Supplemental material to: Analysis of the Term 'Big Data': Usage in Biomedical Publications

A.J. van Altena, P.D. Moerland, A.H. Zwinderman, S.D. Olabarriaga
Department of Epidemiology, Biostatistics and Bioinformatics
Academic Medical Centre of the University of Amsterdam
Email: {a.j.vanaltena, s.d.olabarriaga}@amc.uva.nl

I. APPENDIX A

The six identified biomedical papers that do not use the term "Big Data" in their title, abstract, or keywords are: [1, 2, 3, 4, 5, 6].

II. APPENDIX B

A. *Visual inspection of classification results*

The following documents were selected for inspection, presented here in decreasing order of classification strength:

- TP: [7, 8, 9, 10, 11, 12, 13, 14, 15, 16];
- FN: [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33];
- FP: [34, 35, 36, 37, 38, 39, 40, 41, 42, 43].

Analysis of TP.

Upon reading of the original title and abstract of the TP documents, the following Big Data topics were identified by consensus:

- [7]: Volume, Velocity, Value, Methods, Technology, Impact.
- [8]: Volume, Variety, Veracity, Value, Methods, Technology, Impact;
- [9]: Volume, Technology, Beyond conventional;
- [10]: Volume, Variety, Value, Method, Impact;
- [11]: Volume, Variety, Methods, Technology;
- [12]: Volume, Value;
- [13]: Volume, Variety, Value, Impact;
- [14]: Volume, Velocity, Variety, Value, Methods, Impact;
- [15]: Veracity, Value; and
- [16]: Veracity, Value, Impact.

Analysis of FP.

- [34]: Methods, Variety;
- [35]: Methods, Technology;
- [38]: Variety, Value, Methods;
- [39]: Volume, Value, Methods;
- [40]: Volume, Value, Impact;
- [41]: Variety, Value, Methods, Impact; and
- [42]: Volume, Technology.
- [43]: Value;

REFERENCES

[1] M. Yazdani et al. 'Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 1272–1280. DOI: 10.1109/BigData.2016.7840731.

[2] M. J. Anderson et al. 'Enabling factor analysis on thousand-subject neuroimaging datasets'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 1151–1160. DOI: 10.1109/BigData.2016.7840719.

[3] N. Shah, H. Shah, M. Malensek, S. L. Pallickara and S. Pallickara. 'Network analysis for identifying and characterizing disease outbreak influence from voluminous epidemiology data'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 1222–1231. DOI: 10.1109/BigData.2016.7840726.

[4] T. Yoshida, M. E. Celebi, G. Schaefer and H. Iyatomi. 'Simple and effective pre-processing for automated melanoma discrimination based on cytological findings'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 3439–3442. DOI: 10.1109/BigData.2016.7841005.

[5] J. Wang et al. 'Wearable sensor based human posture recognition'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 3432–3438. DOI: 10.1109/BigData.2016.7841004.

[6] M. K. Lodhi et al. 'A framework to predict outcome for cancer patients using data from a nursing EHR'. In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 3387–3395. DOI: 10.1109/BigData.2016.7840999.

[7] J. Cole, S Newman, F Foertter, I Aguilar and M Coffey. 'Breeding and genetics symposium: Really big data: Processing and analysis of very large data sets'. In: *Journal of animal science* 90.3 (2012), pp. 723–733.

[8] D. U. Pfeiffer and K. B. Stevens. 'Spatial and temporal epidemiological analysis in the Big Data era'. In: *Preventive veterinary medicine* 122.1 (2015), pp. 213–220.

[9] P. Agarwal and K. Owzar. 'Next generation distributed computing for cancer research'. In: *Cancer informatics* 13.Suppl 7 (2014), p. 97.

[10] W. Zhao, W. Zou and J. J. Chen. 'Topic modeling for cluster analysis of large biological and medical datasets'. In: *BMC bioinformatics* 15.11 (2014), S11.

[11] E. A. Mohammed, B. H. Far and C. Naugler. 'Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends'. In: *BioData mining* 7.1 (2014), p. 22.

[12] W. Zhao et al. 'A novel procedure on next generation sequencing data analysis using text mining algorithm'. In: *BMC bioinformatics* 17.1 (2016), p. 213.

[13] B. H. Shirts et al. 'Clinical laboratory analytics: Challenges and promise for an emerging discipline'. In: *Journal of pathology informatics* 6 (2015).

[14] I. El Naqa. 'Perspectives on making big data analytics work for oncology'. In: *Methods* 111 (2016), pp. 32–44.

[15] S. R. Sukumar, R. Natarajan and R. K. Ferrell. 'Quality of Big Data in health care'. In: *International journal of health care quality assurance* 28.6 (2015), pp. 621–634.

[16] A. T. Janke, D. L. Overbeek, K. E. Kocher and P. D. Levy. 'Exploring the potential of predictive analytics and big data in emergency care'. In: *Annals of emergency medicine* 67.2 (2016), pp. 227–236.

[17] J. C. Brand. *Editorial Commentary: Big Data Suggest That Because of a Significant Increased Risk of Postoperative Infection, Steroid Injection Is Not Recommended After Ankle Arthroscopy*. 2016.

[18] D. Glenn. 'Big Data and Pharmacovigilance: The Role of Oncology Nurses.' In: *Clinical journal of oncology nursing* 20.5 (2016), pp. 478–480.

[19] J. K. Wagner, C. Peltz-Rauchman, A. K. Rahm and C. C. Johnson. 'Precision engagement: the PMI's success will depend on more than genomes and big data'. In: *Genetics in Medicine* (2016).

[20] M. Taquet, J. Quoidbach, Y.-A. de Montjoye and M. Desseilles. 'Mapping collective emotions to make sense of collective behavior'. In: *Behavioral and Brain Sciences* 37.1 (2014), pp. 102–103.

[21] N. P. Restifo. 'A big data view of the tumor immunome'. In: *Immunity* 39.4 (2013), pp. 631–632.

[22] M. T. Keane and A. Gerow. 'It's distributions all the way down!: Second order changes in statistical distributions also occur'. In: *Behavioral and Brain Sciences* 37.1 (2014), pp. 87–87.

[23] A. L. Sharp et al. 'In Reply:' in: *Academic Emergency Medicine* 24.3 (2017), pp. 393–393. ISSN: 1553-2712. DOI: 10.1111/acem.13082. URL: http://dx.doi.org/10.1111/acem.13082.

[24] M. L. Lindsey, J. J. Saucerman and K. Y. DeLeon-Pennell. 'Knowledge gaps to understanding cardiac macrophage polarization following myocardial infarction'. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1862.12 (2016), pp. 2288–2292.

[25] Y. Ren et al. 'Assessing the effects of cocaine dependence and pathological gambling using group-wise sparse representation of natural stimulus FMRI data.' In: *Brain imaging and behavior* (2016).

[26] M. L. Berendsen, J. