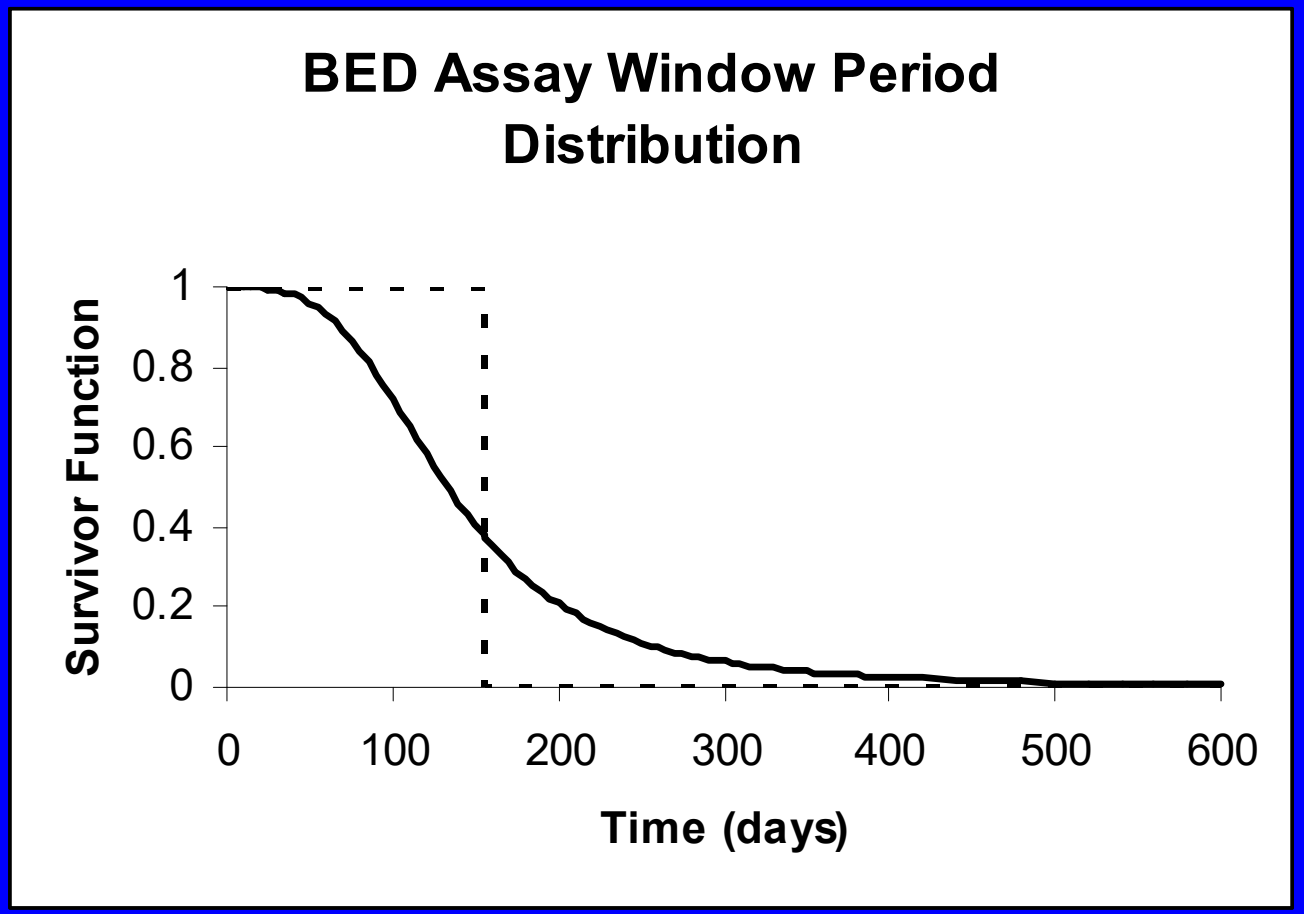


Figure 1. Schematic diagram of serologic testing algorithm for recent HIV seroconversion (STARHS) using the BED assay to determine those recently infected.

BED Assay Window Period Distribution



A Real Simple Model

- Let:
  - N = # new infections per year
  - p_1 = Pr{newly infected tested ≤ 1 year after infection}
  - p_2 = Pr{HIV$^+$ sample receives BED test}
  - p_3 = Pr{BED test reports "recent"| tested ≤ 1 year after infection}
  - R = # "recent" BED test results observed
- Then:

$$E(R) = E(N)\, p_1\, p_2\, p_3$$

so estimate

$$E(N) = E(R) / (p_1\, p_2\, p_3)$$

Example

Data Missing Completely At Random

Data/Parameter Estimate	Repeat Testers	New Testers
Observed HIV^+ Diagnoses	7604	4463
BED Recent Tests	908	298
p_1	0.617	0.240
p_2	0.398	0.426
p_3	0.427	0.427
Estimated Incidence	8660 (7440–10090)	6830 (5660–8210)

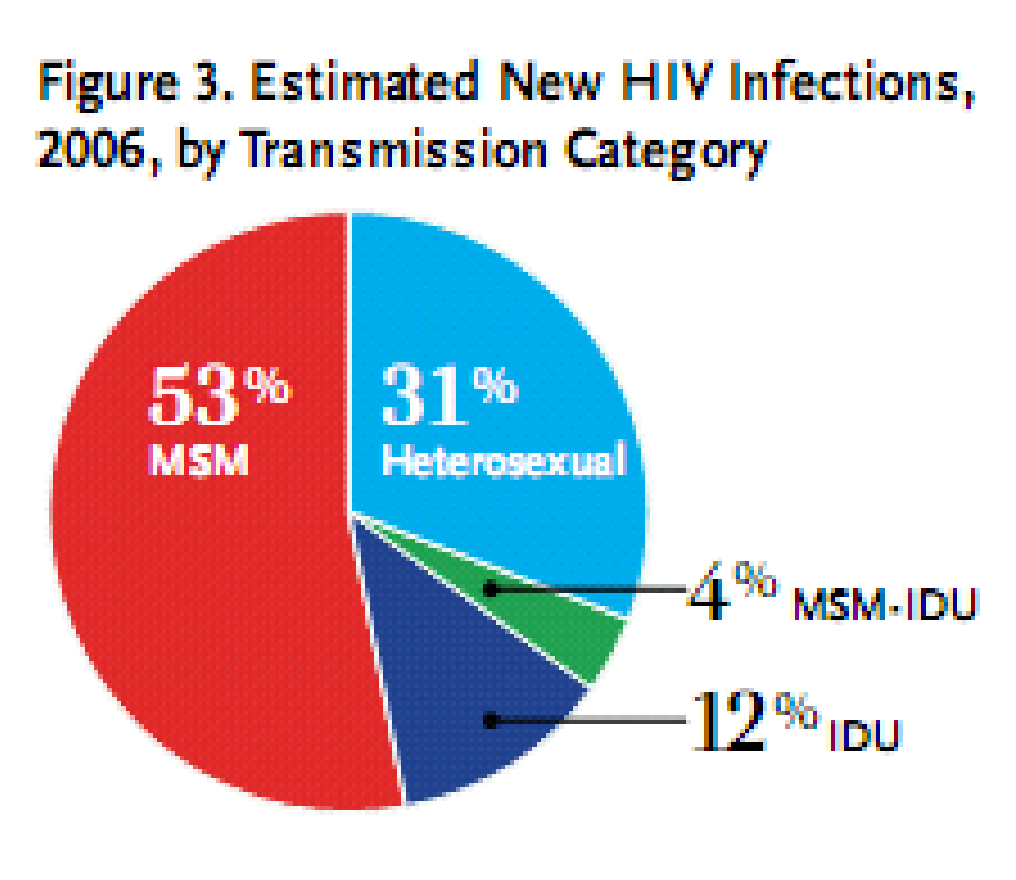
Total Estimated Incidence Among Observed Cases = 8660 + 6830 = 15,490
Test History Available From 12,067 / 33,802 Diagnoses or 35.7%

Raises Incidence Estimate Within 22 States to 15,490/0.357 = 43,400 (34,100–55,200)

Implies National Estimate of 43,400 / 0.73 = 59,500 (46,700–75,600)

- Recall Holmberg (1996) reported 50% of new infections among IDUs, 26% among MSMs, 24% among HETs
- Our results:

