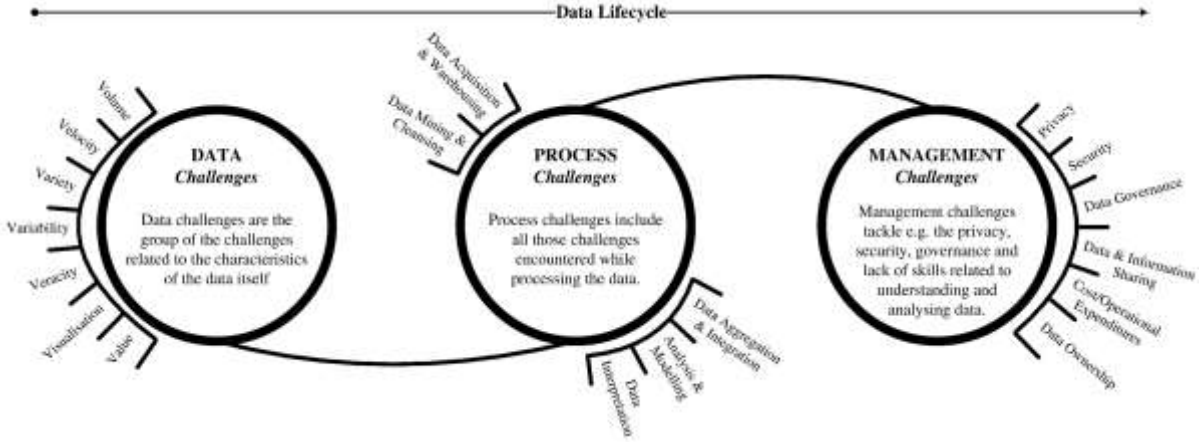
Big data: challenges for research

Samuel S. Allemann, PhD

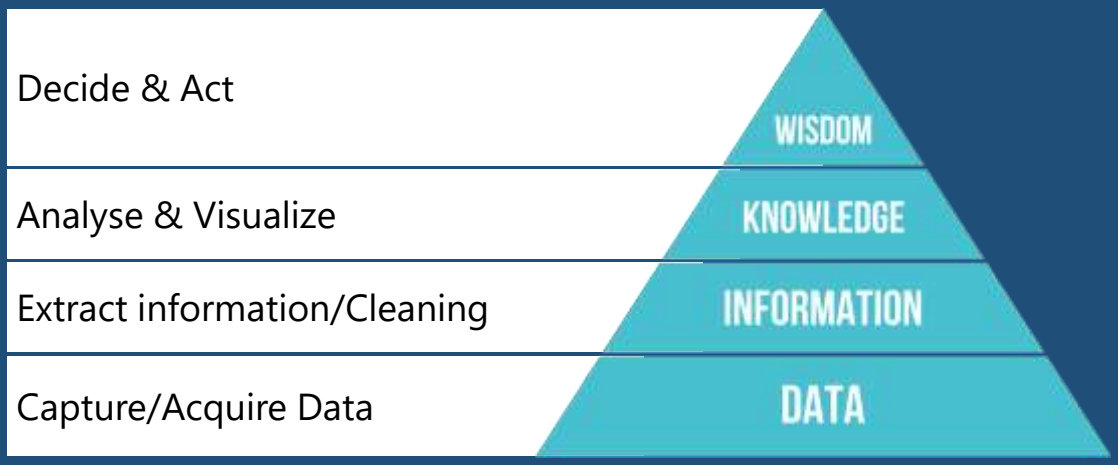
ESCP Symposium: The Digital Revolution | 25.10.2019, Ljubljana

Challenges of Big Data

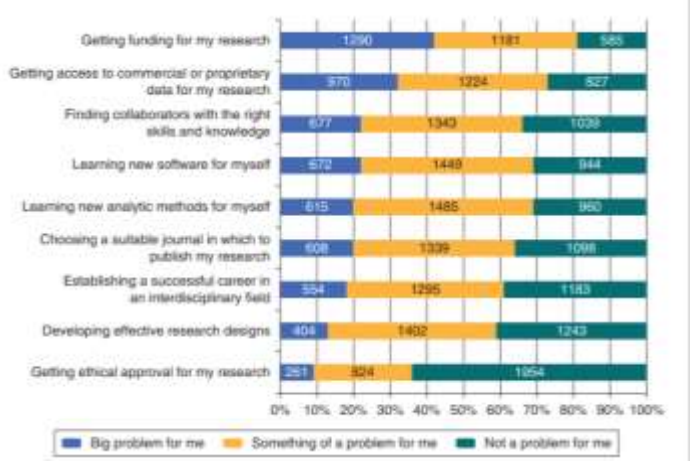


Sivarajah, Uthayasankar et al. *Journal of Business Research* 70 (1. Januar 2017): 263–86. https://doi.org/10.1016/j.jbusres.2016.08.001.

Big Data + Research = Data Science?

