The need for big data & big data methods in pediatrics: lessons from the Global Research in Pediatrics Network of Excellence

Prof. dr. Miriam Sturkenboom
Utrecht University Medical Center, The Netherlands
The traditional way to look at evidence
There’s a gap between what we know and what we do...

45.1% of medicine is not evidence based; it takes 17 years to translate science to practice.

Doctors would have to read approximately 29 hours each workday to keep up with new professional insights; 80% of data is unstructured and each of us will produce 300M books of health-related data in our lifetime.

Shifting landscape?

"Data is fundamentally changing the research enterprise and creating new extraordinary opportunities to learn things that were either un-learnable or would have taken generations."

- Stanford Faculty

There will always be an argument for more research and for better data, but waiting for more data is often an implicit decision not to act or to act on the basis of past practice rather than best available evidence. The goal must be actionable data — data that are sufficient for clinical and public health action that have been derived openly and objectively and that enable us to say, "Here's what we recommend and why."

Growth in Health Care Data

2,314 Exabytes

Data is fundamental to changing the research enterprise and creating new extraordinary opportunities to learn things that were either un-learnable or would have taken generations.

- Stanford Faculty

Evidence for Health Decision Making — Beyond Randomized, Controlled Trials

Jeffrey M. Drazen, M.D., David P. Harrington, Ph.D., John J.V. McMurray, M.D., James H. Ware, Ph.D., and Janet Woodcock, M.D., Editors

Thomas R. Frieden, M.D., M.P.H.
Are there areas for ‘big’ data use in regulatory and clinical decision making?

The EMA said:

"Technological advances in both science and information technology are generating ever-increasing amounts of data on health and medicines. The objective of this workshop was to increase understanding of how big data will impact on our understanding of disease and facilitate medicines development, so that the regulatory community can identify opportunities and address challenges in its use for medicines decision-making. In his opening remarks, Professor Guido Rasi (Executive Director, EMA) emphasised the clear potential of big data to benefit patients.

"However, it is challenging to incorporate these data in a meaningful way into routine regulatory decision-making and importantly to understand how to determine whether the conclusions and associations arising from multiple analyses across varied data sets are causal and not simply spurious coincidence. Workshop participants included patient representatives, healthcare professionals, and representatives from government, industry, and academia, as well as regulators from across the globe."

According to US Food and Drug Administration
What does big data offer?

- **Breadth** – large numbers of individuals get us closer to the underlying source population –

- **Depth** – increasing amount of data on each individual increases the chance that we will have measures of likely confounders

- **Diversity** – different types of data offer the potential to “cross check” findings for any particular data source

- **FDA-Sentinel system**: more than 100 million patient health care data

From: D Martin EMA big data workshop
Need for bigger data and big data approaches in pediatrics to support decision making: some lessons from the Global Research in Pediatrics project (FP-7 EC)

www.grip-network.org
GRiP was created to address the lack of appropriate testing and information on paediatric drugs. GRiP partners are working to reduce the current fragmentation of the efforts to study and develop the use of medicine in children.
In spite of all new trials following pediatric regulations

- Big Health Data in children are generated every day in routine healthcare
  - Spontaneous reports of adverse events
  - Medical records (GPs, paediatricians)
  - Registries (vaccinations)
  - Claims records (pharmacy dispensings, hospitalizations ...)
  - ..... 

- These should be used to study the effects of drugs in children and learn about use, benefits and safety
Attempt to establish global pediatric pharmacoepidemiological platform

1. Literature studies (What type of data/studies used to date)
2. Identification of data sources with ped. data
3. Methods & tools development
4. Proof of concept

GRIP e-learning module in pediatric pharmacoepidemiology & pharmacovigilance
Literature review on safety and effectiveness studies to study current state of art
Inventory: current state of the art in pediatric pharmacoepidemiology?