2006 AIDS diagnoses:
MSM: 43%
HET: 32%
IDU: 18%
MSM-IDU: 5%



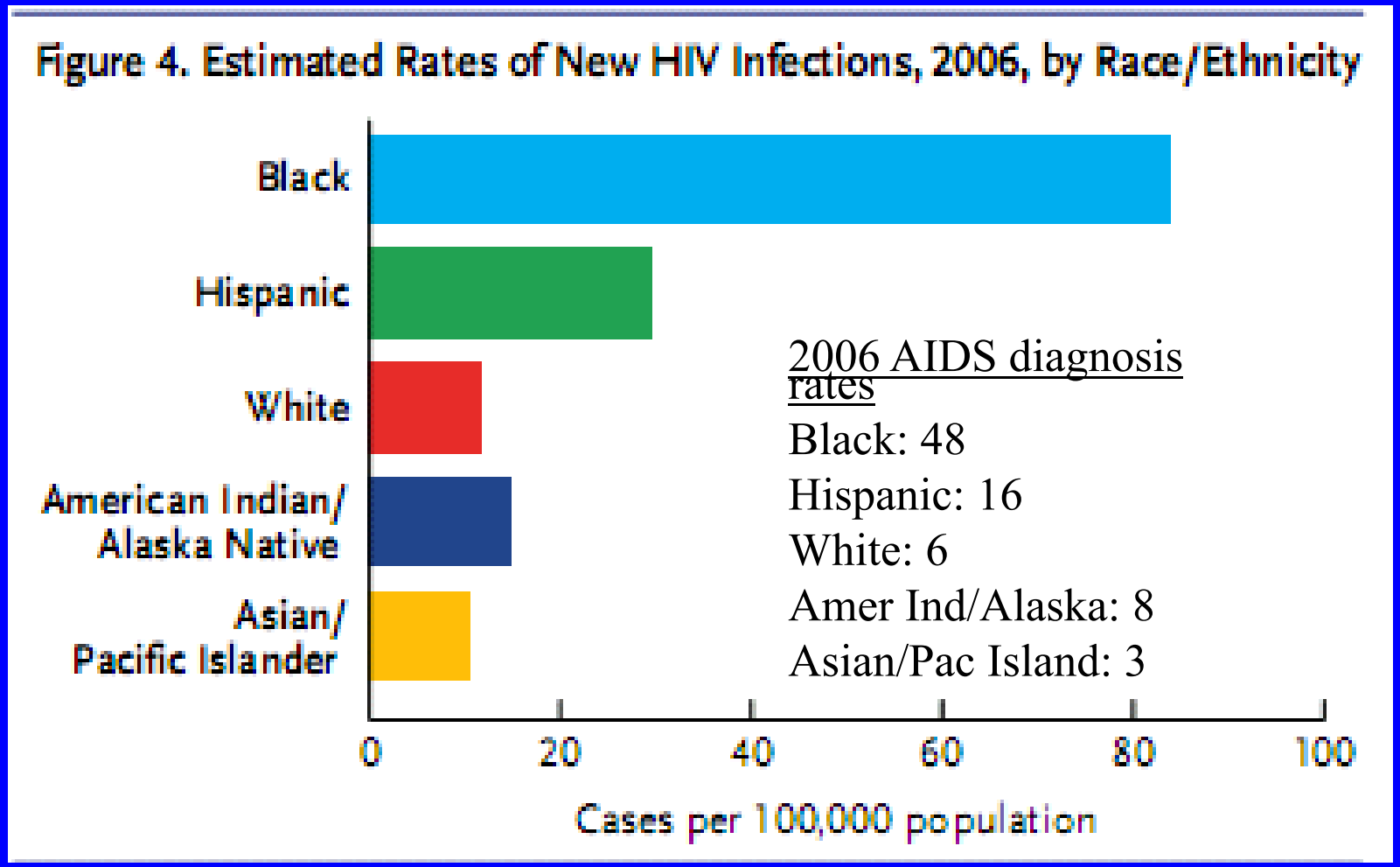


Figure 4. Estimated Rates of New HIV Infections, 2006, by Race/Ethnicity
Black
Hispanic
White
American Indian/
Alaska Native
Asian/
Pacific Islander
0
20
40
60
80
100
Cases per 100,000 population
2006 AIDS diagnosis rates
Black: 48
Hispanic: 16
White: 6
Amer Ind/Alaska: 8
Asian/Pac Island: 3

How Many Drug Injectors in New Haven?

