What is "Big Data"—and how is it impacting the scientific method?

Fernando P. Chaves, MD
Global Head of Medical Affairs
Ortho Clinical Diagnostics

Acknowledgment

Many slides in this presentation are a courtesy of Dr. Warren La Morte, professor of Biostatistics.
Boston University School of Medicine

Evolution of Medical Information

Descriptive Studies

Stage	Study Types
1. Description; Hypothesis Generation Characteristics of disease & subjects	Case Report Case-Series Cross-Sectional Correlational

Analytical Studies

2. Hypothesis testing What factors are associated with disease?	Case-Control Cohort Study
3. Evaluation of Intervention Efficacy of treatment or prevention?	Clinical Trial (Intervention St.)

Test for association by *comparing groups*.

Scurvy: James Lind

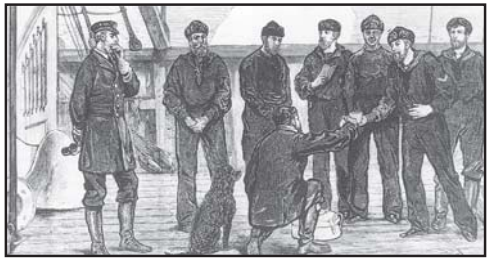
- Vitamin C deficiency: abnormal collagen synthesis
- Capillaries fragile; rupture easily
- Purple blotches on skin
- Hemorrhage into muscles, joints, digestive organs
- Bleeding, swollen gums; anemia; emotional changes
- Jaundice, edema, fever, convulsions, & heart failure
- Can result in death

- Lord Anson's global expedition 1740-1744
  - ➢ 1,955 men set out
  - ➢ 320 die from fevers, infections
  - ➢ Almost 1,000 die from scurvy

◆ **James Lind**, Scottish naval surgeon, suspects citrus fruits can prevent it.

1754: The First Controlled Trial?

12 Sailors with scurvy
- 2 got sea water
- 2 got a quart of cider a day
- 2 got vinegar
- 2 got a mixture containing garlic, radish, Peru balsam, and myrrh
- 2 got oil of vitriol
- 2 got **lemons** and **oranges**



Results:

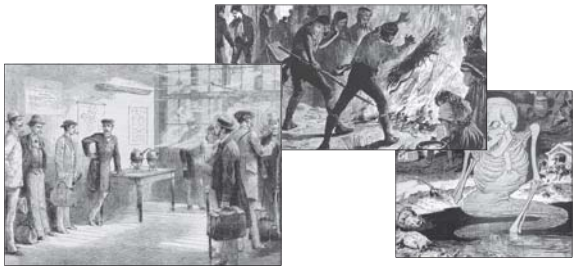
- **6 days later the 2 who received lemons & oranges were fit for duty.**
- **The others remained sick.**
- **Mechanism?**

Epilog:

1795:
- ➢ **The British navy ordered a general issue of lemon juice.**
- ➢ **Within 2 years scurvy disappears from British fleet.**
- ➢ **Later, they switch to limes, giving rise to term "limey."**

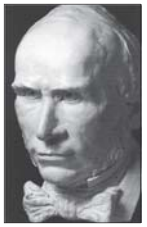
Beliefs About Transmission of Cholera

- Person-to-person
- "Miasmas": mysterious vapors from cemeteries, swamps, cesspools



Cholera: John Snow

- **London physician to poor**
- **Studied cholera outbreaks in 1830s**
- **Since symptoms were GI, not pulmonary, he reasoned it might be transmitted by water or food.**



- ➢ **Cholera is again a problem in London 1848-1854**
- ➢ **Snow conducts a landmark series of studies which:**
  - **a) test his hypothesis about mode of transmission**
  - **b) lead to an intervention**

Water Supply in London: 1800s

- Some families carry water from scattered community pumps
- Companies pump water from Thames River via pipes
  - ➢ Snow compares rates of cholera

Water Company	# Houses	# Cholera Deaths	Death/10,000 Houses
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

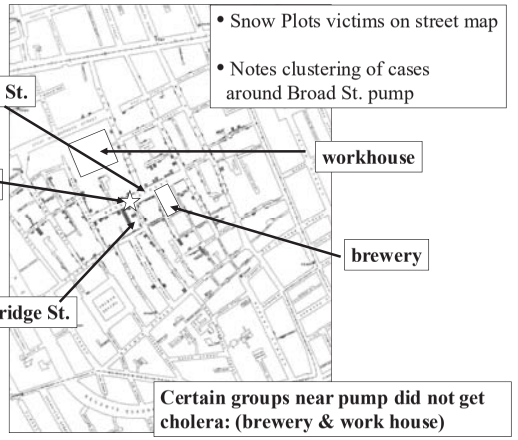
Downstream vs. Upstream

Company	Population	Cholera deaths	Deaths per 100,000
Southwark & Vauxhall (downstream)	167,654	192	114
Lambeth (upstream)	14,832	0	0

A cartoon from the 1800s lampooning the many useless precautions people were urged to take to avoid getting cholera.

Aug. 31, 1854: Cholera Outbreak Near Broad St.

- Cholera outbreak in Broad St. area of London
- By Sept. 4 there are 203 cholera deaths within 250 yards of Broad St. & Cambridge St.
- Residents begin to flee; 75% gone after 6 days



- Snow Plots victims on street map
- Notes clustering of cases around Broad St. pump

Certain groups near pump did not get cholera: (brewery & work house)

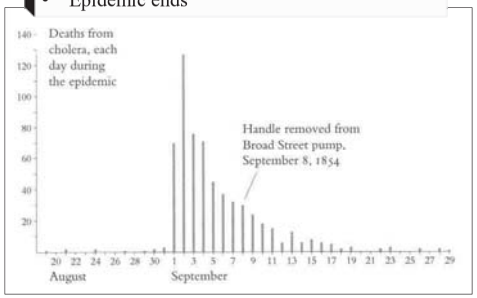
Snow Determines Water Source of Victims

The vast majority used the Broad St. pump.

Persons near Broad St.	
a) Broad St. pump	61
b) Other pump	6
c) Unknown source	6

Sept. 7, 1854

- Snow presents findings to authorities
- Handle removed from pump the next morning
- Epidemic ends



Significance of John Snow's Endeavors

- Based on observation & reason he proposed a new ***hypothesis*** for how cholera was transmitted.
- He tested the hypothesis by collecting data systematically & ***comparing groups*** of people.
- Established an ***association*** between drinking water from Broad St. pump and getting cholera.
- Argued for an ***intervention*** that prevented more cases.

Evolution of Medical Information

Descriptive Studies

Stage	Study Types
1. Description; Hypothesis Generation Characteristics of disease & subjects	Case Report Case-Series Cross-Sectional Correlational

Analytical Studies

Stage	Study Types
2. Hypothesis testing What factors are associated with disease?	Case-Control Cohort Study
3. Evaluation of Intervention Efficacy of treatment or prevention?	Clinical Trial (Intervention St.)

Test for association by *comparing groups*.

Descriptive Studies

Searching for clues.

- Case reports
- Case series
- Cross sectional surveys
- Correlational studies

Describe characteristics & behaviors.
Who gets the disease: age, gender, race, geography, habits, diet, drugs used?

- Alert medical community to new health problems
- Generate hypotheses
- First step for subsequent analytic studies

But, can't establish validity of an association.

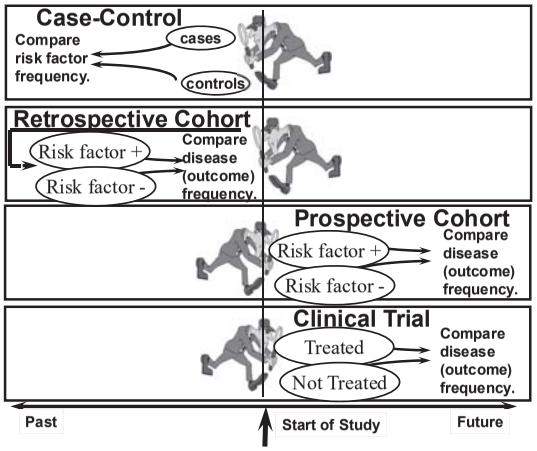
Ignaz Semmelweis, 1847

Clues:
Low rates in home deliveries
a similarity
Low rates in ward 1 (midwives)
a difference
High rates in ward 2 (med students who did dissections)
a similarity
Similar syndrome in a pathologist

Vienna General Hospital "child bed fever"

KEY TAKE HOME MESSAGE

Descriptive studies allow you to
FORMULATE A HYPOTHESIS
Which can be tested in analytical studies.