- Which designs are applied?
- Which data sources are used
- Which methods are used?
Drug safety studies

• Safety studies (268)

Pharmacoepidemiological safety studies in children: a systematic review

Osemeke U. Osokogu1*, Julijana Dukanovic1, Carmen Ferrajolo1, Caitlin Dodd1, Alexandra C. Pacurariu1, Wichor M. Bramer2, Geert ’t Jong3, Daniel Weibel1, Miriam C. J. M. Sturkenboom1 and Florentia Kaguelidou1,4

1 Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
2 Medical Library, Erasmus University Medical Center, Rotterdam, The Netherlands
3 Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, Canada
4 Department of Pediatric Pharmacology and Pharmacogenetics, Hôpital Robert Debré, APHP, Univ Paris 7-Diderot, Sorbonne Paris Cité, EA08, INSERM CIC1426, Paris, France

Results literature review paediatric safety studies

Results literature review paediatric safety studies

- Location: North America (154 [57.5%]) or Europe: (92 [34.3%])
- Only 75 child only studies

- Type of compounds: 147 [54.9%] small molecules, rest vaccines

- Data source:
  - Studies utilizing secondary data: have larger sample sizes
  - Paper medical charts: Main source for
    - exposure (85 [31.7%]) and outcome (122 [45.5%]) data

- Design:
  - Cohort studies: most common (174 [64.9%])
  - SCCS - 30 (11.2%)
  - Case crossover - 4 (1.5%)

Conclusions safety review

Key points

- The number of pharmacoepidemiological safety studies is steadily increasing in pediatrics.
- We identified various challenges including funding, design, type and source of data, mode of data collection, age and geographic spread of the investigated population, studied drugs and outcomes, sample size, control of confounding and reporting of results.
- Pharmacoepidemiological safety studies in children can be improved in several ways including global collaboration.

Results literature review paediatric effectiveness studies

- **Effectiveness studies (164)**

  Dukanovic J et al. Manuscript submitted for publication
Results literature review paediatric effectiveness studies

Fig. 2: Number of pharmacoepidemiological effectiveness studies

Dukanovic J et al. Manuscript submitted for publication
Discrepancy between use of drugs in pediatrics and studies

Comparison between evaluated and routinely utilized drugs

Dukanovic J et al. Manuscript submitted for publication
Conclusions paediatric safety & effectiveness studies

- Studies are conducted mainly in developed countries
- Increased number of studies following the introduction of the BPCA (US) and pediatric legislation (EU)
- Use of more modern methods (propensity scores) especially for effectiveness studies
- Many intermediate outcomes instead of clinical outcomes
- Most studies rely on traditional data collection, opportunity for use of electronic health record data
- Need to use more modern methods (propensity scores) for confounding
- Data pooling needed to achieve desired sample size and ability to look at hard outcomes
- Increased capacity needed for conduct of these studies

Dukanovic J et al. Manuscript submitted for publication
Are there available big health data sources that can be used to generate evidence on the effects of drugs in children?
Spontaneous reporting databases

Spontaneous reporting*

‘Spontaneous’ (or voluntary) reporting of adverse effects is when health professionals or patients decide that they will report suspected harm from a medicine to their local or national pharmacovigilance centre.

VigiBase, the WHO database of individual case safety reports
Spontaneous reports: FAERS (public version)

Distribution of pediatric ICSRs (N = 106,122) within FAERS according to age-category.