- Suppose I = annual # new HIV infections among drug injectors
  - (could estimate via backcalculation from AIDS data)
- Suppose r = annual rate of new HIV infections per injector
  - (could estimate via epidemic model – will discuss December 8 in class)
- Then I / r = # infections/yr / # infections/drug injector/yr
  = # drug injectors!!

How Many Drug Injectors Are There In New Haven?

(EH Kaplan and D Soloshatz, *Math Comput Modeling* 17:109-115, 1993)

Table 1. Estimating the number of drug injectors in New Haven.

	Weibull	Erlang
$\hat{I}(=\hat{\beta}_1)$ (aggregate annual HIV incidence rate among drug injectors)	150.26	141.45
$\hat{\sigma}_{\hat{I}}$ (estimated standard error of $\hat{I}$)	4.41	4.13
$\hat{R}$ (per drug injector annual HIV incidence rate)	0.064	0.064
$\hat{\sigma}_{\hat{R}}$ (estimated standard error of $\hat{R}$)	0.00395	0.00395
$\hat{N}$ (point estimate of the number of drug injectors)	2,350	2,210
$\hat{\sigma}_{\hat{N}}$ (estimated standard error of $\hat{N}$)	160.45	150.90
$\hat{N} - 1.96\,\hat{\sigma}_{\hat{N}}$ (lower 95% confidence limit for $\hat{N}$)	2,030	1,190
$\hat{N} + 1.96\,\hat{\sigma}_{\hat{N}}$ (upper 95% confidence limit for $\hat{N}$)	2,660	2,510

Estimating prevalence of problem drug use at national level in countries of the European Union and Norway

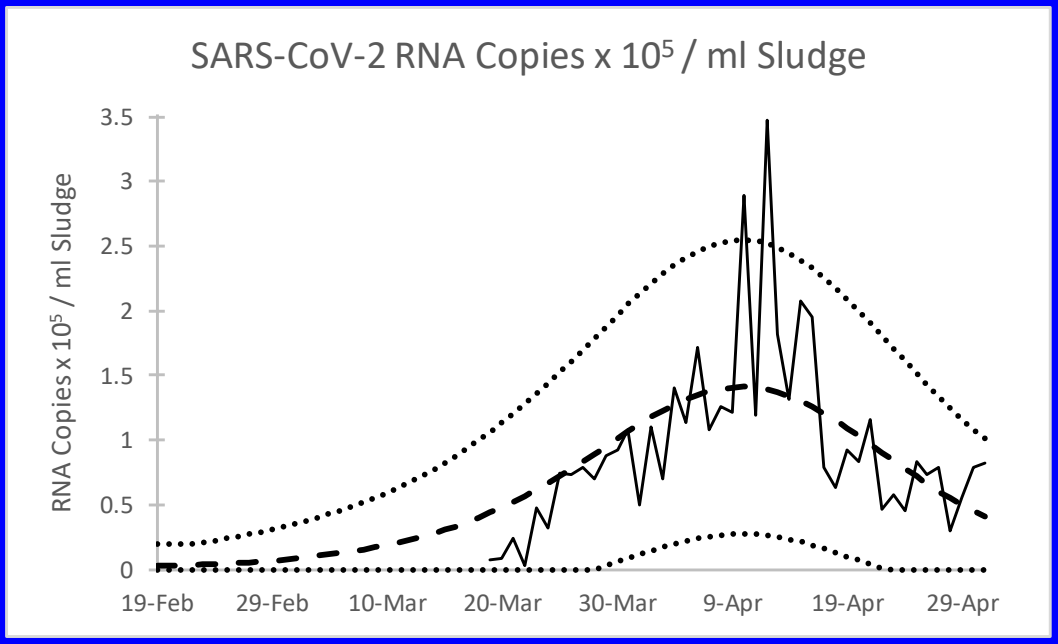
Ludwig Kraus[1], Rita Augustin[1], Martin Frischer[2], Petra Kümmler[1], Alfred Uhl[3] & Lucas Wiessing[4]