Evolution of Medical Information

Descriptive Studies

Stage — **Study Types**

1. **Description; Hypothesis Generation** — Characteristics of disease & subjects → Case Report, Case-Series, Cross-Sectional, Correlational

Analytical Studies

2. **Hypothesis testing** — What factors are associated with disease? → Case-Control, Cohort Study
3. **Evaluation of Intervention** — Efficacy of treatment or prevention? → Clinical Trial (Intervention St.)

Test for association by *comparing groups*.

Why the scientific method has worked well until today?

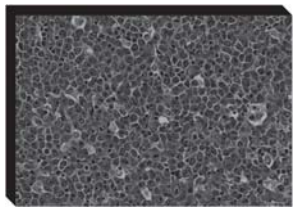
- Data collection and analysis was until recently the most time consuming part of research.
- Investigators simply couldn't afford to go "fishing in the dark" for whatever differences they could find.
- The use of previous observations to generate hypothesis increased the yield of success when expensive and time consuming studies were performed.

So what is different about Big Data

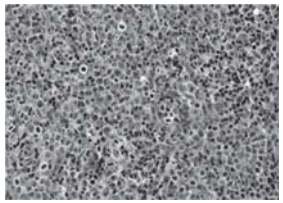
- Modern technology – made possible cheap and fast collection and analysis of vast amounts of data.
- Investigators can now afford to go "fishing in the dark".
- Data and scientific studies are used to **generate** the hypothesis.
- Differences in populations studies are then identified and used to generate biological / theoretical explanations for these differences.

Diffuse Large B-cell Lymphoma

Good response to therapy, good prognosis



Bad response to therapy, bad prognosis



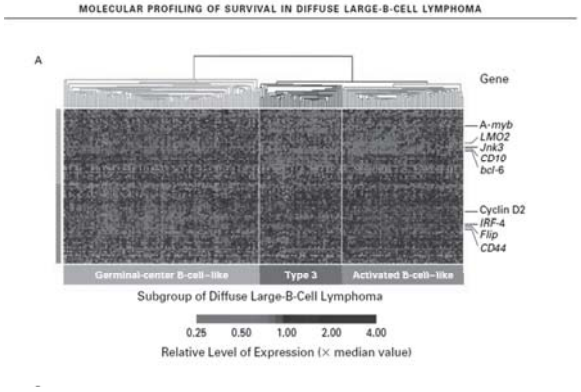
The New England
Journal of Medicine

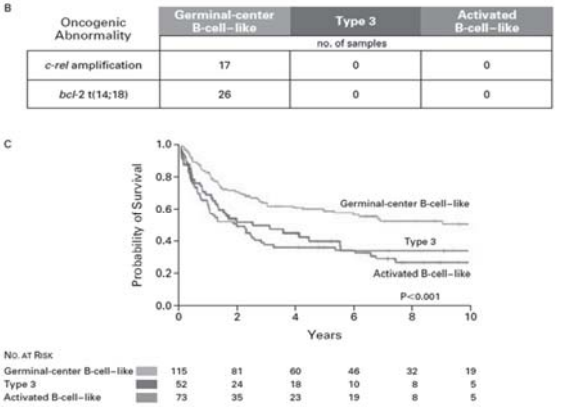
Copyright © 2002 by the Massachusetts Medical Society

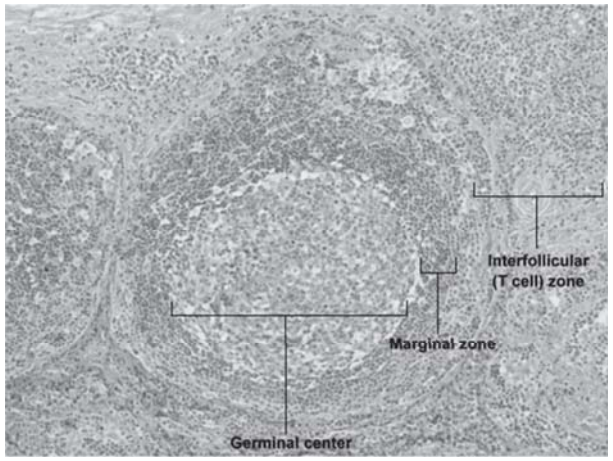
VOLUME 346 — JUNE 20, 2002 — NUMBER 25

THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA

ANDREAS ROSENWALD, M.D., GEORGE WRIGHT, PH.D., WING C. CHAN, M.D., JOSEPH M. CONNORS, M.D., ELIAS CAMPO, M.D., RICHARD I. FISHER, M.D., RANDY D. GASCOYNE, M.D., H. KONRAD MULLER-HERMELINK, M.D., ERLEND B. SMELAND, M.D., PH.D., AND LOUIS M. STAUDT, M.D., PH.D., FOR THE LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT







Methodology for Big Data type studies

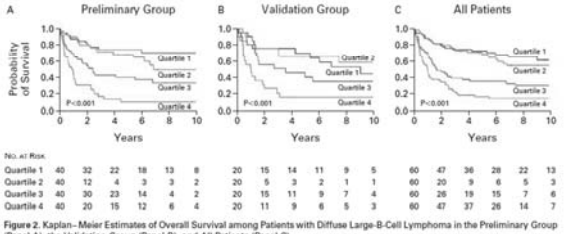


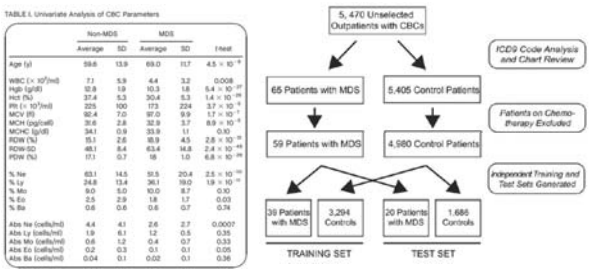
Figure 2. Kaplan–Meier Estimates of Overall Survival among Patients with Diffuse Large-B-Cell Lymphoma in the Preliminary Group (Panel A), the Validation Group (Panel B), and All Patients (Panel C).

RESEARCH ARTICLE

AJH

Automated screening for myelodysplastic syndromes through analysis of complete blood count and cell population data parameters

Philipp W. Raess,[1] Gert-Jan M. van de Geijn,[2] Tjin L. Njo,[2] Boudewijn Klop,[3] Dmitry Sukhachev,[4] Gerald Wertheim,[1,5] Tom McAleer,[1] Stephen R. Master,[1]* and Adam Bagg[1]



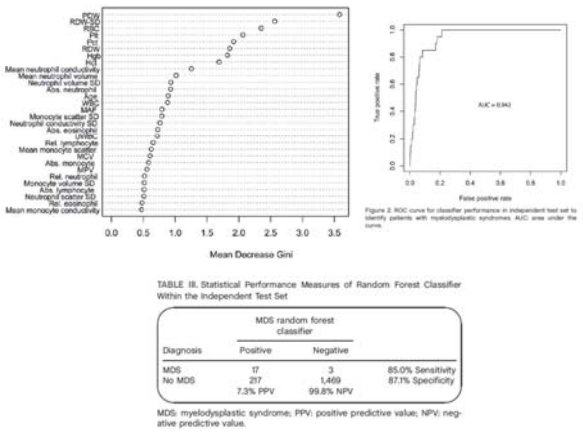


Figure 2. ROC curve for classifier performance in independent test set to identify patients with myelodysplastic syndromes. AUC: area under the curve.

TABLE III. Statistical Performance Measures of Random Forest Classifier Within the Independent Test Set

	MDS random forest classifier		
Diagnosis	Positive	Negative	
MDS	17	3	85.0% Sensitivity
No MDS	217	1,469	87.1% Specificity
	7.3% PPV	99.8% NPV	

MDS: myelodysplastic syndrome; PPV: positive predictive value; NPV: negative predictive value.