<table>
<thead>
<tr>
<th></th>
<th>Total N = 106,122 (%)</th>
<th>≤ 27 days N = 4,717 (4.4%)</th>
<th>28 days–23 months N = 16,096 (15.2%)</th>
<th>2–11 years N = 47,248 (44.5%)</th>
<th>12–17years N = 38,061 (35.9%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>54,768 (54.5%)</td>
<td>2,114 (54.1%)</td>
<td>7,921 (55.3%)</td>
<td>27,075 (59.9%)</td>
<td>17,658 (47.7%)</td>
</tr>
<tr>
<td>Mean age (95%CI)</td>
<td>9.1 (9.0–9.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reported drugs</td>
<td>236,491</td>
<td>12,180 (5.2%)</td>
<td>34,575 (14.6%)</td>
<td>103,988 (44.0%)</td>
<td>85,748 (36.3%)</td>
</tr>
<tr>
<td>Drugs/ICSR [median (IQR)]</td>
<td>1 (1–3)</td>
<td>1 (1–3)</td>
<td>1 (1–3)</td>
<td>1 (1–3)</td>
<td>1 (1–3)</td>
</tr>
<tr>
<td>Reported events</td>
<td>397,220</td>
<td>21,265 (5.4%)</td>
<td>59,306 (14.9%)</td>
<td>173,395 (43.7%)</td>
<td>143,254 (36.1%)</td>
</tr>
<tr>
<td>Events/ICSR [median (IQR)]</td>
<td>1 (1–1)</td>
<td>1 (1–2)</td>
<td>1 (1–2)</td>
<td>1 (1–1)</td>
<td>1 (1–1)</td>
</tr>
</tbody>
</table>

Spontaneous reports EUDRAVIGILANCE (Academic version)

Table 1 Description of pediatric ADR reports by age categories in EUDRAVIGILANCE

<table>
<thead>
<tr>
<th>Age group</th>
<th>Number of DECs, n (%) full set</th>
<th>Number of DECs, n (%) vaccines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infants: 0 days-23 months</td>
<td>402,817</td>
<td>208,658</td>
</tr>
<tr>
<td>Children: 2-11 years</td>
<td>406,136</td>
<td>72,271</td>
</tr>
<tr>
<td>Adolescents: 12-17 years</td>
<td>368,422</td>
<td>60,064</td>
</tr>
<tr>
<td>Total</td>
<td><strong>1,177,375</strong></td>
<td><strong>340,993</strong></td>
</tr>
</tbody>
</table>

Dodd CN et al., manuscript in preparation
Spontaneous reports VAERS: public version

<table>
<thead>
<tr>
<th>Age group</th>
<th>Number of vaccine-event combinations, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infants: 0 days-23 months</td>
<td>848,365 (54%)</td>
</tr>
<tr>
<td>Children: 2-11 years</td>
<td>437,082 (28%)</td>
</tr>
<tr>
<td>Adolescents: 12-17 years</td>
<td>271,216 (17%)</td>
</tr>
<tr>
<td>Total</td>
<td>1,556,663 (100%)</td>
</tr>
</tbody>
</table>
Conclusion SRS

- Millions of spontaneous reports are available for pediatrics in publicly accessible datasources.
- Each source has different structure.
- Methods for cleaning, deduplication and pooling of data might improve ability to do data mining in pediatrics specifically.
Population based pediatric health care data
Identify healthcare databases comprising paediatric data (2012)

Published: Ferrajolo C. https://repub.eur.nl/pub/77131/
Chapter 6
Identify healthcare databases comprising paediatric data (2012)

64 responses out of 125, 34 willing to participate in GRIP: Health care data on more than 50 million children

Published: Ferrajolo C. https://repub.eur.nl/pub/77131/ Chapter 6
Conclusion on availability of paediatric data for use, effectiveness and safety studies

- **Spontaneous reports**: millions of reports on pediatrics are publicly available

- **Health care records**: Data on many children available around the world, databases with >50 million children willing to collaborate
We need to pool and combine to increase ability to detect in pediatrics
EU-ADR network: 8 databases, 5 million children, 2170 different drugs, 25 million PY follow-up

“The 1.6 million PYs of exposure were distributed over 2170 individual drugs, compared with 2289 for the overall population (all ages) in the database network. Of these, only 18 represented 50% and 158 drugs represented 90% of the total drug exposure time.

Drug exposure in person-years by age. Note: Drug exposure is aggregated on the first ATC level (anatomical main group). ‘Other’ represents all other drug groups with a total exposure of <5000 PYs. • respiratory, • anti-Infectives, • dermatological, • genitourinary, • alimentary, • neurologic, • blood, • other (<5 000 PYs)

Methods in electronic health care databases: size & power for paediatric studies within EU-ADR databases with 5 million children

Global collaboration is needed

Methods & tools to mine big health data in pediatrics
methods for paediatric signal detection in spontaneous reporting databases
Goal

• What are the best methods to mine for safety signals in children in spontaneous reporting databases?

