Incomplete, Missing, and Wrong:
Working with Dirty Healthcare Data from a Clinical Investigator’s Perspective

Dr. Timothy Imler, MD. MS.
Director of Data Core; Regenstrief Institute
Assistant Professor of Medicine; Division of Gastroenterology; Indiana University
Research Scientist; Center of Biomedical Informatics; Regenstrief Institute

Conflicts

• No conflicts related to this presentation

https://projects.propublica.org/docdollars/
Objectives

1. Describe the healthcare data set (INPC) that is available at Regenstrief
2. Provide examples of incomplete, missing, and wrong data based on real research questions
3. Explain how we work within these constraints to provide meaningful clinical research
4. Discuss how researchers can become involved with INPC research

Incomplete, Missing, and Wrong

{W56.22xA}
Struck by orca, initial encounter

icd10illustrated.com
Objective 1

Describe the healthcare data set (INPC) that is available at Regenstrief

INPC as a National Model

- Number of distinct patients all time: 13,249,595
- Number of distinct patients in 2013: 4,287,896
- Number of distinct patients in 2014: 4,649,836
- Number of data points all time: 4,166,691,607
- Number of text reports all time: 165,203,864
- Number of institutions all time: 103
- Number of hospitals all time: 80
INPC as a National Model

INPC as a National Model
INPC as a National Model

Multimedia Content

SCHOOL OF MEDICINE
INDIANA UNIVERSITY
Multimedia Content

colon AND (adenoma OR SSP OR SSA OR "serrated") AND NOT ("no adenoma"=3 OR "no serrated"=3 OR "negative adenoma"=3 OR "negative serrated"=3)

Final Pathologic Diagnosis:
A. Colon, ascending right, biopsy:
   Fragments of tubular adenoma.
B. Colon, ascending, biopsy:
   Fragments of serrated polyp (sessile serrated adenoma).

Pharmacy Data
Objective 2

Provide examples of incomplete, missing, and wrong data based on real research questions
Incomplete, Missing, and Wrong

• Healthcare data are challenging to analyze due to the “messiness” of the data
  – Multiple sources
  – Differing data structures
  – Competing interests in utilization of data

Example: “You only live twice”

Data Point: Age at event

Which is the real date of birth?
Example: “Under Pressure”

Data Point: Blood pressure

80 (Systolic)

140 (Diastolic)

How can this be real?

Example: “Too Sweet”

Data Point: Patients with DM2

DM2 Code → 45% (DM1)

Why are they miscoded?
Example: “Only the good die young”

Data Point: SIDS age

27 years

How could this happen?

Example: “Weightless”

Data Point: Weight/BMI

<table>
<thead>
<tr>
<th>Obs</th>
<th>HT_IN</th>
<th>HT_CM</th>
<th>WT_LBS</th>
<th>WT_KG</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>1271</td>
<td>5.1</td>
<td>0</td>
<td>330.8</td>
<td>162.8</td>
<td>54.2</td>
</tr>
<tr>
<td>564</td>
<td>5.9</td>
<td>175</td>
<td>103.9</td>
<td>99.8</td>
<td>33.65</td>
</tr>
<tr>
<td>3053</td>
<td>6</td>
<td>180</td>
<td>341</td>
<td>160.3</td>
<td>47</td>
</tr>
<tr>
<td>6980</td>
<td>24</td>
<td>61</td>
<td>204</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>7652</td>
<td>25</td>
<td>64</td>
<td>207</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>5499</td>
<td>59</td>
<td>139</td>
<td>129</td>
<td>62</td>
<td>1.19</td>
</tr>
<tr>
<td>933</td>
<td>60</td>
<td>130</td>
<td>129</td>
<td>62</td>
<td>1.19</td>
</tr>
<tr>
<td>3799</td>
<td>63</td>
<td>243.9</td>
<td>110.6</td>
<td>2.22</td>
<td></td>
</tr>
<tr>
<td>5648</td>
<td>63</td>
<td>167</td>
<td>84.8</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>1969</td>
<td>65</td>
<td>225</td>
<td>168</td>
<td>9.97</td>
<td></td>
</tr>
<tr>
<td>3008</td>
<td>65</td>
<td>59.9</td>
<td>220</td>
<td>167</td>
<td>2.49</td>
</tr>
<tr>
<td>4409</td>
<td>65.5</td>
<td>165</td>
<td>164.7</td>
<td>62.1</td>
<td>0</td>
</tr>
<tr>
<td>2560</td>
<td>66</td>
<td>204.6</td>
<td>122.4</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>8368</td>
<td>69</td>
<td>178</td>
<td>100.8</td>
<td>82</td>
<td>0.36</td>
</tr>
<tr>
<td>2408</td>
<td>71</td>
<td>185</td>
<td>205.6</td>
<td>60.3</td>
<td>0.55</td>
</tr>
<tr>
<td>6037</td>
<td>71</td>
<td>185</td>
<td>37.123</td>
<td>15.839</td>
<td></td>
</tr>
<tr>
<td>3427</td>
<td>137</td>
<td>196</td>
<td>180.5</td>
<td>100.5</td>
<td></td>
</tr>
</tbody>
</table>