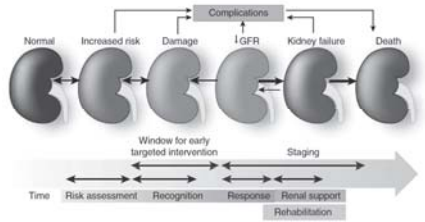
Acute Kidney Injury & Current Diagnosis

- Per the 2012 KDIGO guideline, acute kidney injury (AKI) is defined as:
  - Increase in serum creatinine (sCr) by ≥0.3 mg/dl (≥26.5 μmol/l) within 48 hours; or
  - Increase in sCr to ≥1.5 times baseline, which is known or presumed to have occurred within the prior 7 days; or
  - Urine volume <0.5 ml/kg/h for 6 hours.
- AKI is further staged (stage 1-3) based on severity
- The current AKI diagnosis relies largely on functional markers, which are not sensitive and often lead to a late or inaccurate diagnosis with adverse outcomes and high mortality

KDIGO Clinical Practice Guideline for Acute Kidney Injury. 2012: http://www.kdigo.org/clinical_practice_guidelines/pdf/KDIGO%20AKI%20Guideline.pdf
Lewington AJP, Cerda J, Mehta RL. Raising Awareness of Acute Kidney Injury: A Global Perspective of a Silent Killer. Kidney Int. 2013;84(3):457-467.

The Pathogenic Progression of AKI Is Often Silent

- AKI refers to a spectrum of sub-clinical and clinical conditions and the development of AKI is often silent without early warning signs or symptoms
- Serum Creatinine does not change until ~50% of kidney function loss



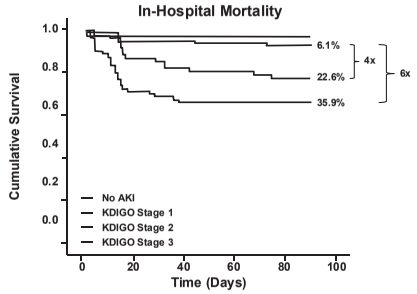
Lewington AJP, Cerda J, Mehta RL. Raising Awareness of Acute Kidney Injury: A Global Perspective of a Silent Killer. Kidney Int. 2013;84(3):457-467.
Marrensson J, Martling CR, Bell M. *Br J Anesth*, 2012, 109 (6): 843-850

Acute Kidney Injury is Prevalent

- Acute kidney injury (AKI) is a common, under-recognized, but potentially reversible loss of kidney function that develops abruptly (over hours to days)
  - 7-18% of hospitalized patients developed AKI
  - Up to 50% of critically ill patients develop some stage of AKI
  - ~2 million AKI-related deaths annually worldwide
- Common causes include major surgery, sepsis, circulatory shock, nephrotoxic drugs, radiocontrast agent, etc.
- Estimated annual costs to US healthcare system attributable to hospital acquired AKI is > $10 billion

Lewington AJP, Cerda J, Mehta RL Raising Awareness of Acute Kidney Injury: A Global Perspective of a Silent Killer. *Kidney Int.* 2013;84(3):457-467.
Mandelbaum T, Scott D, et al. Outcome of Critically Ill Patients with Acute Kidney Injury using the AKIN Criteria. *Crit Care Med.* 2011;39(12):2659–2664.
Ali T, Khan I, et al. Incidence and Outcomes in Acute Kidney Injury: A Comprehensive Population-Based Study. *J Am Soc Nephrol.* 2007;18(4):1292–1298.
Chertow GM, et al. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J. Am. Soc. Nephrol.* 2005;16:3365-3370.

Short Term Survival Decreased with Increased AKI Severity



Aitken E, Carruthers C, Gall L, Kerr L, Geddes C, Kingsmore D. Acute Kidney Injury: Outcomes and Quality of Care. *Q J Med.* 2013;106:323-332.