• Comparison of performance of different data mining methods in spontaneous reporting databases
GRIP workflow

- Creation of reference sets
- Cleaning & completion of VAERS, FAERS and EUDRAVIGILANCE sets (machine learning)
- Terminology mapping
- Running and analysis
Pediatric Drug Safety Signal Detection: A New Drug–Event Reference Set for Performance Testing of Data-Mining Methods and Systems

Osemeke U. Osokogu · Federica Fregonese · Carmen Ferrajolo · Katia Verhamme · Sandra de Bie · Geert ’t Jong · Mariana Catapano · Daniel Weibel · Florentia Kaguelidou · Wichor M. Bramer · Yingfen Hsia · Ian C. K. Wong · Madlen Gazarin · Jan Bonhoeffer · Miriam Sturkenboom
### Table 2: Classification of each drug–event pair as positive control (green: PC1 or PC2) or negative control (red: NC2)

<table>
<thead>
<tr>
<th>Selected Adverse Events</th>
<th>Bullous eruption</th>
<th>Aplastic anemia</th>
<th>Agranulocytosis</th>
<th>Thrombocytopenia</th>
<th>Psychosis</th>
<th>Suicide</th>
<th>Vent. arrhythmia</th>
<th>Sudden death</th>
<th>QT prolongation</th>
<th>Venous thromboembolism</th>
<th>Anaphylaxis</th>
<th>Seizure</th>
<th>Acute kidney injury</th>
<th>Acute liver injury</th>
<th>Sepsis</th>
<th>SIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>flucloxacillin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clarithromycin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doxycline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lopinavir</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>isoniazid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>praziquantel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mebendazole</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quinine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fluoxetine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>montelukast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>isoretinoin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>loperamide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>domperidone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>methylphenidate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fluoxetine f/H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cyproterone f/H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** Vent. - ventricular; SIDS - Sudden Infant Death Syndrome; eth.est. - ethinylestradiol
Reference set for performance testing of pediatric vaccine safety signal detection methods and systems

Yolanda Brauchli Pernus, Cassandra Nan, Thomas Verstraeten, Mariia Pedenko, Osemeke U. Osokoglu, Daniel Weibel, Miriam Sturkenboom, Jan Bonhoeffer, on behalf of the GRIP consortium

Brighton Collaboration Foundation, Switzerland
P95 Pharmacovigilance and Epidemiology Services, Leuven, Belgium
Erasmus University Medical Center, The Netherlands
University Children’s Hospital Basel, University Basel, Switzerland

Received 1 October 2015, Accepted 5 October 2015, Available online 20 October 2015

doi:10.1016/j.vaccine.2015.10.013
Cleaning spontaneous reporting datasets: enhancing quality of big data
Cleaning steps

• Creation of GRIP spontaneous reporting common data model
• Deduplication of records (within and between FAERS/VAERS/Eudravigilance)
• Coding of events (MEDDRA) and drugs (ATC)
  – Mapping tool (Machine learning) for coding of drug names
• Transfer of FAERS, VAERS and EUDRAVIGILANCE in common data model
testing automated signal identification methods (example)

FAERS (FDA adverse events)
EUDRAVIGILANCE (EMA)
VAERS (CDC adverse events vaccines)
No difference in performance of methods on FAERS, adjustment for age worsens performance

Fig. 3 Performance of signal detection algorithms within the entire pediatric population

<table>
<thead>
<tr>
<th>SDA</th>
<th>Number of reports</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50.62</td>
<td>67.57</td>
<td>50.62</td>
<td>67.57</td>
<td>0.634</td>
</tr>
<tr>
<td>PRR</td>
<td></td>
<td>64.71</td>
<td>63.27</td>
<td>37.93</td>
<td>83.78</td>
<td>0.731</td>
</tr>
<tr>
<td>EBGM</td>
<td></td>
<td>62.50</td>
<td>58.62</td>
<td>17.24</td>
<td>91.89</td>
<td>0.745</td>
</tr>
</tbody>
</table>

*After age adjustment*

| PRR       |                   | 66.67       | 62.75       | 34.48| 86.49| 0.688|
| EBGM      |                   | 50.00       | 56.90       | 13.79| 89.19| 0.683|

SDA—signal detection algorithm; PRR = Proportional reporting ratio; EBGM = Empirical Bayes Geometric Mean; AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value

*adjusted PRR/ROR values calculated by combining the individual estimates from each age stratum into one measure according to the Mantel-Haenszel approach.