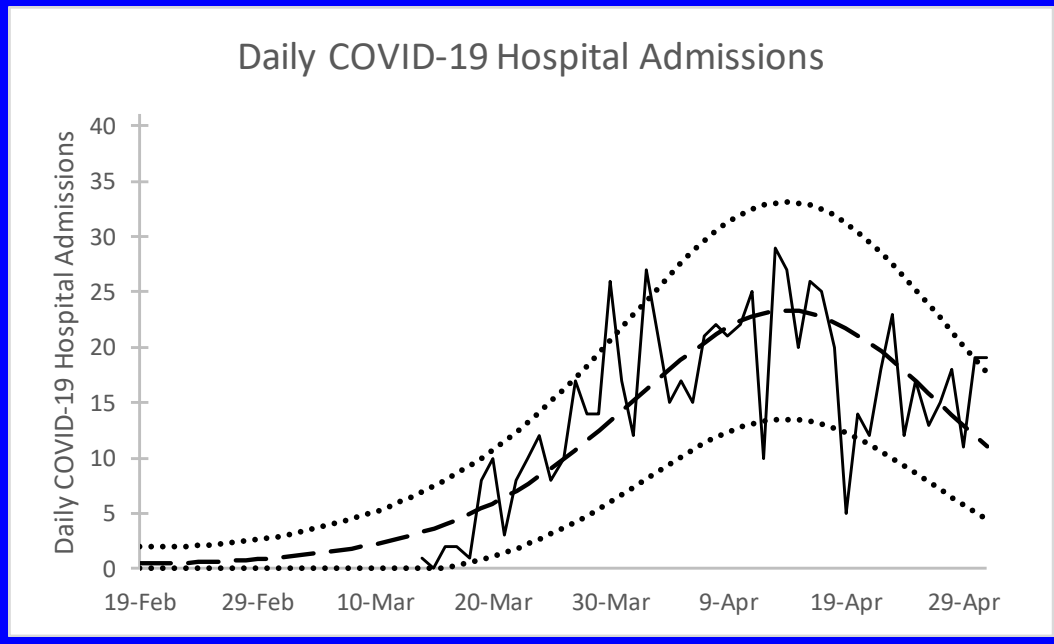
Addiction, 98, 471–485 2003

Table 1 National prevalence estimates of problem opiate use according to method in the EU and Norway (absolute numbers).

Country	*Treatment multiplier*	*Police multiplier*	*Mortality multiplier*	*HIV multiplier*
Austria 1995* [37]; 2000**			12 000–23 000 (3)[d]**	
Belgium 1995 [60]				20 000 (10 300–46 300)[f]
Denmark 1996			15 400 (3)[d]	
Finland 1999 [61]			7 000–14 000 (3)[d]	
France 1995*; 1999**	180 000[a]**	150 000 (1)[a]**		141 000–177 000[f]*
Germany 2000	166 000–198 000	153 000–190 000 (2)	127 000–169 000 (3)[d]	
Ireland 1995*; 1996** [58]			4 700 (4)[e]*	
			7 900 (4) (5)[e]*	
Italy 1996* [28]; 1999**	277 000**	281 000 (2)**		214 000–272 000[f]*
Luxembourg 1999 [25]		2 620 (1)	1 330–1400 (3)[d]	1 780[f]
		2 210–2480 (2)	2 090–2150 (3) (5)[d]	
the Netherlands 1998*; 1999**	26 000–30 300**	25 800–34 300 (1)[c]*		
Norway 2000			10 500–14 000 (4)[e]	
Portugal 1999*; 2000**	41 700–48 700[c]**	49 900–56 200 (1)[c]*	18 500–36 900 (3)[d]*	22 700–33 600[f]*
Spain 1998	177 800		84 000 (4)[e]	
Sweden 1998 [62]				
UK 1996 [59]	243 800		161 100 (3)[d]	161 200[f]

Empirical Results





- Estimated R_0 = 2.38 (std error 0.10); $s(0)$ = 0.984 (std error 0.003)

Bifurcation Model for 1st Covid Wave

- Interpret the epidemic model as applying only to *exposed* population; stay-at-home assumed safe
- Of the New Haven area's 200K population, how many locked down, how many remained exposed, and of those how many got infected?
- Use the aligning indicators model to estimate answers