AKI Drives Up Length of Stay, Hospital Costs, and Mortality

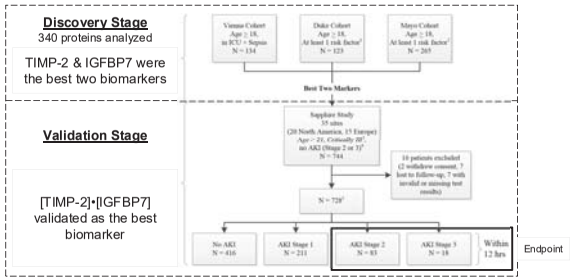
- A study on 50,314 adult surgical patients admitted to the University of Florida for longer than 24 hours between Jan 2000 and Nov 2010
- Incremental hospital costs and mortality associated with AKI following major surgery

	Mild AKI RIFLE-R (Risk)	Moderate AKI RIFLE-I (Injury)	Severe AKI RIFLE-F (Failure)
Length-of-Stay (days)	+4.0	+7.7	+13.2
Cost per Patient ($)	+$10,700	+$21,400	+$38,200
Hospital Mortality (%)	+0.7%	+1.5%	+4.7%

Increasing Severity

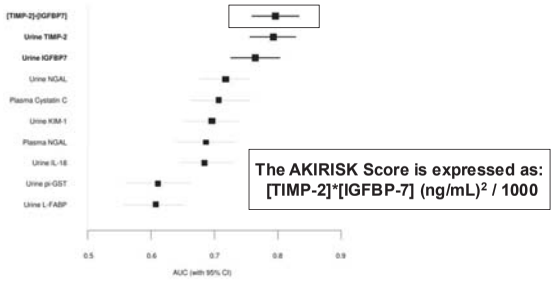
Hobson CE, *et al.* Cost and Mortality Associated With Postoperative Acute Kidney Injury. *Annals of Surgery.* 2014;00:1–8

The Discovery and Validation of Two New Markers



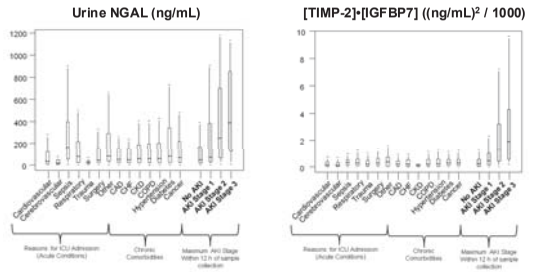
Kashani *et al.* Critical Care 2013, 17:R25

TIMP-2 and IGFBP-7 Outperform Existing Biomarkers



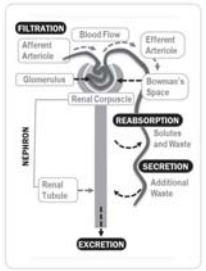
Kashani K, *et al.*, Discovery and validation of cell cycle arrest biomarkers in human acute kidney injury, *Critical Care* 2013, 17:R25

[TIMP-2]•[IGFBP-7] Demonstrated a Compelling Specificity Profile



Kashani K, *et al.*, *Critical Care* 2013, 17:R25

Biological rationale explained **after** the identification of the biomarkers



- Tissue inhibitor of metalloproteinase 2 (TIMP-2) and insulin-like growth factor binding protein 7 (IGFBP-7) are produced by renal tubular cells and are involved in G1 cell cycle arrest
- TIMP-2 and IGFBP-7 are known to be involved in the response to a wide variety of insults (inflammation, oxidative stress, ultraviolet radiation, drugs, and toxins) and are elevated in earliest stages of stress, which may explain why they correspond to risk of AKI.

Kellum JA and Chawla LS. Cell cycle arrest and acute kidney injury: the light and the dark side, *Nephrol Dial Transplant* 2016;31(1):16-22.

Research Article aci Applied Clinical Informatics 212

Diagnostic Performance of Electronic Syndromic Surveillance Systems in Acute Care

A Systematic Review

M. Kashiouris[1,2]; J.C. O'Horo[1,2]; B.W. Pickering[2,3]; V. Herasevich[2,3]

[1]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA;
[2]M.E.T.R.I.C., Mayo Clinic, Rochester, Minnesota, USA;
[3]Department of Anesthesiology, Division of Critical Care Medicine, Mayo Clinic, Rochester, Minnesota, USA

Table 2 Summary of performance of ESS studies in comparison with the gold standard. Study ID numbers correspond to the labels in the likelihood matrix (figure 2)