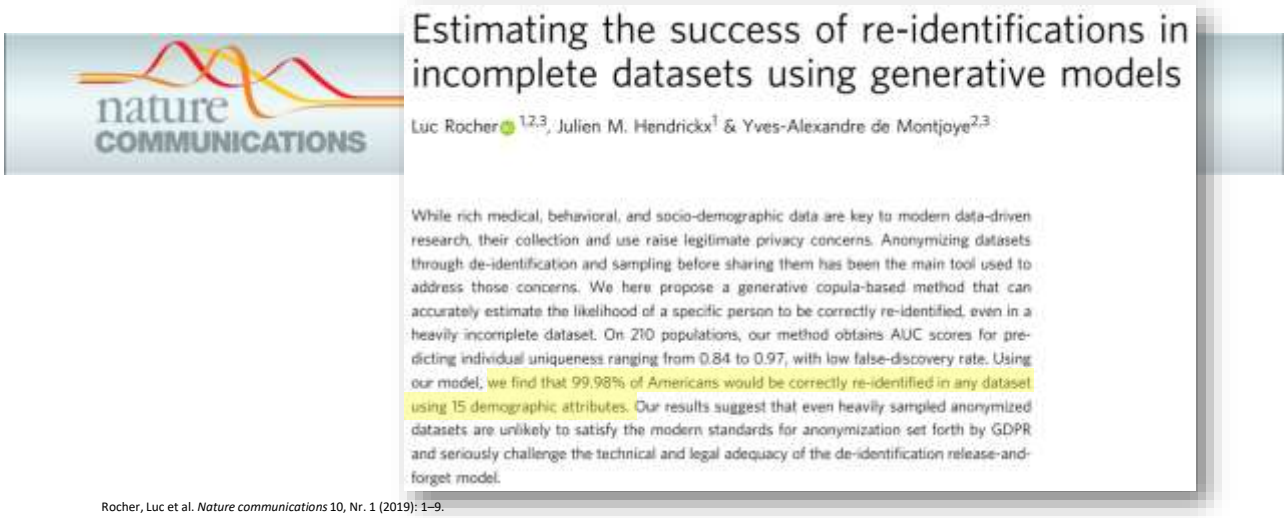


Challenges facing big data researchers



Metzler, Katie, David A Kim, Nick Allum, und Angella Denman. „Who is doing computational social science? Trends in Big Data research", 2016.

Privacy



Rocher, Luc et al. *Nature communications* 10, Nr. 1 (2019): 1–9.

FAIR Data

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE

Wilkinson et al. *Scientific data* 3 (2016).

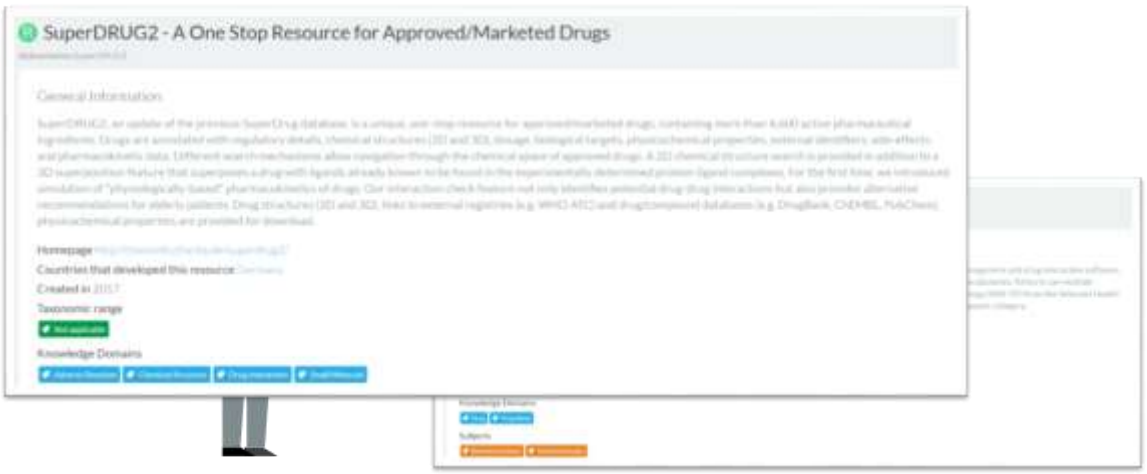
All of Us
RESEARCH PROGRAM

There's a gap in medical research that only you can fill.

The *All of Us* Research Program has a simple mission. We want to speed up health research breakthroughs. To do this, we're asking one million people to share health information. In the future, researchers can use this to conduct thousands of health studies.

All of Us Research Program Investigators. *New England Journal of Medicine* 381, Nr. 7 (2019): 668–76.

Fairshare

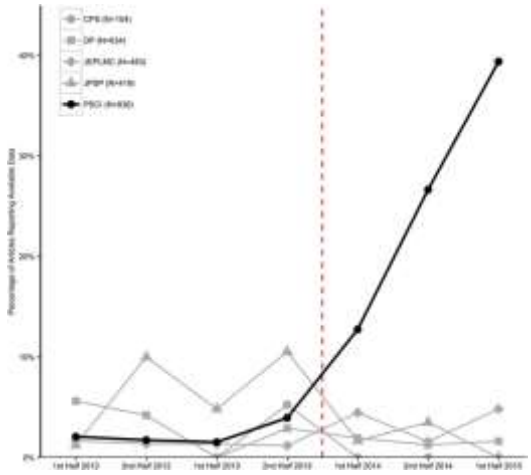


Open Science Badges

https://cos.io/

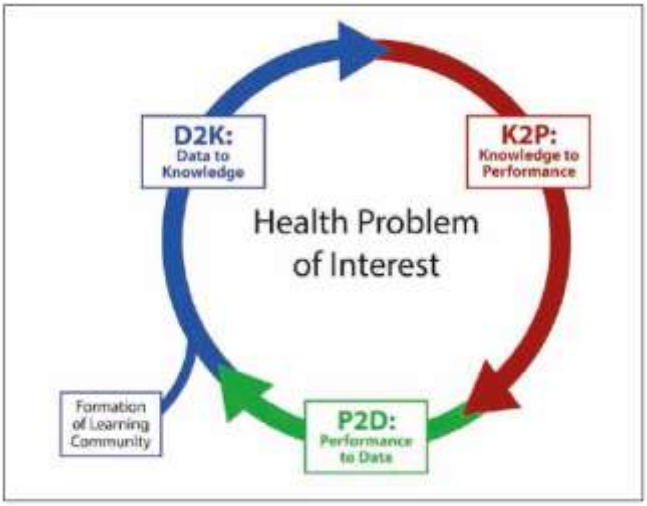
Rowhani-Farid, Anisa, Michelle Allen, und Adrian G. Barnett. *Research Integrity and Peer Review* 2, Nr. 1 (5. Mai 2017): 4. https://doi.org/10.1186/s41073-017-0028-9.

Open Science Badges Success



Kidwell MC et al. (2016) PLOS Biology 14(5): e1002456. https://doi.org/10.1371/journal.pbio.1002456 https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002456

Learning Healthcare System



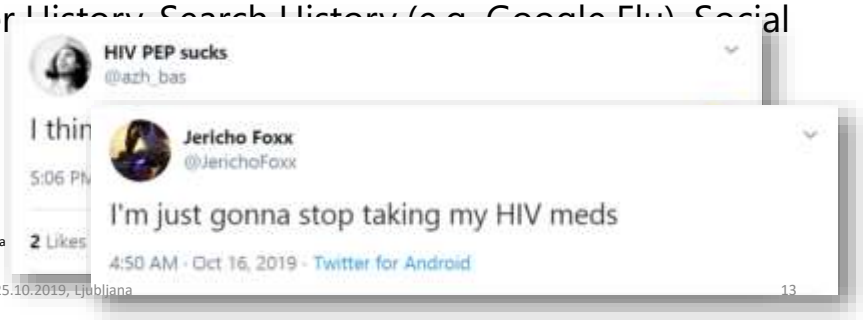
Friedman, C. P., J. C. Rubin, und K. J. Sullivan. *Yearbook of Medical Informatics* 26, Nr. 1 (August 2017): 16–23. https://doi.org/10.15265/IY-2017-004.

Medication Adherence and Big Data?

„Medicine Use and Spending in the U.S.: A Review of 2018 and Outlook to 2023". IQVIA Institute for Human Data Science, Mai 2019. https://www.iqvia.com/insights/the-iqvia-institute/reports/medicine-use-and-spending-in-the-us-a-review-of-2018-and-outlook-to-2023.

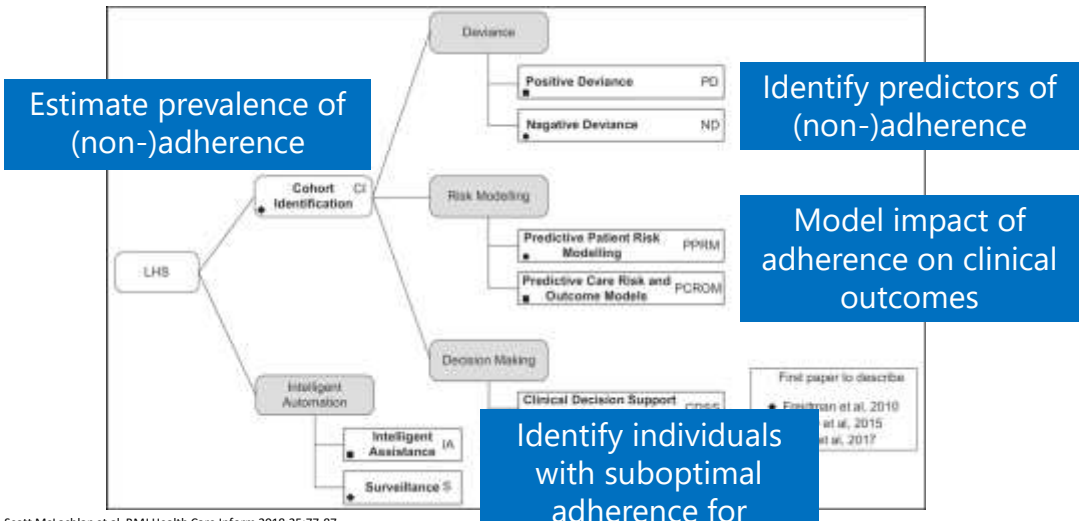
Capture Adherence Big Data

- Administrative Data (Claims, EHR → Electronic Healthcare Data)
- Mobile Phones: Sensor data, Ecological momentary assessment
- Biomedical Data (e.g. Proteus Digital Pill, laboratory & point-of-care testing)
- Internet: Browser History, Search History (e.g. Google Flu), Social media



Huang, Tao, Liang Lan, Xuexian Fang, Peng An, Junxia
https://doi.org/10.1016/j.bdr.2015.02.002.

Adherence: LHS Opportunities



Estimate prevalence of (non-)adherence

Identify predictors of (non-)adherence

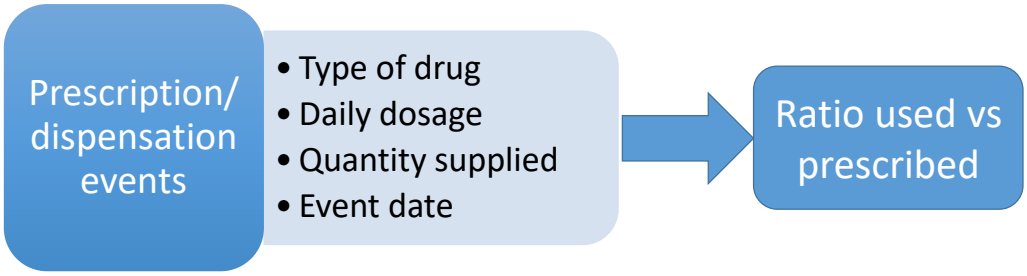
Model impact of adherence on clinical outcomes

Identify individuals with suboptimal adherence for targeted interventions

Scott McLachlan et al. BMJ Health Care Inform 2018;25:77-87

Clinical Epidemiology / Health Services

- EHD-based adherence is a key indicator in routine care



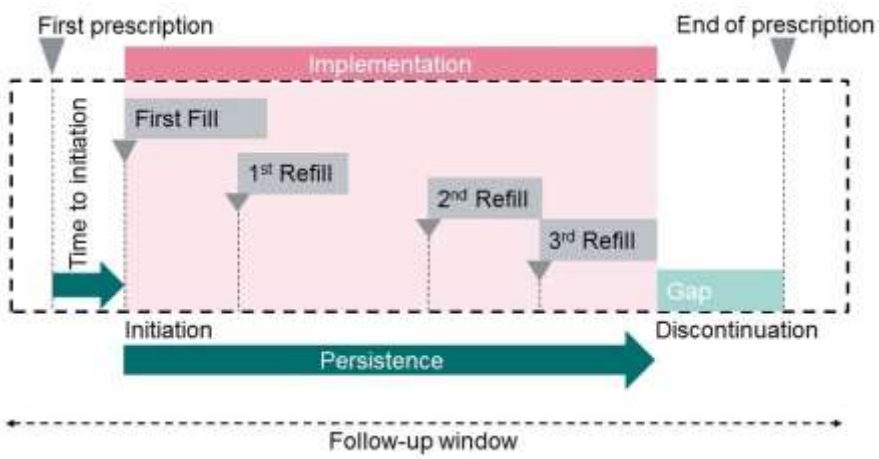
15

Method-related variation in adherence to sibutramine

Measure	Formula	Value	Result (Standard Deviation)
CMA[17]	cumulative days' supply of medication obtained/total days to next fill or to end of observation period	adherence value for cumulative time period	0.635 (0.29)
CMG[17]	total days of treatment gaps/total days to next fill or end of observation period	nonadherence value for cumulative period, winsorized at zero	0.370 (0.28)
CMOS[11]	total days of treatment gaps (+) or surplus[a] (–)/total days in observation period	nonadherence value for cumulative period, allowing for surplus	0.365 (0.29)
CR[24]	(total days supplied – last days' supply)/(last claim date – first claim date) × 100	adherence value for period between fills	84.4% (0.22)[b]
CSA[17]	days' supply obtained at beginning of interval/days in interval	adherence value for interval of study participation	1.097 (1.73)
DBR[22]	1 – {[(last claim date – first claim date) – total days' supply]/(last claim date – first claim date)} × 100	overall adherence percentage	104.8% (38.6)
MPR[16]	days' supply: days in period	ratio of medication available	0.635:1 (0.29)
MPRm[26]	[total days supplied/(last claim date – first claim date + last days' supply)] × 100	adherence percentage, adjusted to include final refill period	86.6% (16.6)
MRA[23]	(total days' supply/total number of days evaluated) × 100	overall adherence percentage	83.5% (29.1)
PDC[17]	(total days supply/total number of days evaluated) × 100%, capped at 1.0[a]	percentage of days with medication available	63.0% (28.3)
RCR[25]	[(sum of quantity dispensed over interval/quantity to be taken per day) × 100]/number of days in interval between first and last refill	overall adherence percentage	104.8% (38.6)

LM Hess - 2006. https://doi.org/10.1345/aph.1H018.

Adherence Taxonomy



Adapted from Vrijens, Bernard, et al. *British Journal of Clinical Pharmacology* 73, Nr. 5 (Mai 2012): 691–705. https://doi.org/10.1111/j.1365-2125.2012.04167.x.

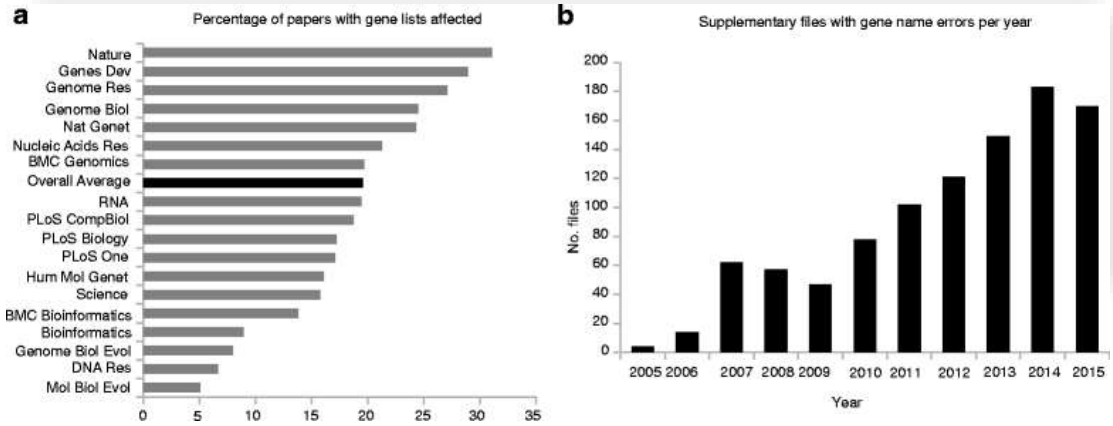
Challenge in Adherence Big Data research

- Too much data to easily comprehend
- Too much variation & variability to easily standardize
- Too complex to easily visualize and communicate

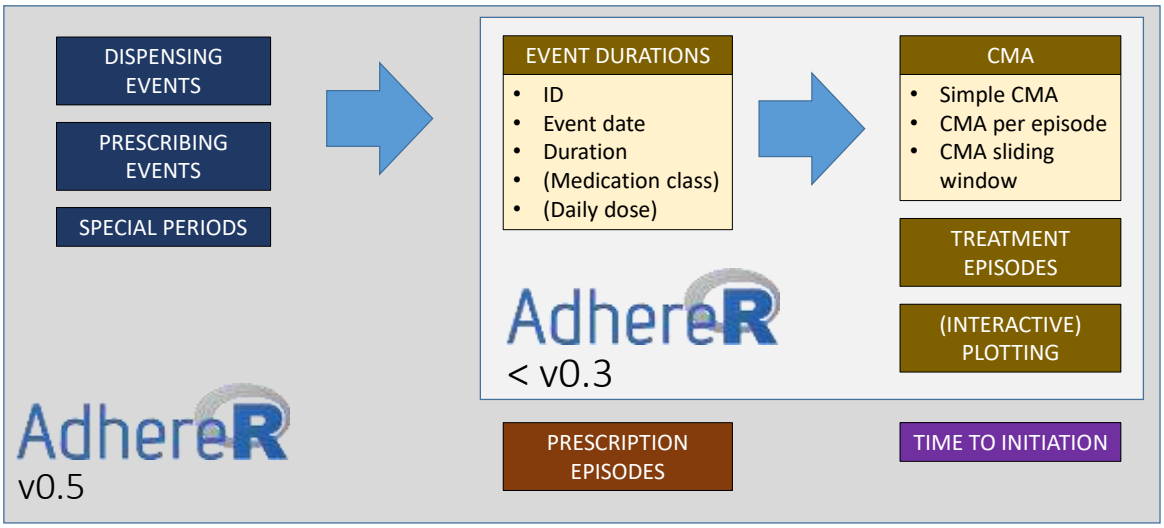
What tools do you use?

https://twitter.com/statsepi/status/795574223439876100?s=09

The challenges of using Excel

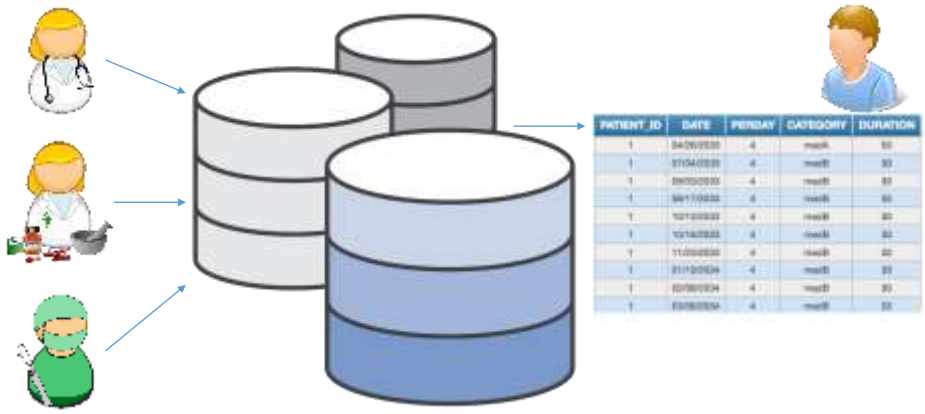


Ziemann, Mark, Yotam Eren, und Assam El-Osta. *Genome biology* 17, Nr. 1 (2016): 177.

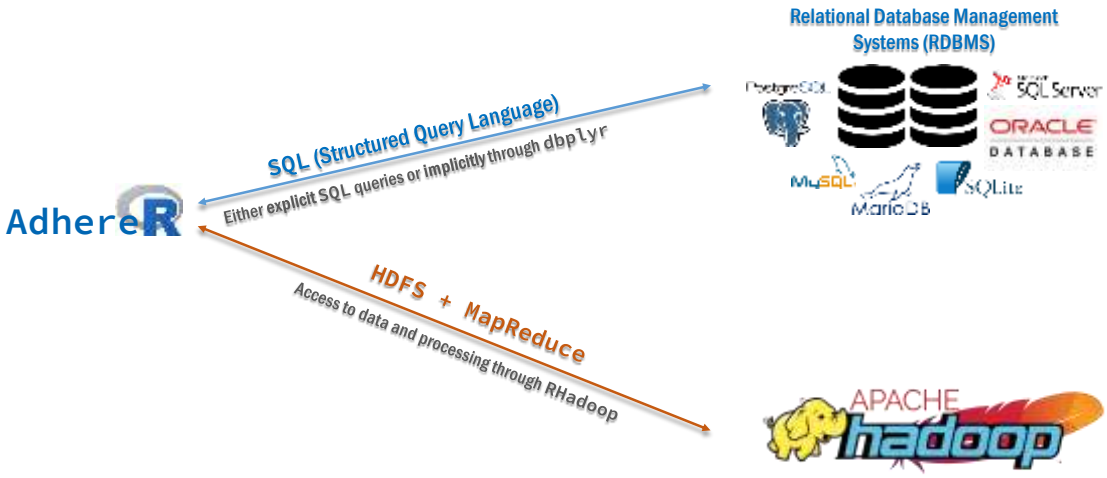


Dima, A, S Allemann, und D Dediu. „AdhereR: An Open Science Approach to Estimating Adherence to Medications Using Electronic Healthcare Databases." *Studies in health technology and informatics* 264 (2019): 1451–52.

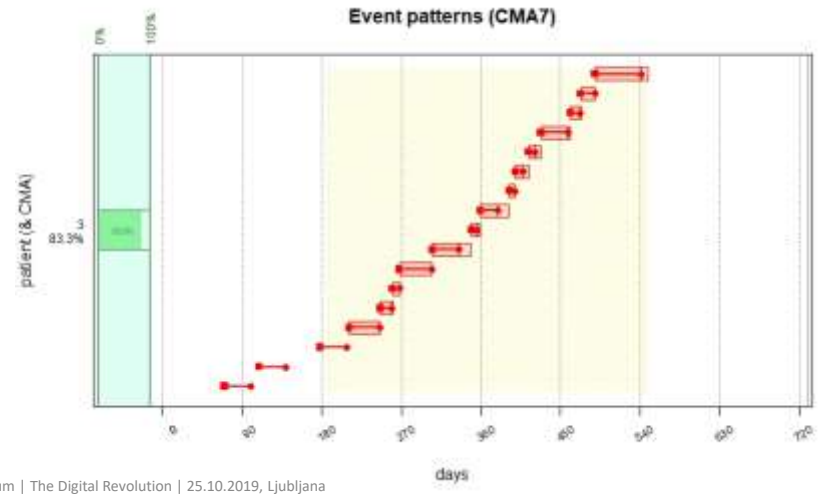
Step 1: Acquire Data



AdhereR & databases: an overview



Step 2: Extract Information



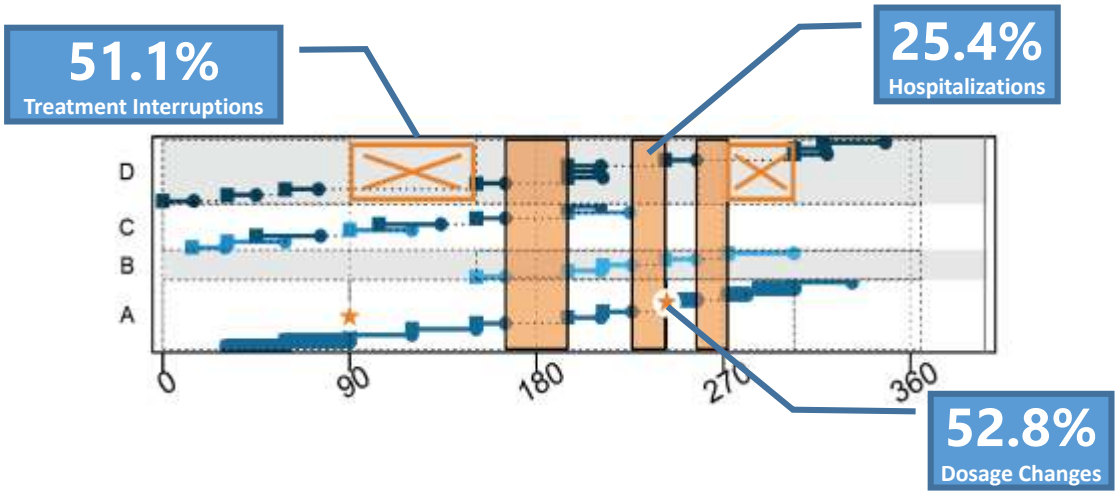
Combine multiple data sources

Functions to:

1. automatically select the last prescribed dose to calculate supply duration,
2. check for prescription changes, hospitalizations, and other treatment interruptions during this period,
3. adjust the supply duration based on prescription changes and hospitalizations

Samuel S. Allemann, Dan Dediu, Alexandra L Dima. Data preparation for computing adherence to medication in AdhereR. 2019 [cited 2019 September 09]. Available from: https://ddediu.github.io/AdhereR/compute_event_durations/compute_event_durations.html

Patients with Cystic Fibrosis



Allemann S. 11th PCNE Working Conference in Egmond aan Zee (Netherlands), 6-9 Feb 2019. Rouzé H. Patient Preference and Adherence. 2019;2019(13):1497—1510.

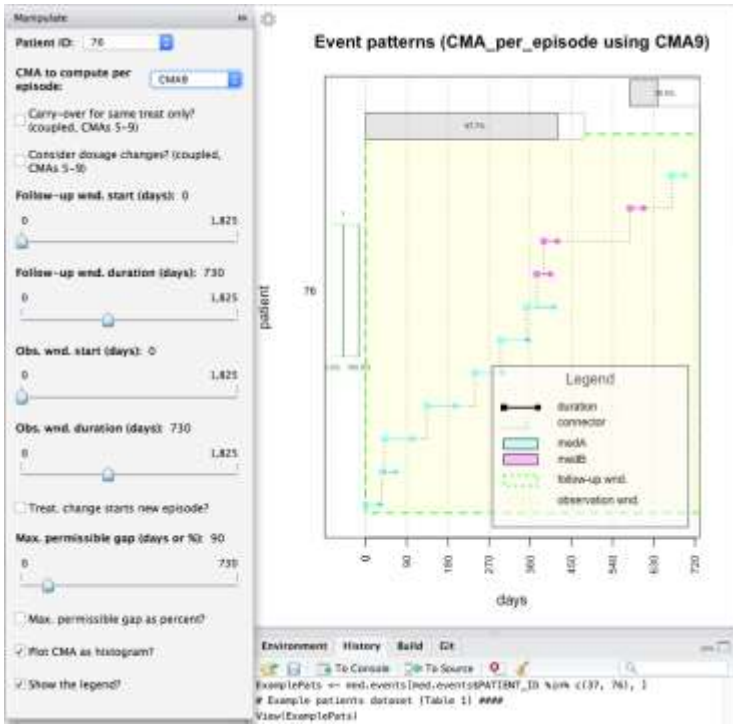
Interactive Visualization

Medication history per patient

→ validation of algorithms

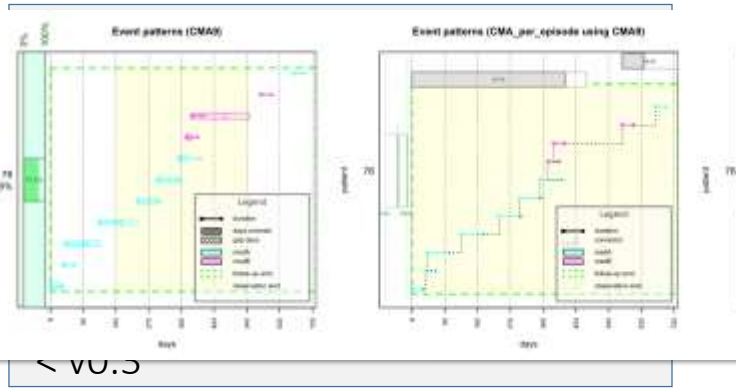
→ identification of unusual patterns

→ hypothesis generation

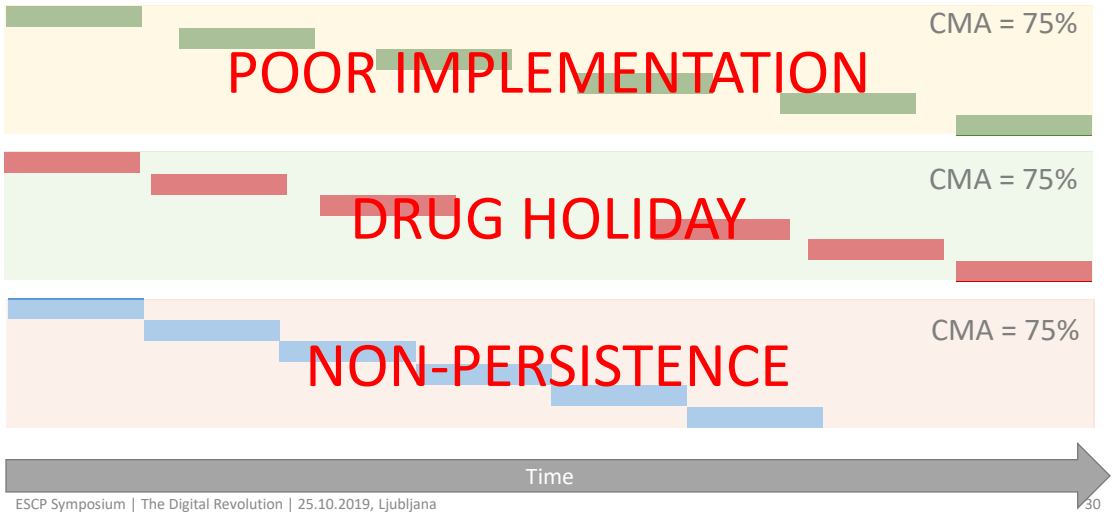


Step 3: Analyse & Visualize