Back-of-the-Envelope Bifurcation

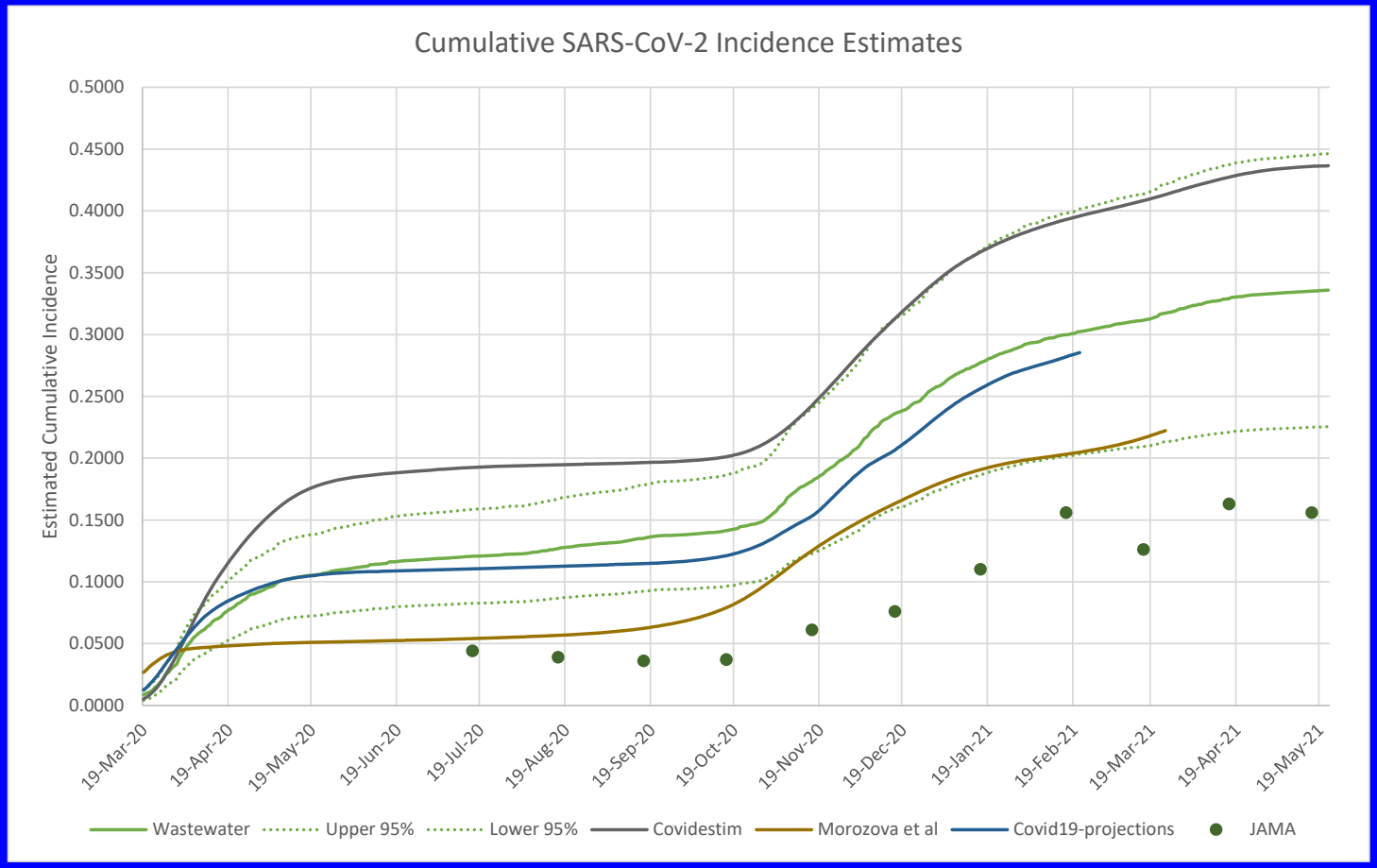
- Let ϕ = Pr{Infected during outbreak | exposed}
- Total hospitalizations $\mathbf{H} = k_H \phi$ (see eqs (16) and (21) in paper)
- Let **N, C** = total number exposed, total diagnosed COVID-19 cases
- Write $\mathbf{H} = \mathbf{N}\phi \times \frac{\mathbf{C}}{\mathbf{N}\phi} \times \frac{\mathbf{H}}{\mathbf{C}}$ (verify that RHS indeed = **H**)
- Equate the two expressions for **H** and simplify to obtain