Study ID	Author and Year	Electronic Syndromic Surveillance Purpose	Type of Study	N of decision points	Sensitivity		Specificity		Positive Likelihood Ratio		Negative Likelihood Ratio	
						95% CI		95% CI		95% CI		95% CI
1	Sloutsr et al 2006 (19)	Seizures (distinguish between seizure and nonseizure epochs)	Retrospective	150	0.89	0.75 to 0.97	0.50	0.40 to 0.60	1.79	1.44 to 2.22	0.21	0.08 to 0.54
2	Kho et al 2007 (20)	Identification of rapid patient deterioration	Prospective, nonrandomized	1878	0.54	0.39 to 0.68	0.71	0.69 to 0.73	1.86	1.43 to 2.43	0.65	0.48 to 0.88
3	Azzam et al 2009 (13)	Acute Lung Injury	Retrospective	199	0.87	0.75 to 0.95	0.89	0.83 to 0.94	7.92	4.93 to 2.72	0.15	0.07 to 0.30
4	Mojoli et al 2010 (17)	Patient Ventilator asynchrony	Retrospective	3016	0.99	0.98 to 1.00	0.98	0.97 to 0.99	49.09	36.76 to 65.56	0.01	0.01 to 0.02
5	Garcia-Esquirol et al 2010 (16)	Patient Ventilator asynchrony	Retrospective	1001	0.91	0.88 to 0.94	0.92	0.89 to 0.94	11.03	8.60 to 14.14	0.09	0.07 to 0.14
6	Herasevich et al 2011 (18)	Ventilator-Induced Lung Injury	Retrospective	304	1.00	0.97 to 1.00	0.81	0.74 to 0.86	5.11	3.79 to 6.90	<0.01	<0.01 to 0.08
7	Koenig et al 2011 (14)	Acute Lung Injury	Prospective, nonrandomized	1269	0.96	0.91 to 1.00	0.98	0.97 to 0.98	45.34	39.61 to 59.76	0.01	0.01 to 0.16
8	Schmidt et al 2012 (15)	Acute Lung Injury	Retrospective	161	0.47	0.37 to 0.57	0.92	0.80 to 0.98	5.62	2.16 to 4.68	0.58	0.48 to 0.70
9	Hooper et al 2012 (12)	Systemic Inflammatory Response Syndrome	Prospective, randomized	529	0.99	0.97 to 1.00	0.92	0.88 to 0.95	11.65	7.88 to 17.22	0.01	<0.01 to 0.04
10	Leth et al 2006 (22)	Septicemia, UTI, pneumonia, postoperative wound infection	Prospective, nonrandomized	2820	0.77	0.70 to 0.83	0.99	0.99 to 0.99	85.13	56.71 to 27.79	0.23	0.17 to 0.30
11	Pokorny et al 2006 (21)	Various types of nosocomial infections	Retrospective	1044	1.00	0.98 to 1.00	0.54	0.50 to 0.57	2.14	2.00 to 2.31	<0.01	<0.01 to 0.08

Published in final edited form as:
J Hosp Med. 2015 January ; 10(1): 26–31. doi:10.1002/jhm.2259.

Development, Implementation and Impact of an Automated Early Warning and Response System for Sepsis

Craig A. Umscheid, MD, MSCE[1,2,3,4,5,6], Joel Betesh, MD[1], Christine VanZandbergen, PA, MPH[7], Asaf Hanish, MPH[1], Gordon Tait, BS[7], Mark E. Mikkelsen, MD, MSCE[2,3], Benjamin French, PhD[1,3,4,5], and Barry D. Fuchs, MD, MS[2]

[1]Center for Evidence-based Practice, University of Pennsylvania, Philadelphia, PA
[2]Department of Medicine, University of Pennsylvania, Philadelphia, PA
[3]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA
[4]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA
[5]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA
[6]Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA
[7]Information Services, University of Pennsylvania, Philadelphia, PA