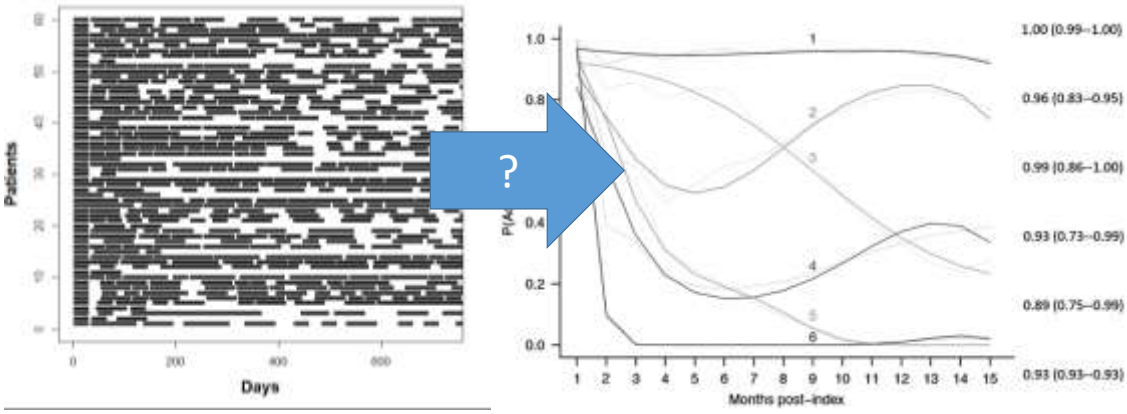
PATIENT_ID	DATE	PERDAY	CATEGORY	DURATION
76	12/13/2035	20	medA	30
76	01/18/2036	20	medA	30
76	01/23/2036	2	medA	60
76	04/25/2036	2	medA	60
76	08/08/2036	2	medA	60
76	10/03/2036	2	medA	60
76	11/29/2036	2	medA	60
76	12/21/2036	6	medB	30
76	01/05/2037	6	medB	30
76	07/13/2037	6	medB	30
76	10/11/2037	2	medA	30



Step 4: Decide & Act

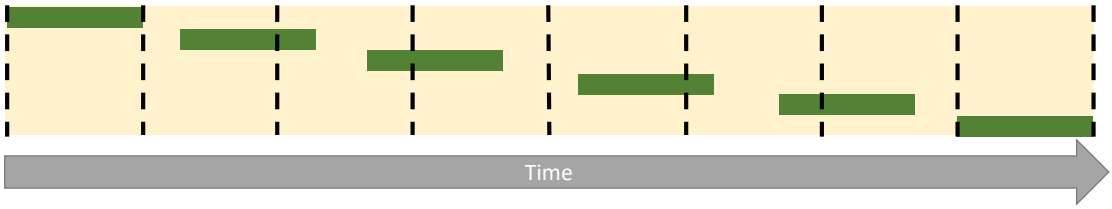


Adherence Trajectories



Franklin, Shrank, Pakes, et al. *Medical Care* 51, no. 9 (September 2013): 789–96.

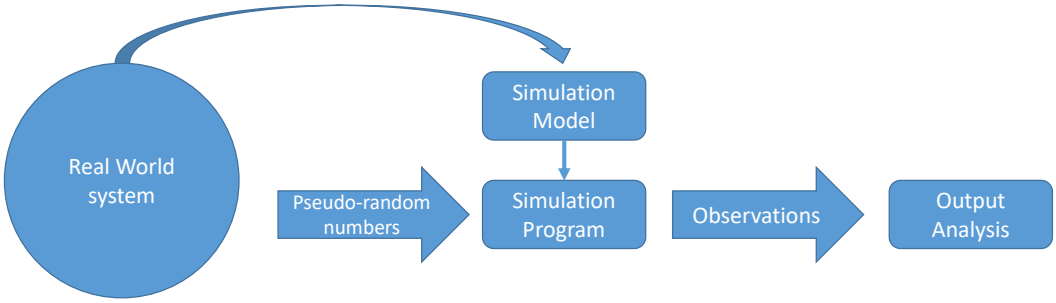
Adherence estimates for sliding windows



- Which adherence estimation within windows?
- Which window size?
- Which overlap between windows?

Simulation study

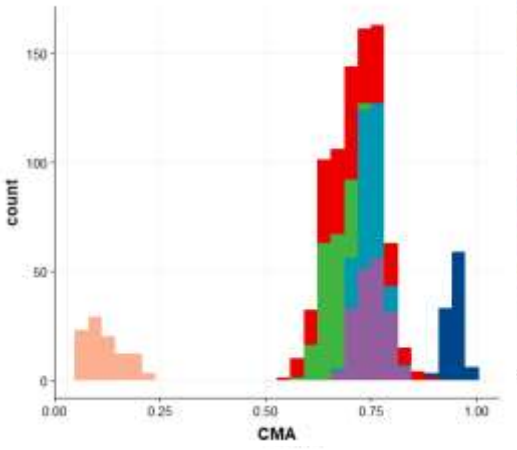
- assess the performance of a variety of methods and parameters in relation to a known state



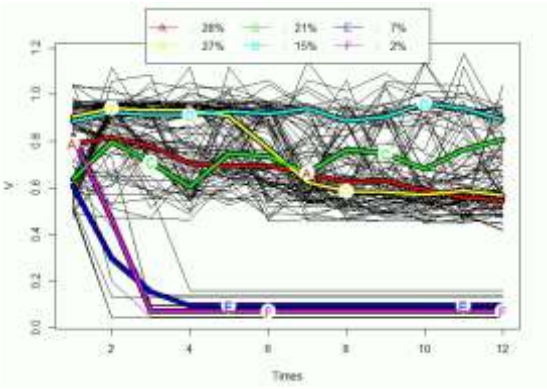
Burton, Altman, Royston, et al. *Statistics in Medicine* 25, no. 24 (December 30, 2006): 4279–92.