Impact of age stratification, some signals unmasked

Fig. 2 Variation of PRR and EBGM estimates across pediatric specific strata —selected examples

Recommendations for pediatric signal detection on FAERS

- The Signal detection algorithms showed good performance on pediatric data and can be utilized for pediatric signal detection.
- Age adjustment did not improve the performance of the SDAs.
- Age stratification showed that some signals may be detected only in specific pediatric age groups. For routine surveillance, checking for effect modification across age-strata may generate useful information.
methods for conducting studies in big health care databases
methods for conducting studies in electronic health care databases: how to estimate incidence & prevalence in children given dynamic populations
Explanation of issue

- Health care databases comprise
  - Population file
  - All drugs prescribed/dispensed
  - Events (primary care, hospitalization)
- On registered population
- However population is dynamic and we only see a fraction of the ‘life’

Osokogu: https://repub.eur.nl/pub/95504/
Methods in electronic health care databases: estimation of incidence & prevalence:
impact of episode duration on incidence estimation

![Graph showing the incidence of acute otitis media across different age categories and episode durations.](https://repub.eur.nl/pub/95504/)

Osokogu: https://repub.eur.nl/pub/95504/
Methods in electronic health care databases: estimation of incidence & prevalence: recommendations for studies

Outcome

Transient/recurrent

Testing assumptions regarding duration of an episode

Common outcome

Do the assumptions impact the incidence or prevalence?

YES - BOTH

Rare outcome

DO THE ASSUMPTIONS IMPACT THE INCIDENCE OR PREVALENCE?

NO

Chronic

Testing assumptions regarding length of the run-in period

Common outcome

Do the assumptions impact the incidence or prevalence?

YES - INCIDENCE

Rare outcome

Do the assumptions impact the incidence or prevalence?

YES - INCIDENCE

Figure 4: Summary of the impact of assumptions on the investigated outcomes

Osokogu: https://repub.eur.nl/pub/95504/
methods for conducting studies in big health care databases:

how to best adjust for confounding in pediatric observational studies?
Propensity scores

• Propensity score: statistical model that predicts the ‘assignment of treatment based on covariates’
• Allows for matching on this score to create balance between the different treatment groups
• Not much used in pediatrics and not clear what the ‘look back period’ should be (co-morbidity may be acute)
• Used an example regarding effectiveness of asthma medication re exacerbations in IPCI

Osokogu: https://repub.eur.nl/pub/95504/
Design of methods study on look back period for construct of propensity score

Study entry:
Diagnosis of asthma + incident prescription of 'ICS-LABA' (fixed or loose) after a run-in period of one year

- 1 week
- 1 month
- 3 months
- 1 year
- Full history
The impact of different look back periods and the choice of the way to implement the PS are important. The results on a matched analysis are comparable to clinical trial data on the comparison between fixed and loose ICS+LABA combinations in preventing worsening of asthma.

Osokogu: https://repub.eur.nl/pub/95504/
Summary

• ‘Big data’ is a great & challenging opportunity also in pediatrics, many data sources are available

• Many applications can be found/exist where big data analysis may assist clinical and regulatory decision making

• Computing and data facilities for distributed systems need to be improved but great developments are on the way, in pediatrics global collaboration is needed!

• Collaboration needed between Data Scientists, pharmacologists, regulators, epidemiologists, pediatricians to improve the field

• Machine learning methods can help, but the human mind will remain necessary for interpretation & generalization
Thank you