How do you interpret?
Example: “Yin and Yang”

Data Point: Preterm birth

Huh?

Incomplete, Missing, and Wrong

{V91.07xD}
Burn due to water-skis on fire, subsequent encounter

icd10illustrated.com
Explain how we work within these constraints to provide meaningful clinical research.

Example: “You only live twice”

Data Point: Age at event

Exclude or assume...
Example: “Under Pressure”

Data Point: Blood pressure

80 (Systolic)
140 (Diastolic)

Exclude or assume…

Example: “Too Sweet”

Data Point: Patients with DM2

DM2 Code → 45% (DM1)

Exclude or assume…
Example: “Weightless”

Data Point: Weight/BMI

<table>
<thead>
<tr>
<th>Obs</th>
<th>HT_IN</th>
<th>HT_CM</th>
<th>WT_LBS</th>
<th>WT_KG</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>89.5</td>
<td></td>
</tr>
<tr>
<td>1271</td>
<td>5.1</td>
<td>183.8</td>
<td>262.8</td>
<td>54.2</td>
<td></td>
</tr>
<tr>
<td>864</td>
<td>5.9</td>
<td>175</td>
<td>103.9</td>
<td>95.8</td>
<td>33.65</td>
</tr>
<tr>
<td>3033</td>
<td>6.0</td>
<td>160.3</td>
<td>100.3</td>
<td>47.2</td>
<td></td>
</tr>
<tr>
<td>6658</td>
<td>24.6</td>
<td>61</td>
<td>264.2</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>7632</td>
<td>25.6</td>
<td>64</td>
<td>207.5</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>9469</td>
<td>58.6</td>
<td>129</td>
<td>63</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>923</td>
<td>60.0</td>
<td>130</td>
<td>123.7</td>
<td>288</td>
<td></td>
</tr>
<tr>
<td>3799</td>
<td>63.0</td>
<td>243.9</td>
<td>110.8</td>
<td>2.22</td>
<td></td>
</tr>
<tr>
<td>5648</td>
<td>63.0</td>
<td>157</td>
<td>84.8</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>1669</td>
<td>65.0</td>
<td>225</td>
<td>108</td>
<td>9.97</td>
<td></td>
</tr>
<tr>
<td>3808</td>
<td>65.9</td>
<td>250</td>
<td>167</td>
<td>2.49</td>
<td></td>
</tr>
<tr>
<td>4409</td>
<td>65.5</td>
<td>165</td>
<td>164.7</td>
<td>62.1</td>
<td>0</td>
</tr>
<tr>
<td>2560</td>
<td>69.0</td>
<td>264.6</td>
<td>122.4</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>8548</td>
<td>69.1</td>
<td>185.8</td>
<td>82</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2498</td>
<td>71.1</td>
<td>205.6</td>
<td>90.3</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>6027</td>
<td>71.1</td>
<td>180.34</td>
<td>37.123</td>
<td>15.839</td>
<td></td>
</tr>
<tr>
<td>3427</td>
<td>137</td>
<td>196</td>
<td>221.5</td>
<td>100.5</td>
<td></td>
</tr>
</tbody>
</table>

Exclude or assume...

Example: “Yin and Yang”

Data Point: Preterm birth

Preterm Birth Rates vs LMP Months 2003

Preterm Birth Rates vs LMP Day

Use a different (better) data set?
Objective 4

Discuss how researchers can become involved with INPC research
Getting Data

https://i2b2ctsi.regenstrief.org

Getting Data

https://www.indianactsi.org/regenstrief/feasibilityassessment
Getting Data

Questions???