	Hospitals A-C		
	Pre	Post	p-value
N alerts	101	345	
→ 500cc IV bolus order <3hrs	92 (13%)	142 (28%)	<.01
IV/PO antibiotic order <3hrs	75 (13%)	123 (23%)	<.01
IV/PO sepsis antibiotic order <3hrs	61 (10%)	83 (16%)	<.01
Lactic acid order <3hrs	57 (10%)	128 (23%)	<.01
Blood culture order <3hrs	68 (11%)	99 (18%)	<.01
Blood gas order <6hrs	53 (9%)	59 (11%)	0.28
CBC or BMP <6 hrs	247 (42%)	219 (40%)	0.65
Vasopressor <6hrs	17 (3%)	21 (4%)	0.35
Bronchodilator administration <6hrs	71 (12%)	64 (12%)	0.92
RBC, plasma or platelet transfusion order <6hrs	31 (5%)	52 (10%)	<.01
Naloxone order <6hrs	0 (0%)	1 (0%)	0.30
AV node blocker order <6hrs	33 (6%)	33 (6%)	0.70
Loop diuretic order <6hrs	33 (6%)	28 (5%)	0.38
CXR <6hrs	92 (15%)	113 (21%)	0.02
CT head, chest or abd <6hrs	29 (5%)	34 (6%)	0.31
Cardiac monitoring (EKG or telemetry) <6hrs	70 (12%)	90 (17%)	0.02

Improved patient care for sepsis thanks to the development of automated surveillance systems.

Penn Medicine News

News Releases | News Blog | Publications & Special Projects | Internal Newsletters

Home - News Releases - Penn Medicine's "Sepsis Sniffer" Generates Faster Sepsis Ca...

News Release

Penn Medicine's "Sepsis Sniffer" Generates Faster Sepsis Care and Suggests Reduced Mortality

Early identification and intervention lead to better results for hospital inpatients

October 09, 2014

PHILADELPHIA — An automated early warning and response system for sepsis developed by Penn Medicine experts has resulted in a marked increase in sepsis identification and care, transfer to the ICU, and an indication of fewer deaths due to sepsis. A study assessing the tool is published online in the *Journal of Hospital Medicine*.

Craig A. Umscheid, MD, MSCE

Sepsis is a potentially life-threatening complication of an infection; it can severely impair the body's organs, causing them to fail. There are as many as three million cases of severe sepsis and *750,000 resulting deaths in the United States annually. Early detection and treatment, typically with* antibiotics and intravenous fluids, is critical for survival.

The Penn prediction tool, dubbed the "sepsis sniffer," uses laboratory and vital-sign data (such as body temperature, heart rate, and blood pressure) in the electronic health record of hospital inpatients to identify those at risk for sepsis. When certain data thresholds are detected, the system automatically sends an electronic communication to physicians, nurses, and other members of a rapid response team who quickly perform a bedside evaluation and take action to stabilize or transfer the patient to the intensive care unit if warranted.

Penn Medicine News

News Releases | News Blog | Publications & Special Projects | Internal Newsletters

Craig A. Umscheid, MD, MSCE

Related Links

Perelman School of Medicine at the University of Pennsylvania

University of Pennsylvania Health System

inpatients to identify those at risk for sepsis. When certain data thresholds are detected, the system automatically sends an electronic communication to physicians, nurses, and other members of a rapid response team who quickly perform a bedside evaluation and take action to stabilize or transfer the patient to the intensive care unit if warranted.

The study developed the prediction tool using 4,575 patients admitted to the University of Pennsylvania Health System (UPHS) in October 2011. The study then validated the tool during a pre-implementation period from June to September 2012, when data on admitted patients was evaluated and alerts triggered in a database, but no notifications were sent to providers on the ground. Outcomes in that control period were then compared to a post-implementation period from June to September 2013. The total number of patients included in the pre and post periods was 31,093.

In both the pre- and post-implementation periods, four percent of patient visits triggered the alert. Analysis revealed 90 percent of those patients received bedside evaluations by the care team within 30 minutes of the alert being issued. In addition, the researchers found that the tool resulted in:

- A two to three-fold increase in orders for tests that could help identify the presence of sepsis
- A 1.5 to two-fold increase in the administration of antibiotics and intravenous fluids
- An increase of more than 50 percent in the proportion of patients quickly transferred to the ICU
- A 50 percent increase in documentation of sepsis in the patients' electronic health record

The study found a lower death rate from sepsis and an increase in the number of patients successfully discharged home, although these findings did not reach statistical significance.

Take Home Message

- Big Data doesn't mean "lots of data".
- It requires a reversal of the traditional scientific method.
- This was made possible thanks to new technology for data collection and analysis.
- Big Data is increasing significantly the rate of scientific discovery, and the time required for these discoveries to bring real life benefits to patients.
- It`s here to stay, so in science as in life:

Keep an open mind – you may like what you will learn

Thank you