$\mathbf{N} = \frac{k_H}{\frac{\mathbf{C}}{\mathbf{N}\phi} \times \frac{\mathbf{H}}{\mathbf{C}}}$ and note that $\frac{\mathbf{C}}{\mathbf{N}\phi}$ = cases per infection while $\frac{\mathbf{H}}{\mathbf{C}}$ = fraction of cases admitted to the hospital

Back-of-the-Envelope Bifurcation

- Estimated k_H = 1006.6 from maximum likelihood (s.e. 56.8)
- CDC estimated infections per case for Connecticut as 6 (s.e. 1.8) from a seroprevalence study March 23 – May12; this is $\frac{N\phi}{C}$
- Observed 734 hospital cases and 2,674 diagnosed cases for study population; this estimates $\frac{H}{C}$
- Taken together, estimate **N** = 22K (95% CI 18.6 K to 26K)
- Bifurcation model suggests about 11% of total population exposed
- Also suggests 18.6K infected (or 9.3% of total population)

Cumulative Infections as of May 1, 2020: Bifurcation Model: 9.3% (Morozova et al: 5%; covidestim: 14.8%; covid19-projections.com: 9.6%)



Typos in a Book

- Two proofreaders have been hired to independently read a book manuscript in search of typos
- The first (second) proofreader will catch any true typo with probability p (q)
- The first proofreader catches 50 typos, the second catches 40, and the same 10 typos were detected by both proofreaders
- Can you estimate the total number of typos in the book?

Typos in a Book

- Let the total number of typos = n
- E[Typos found by first proofreader] = np (observed 50)
- E[Typos found by second proofreader] = nq (observed 40)
- E[Typos found by both] = npq (observed 10)
- Here comes the cool part: np x nq / (npq) = n (!!!)
- Plug in the data: estimate n = 50 x 40 / 10 = 200 typos (!!)
- This is an example of what is called *capture recapture*

American Journal of Epidemiology
Vol. 141, No. 6
Printed in U.S.A.

Use of Capture-Recapture to Estimate the Prevalence of Opiate Addiction in Barcelona, Spain, 1989

Antònia Domingo-Salvany,[1] Richard L. Hartnoll,[1] Andrew Maguire,[1] J. M. Suelves,[2] and J. M. Antó[1]

TABLE 1. Distribution of unique individuals by occurrence of emergency room episodes in different combinations of trimesters (substudy one, *n* = 2,075), Barcelona, Spain, 1989

	TR1* yes, TR2* yes	TR1 yes, TR2 no	TR1 no, TR2 yes	TR1 no, TR2 no
TR3* yes, TR4* yes	29	35	35	96
TR3 yes, TR4 no	48	58	80	400
TR3 no, TR4 yes	25	77	50	376
TR3 no, TR4 no	97	357	312	†

* TR1, first trimester, TR2, second trimester; TR3, third trimester; TR4, fourth trimester.
† To be estimated.

199 527 477

		TRI2? Yes	No
TR1?	Yes	199	527
	No	477	
	nhat	2466.211	
	var	15655	
	st dev	125.1199	

Var=nhat*b*c/a^2

TABLE 3. Estimated number of addicts (aged 15–44 years) on the basis of different combinations of two trimesters of emergency room episodes, Barcelona, Spain, 1989

Trimesters	Estimated no. of addicts*	SE*
TR1† and TR2†	2,466	125.1
TR2 and TR3†	2,750	145.8
TR3 and TR4†	2,896	153.5
TR1 and TR3	3,335	198.0
TR2 and TR4	3,516	238.9
TR1 and TR4	3,162	189.2

* Number and standard error (SE) calculated through the following formulae: $N = (a + b)(a + c)/a$, $\mathrm{Var}(N) = (a + b)(a + c)bc/a^3$; where *a* is the overlap cell and *b* and *c* are unique individuals in each sample (13). The numbers of individuals in each cell are derived from table 1.
† TR1, first trimester, TR2, second trimester; TR3, third trimester; TR4, fourth trimester.

Number in first trimester = 199 + 527 = 726; number in second trimester = 199 + 477 = 676; number in both = 199.
So, capture recapture estimate = 726 x 676 / 199 = 2,466