Simulated refill patterns

- **Group 1:** *"High adherence"*
- **Group 2:** *"Erratic adherence"*
- **Group 3:** *"Gradual decline"*
- **Group 4:** *"Intermittent adherence"*
- **Group 5:** *"Partial drop-off"*
- **Group 6:** *"Non-persistence"*



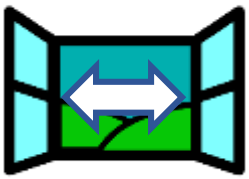
Clustering trajectories with *kml*



- K-means clustering for longitudinal data
  - No prior information about groups required
  - clustering of trajectories that do not follow polynomial or parametric functions
- Euclidean distance
- 20 re-rolls with different starting conditions

Genolini and Falissard. *Computer Methods and Programs in Biomedicine* 104, no. 3 (December 1, 2011): e112–21.

Simulation Scenarios



- **Window Size**
  - 7, 14, and each multipl… …il 720 days (observation perio…

- **Overlap**
  - 0-90% o… …ze (10%-intervals)

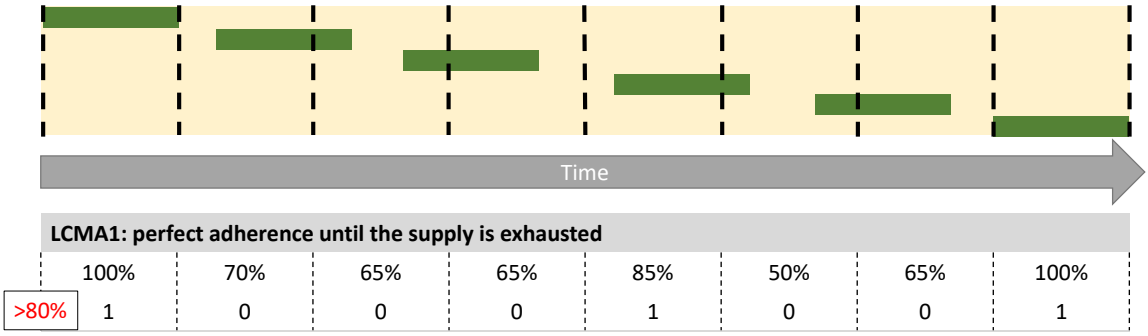
> 200 Mio patient-years

Performance analysis

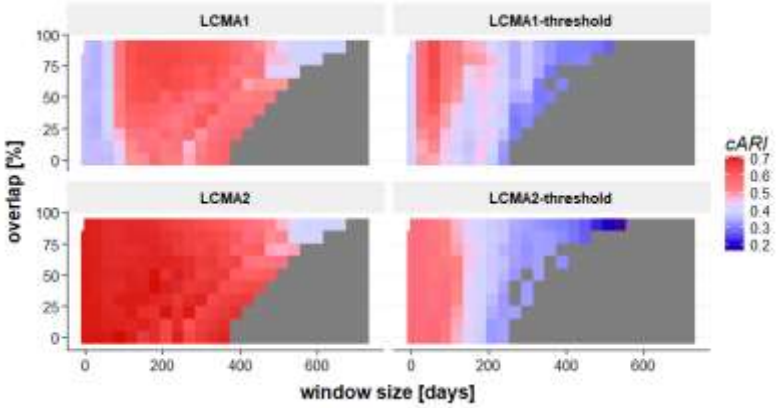
- Baseline comparison: simple k-means clustering with average CMA
- Measurement of agreement between pre-assigned group and identified cluster
- Adjusted Rand Index: value between 0 and 1
  - 0: not better than chance
  - 1: perfect agreement with pre-specified group allocation

Hubert, and Arabie. *Journal of Classification* 2, no. 1 (1985): 193–218.

Adherence estimates for sliding windows

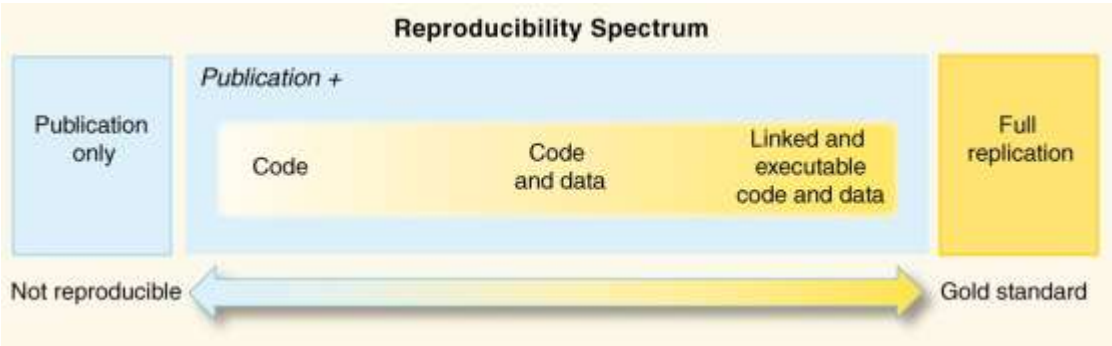


Impact of window size and overlap on classification accuracy



Allemann et al. *Frontiers in pharmacology* 10 (2019): 383.

Open Science /



Roger D. Peng Science 2011;334:1226-1227

Conclusions

AdhereR

- Acquires data from various, distributed sources
- Supports aggregation and information extraction from different data sources
- Allows interactive calculations & visualization
- Supports decisions on calculation methods

Tool for reproducible (big data) research on adherence to medications!

Alexandra Dima

Dan Dediu

Samuel Allemann
s.allemann@unibas.ch

Thank you!

AdhereR

www.adherer.eu

University of Basel

FNSNF Fonds National Suisse Schweizerischer Nationalfonds Fondo Nazionale Svizzero Swiss National Science Foundation

UNIVERSITÉ DE LYON

Lyon 1

université lumière lyon 2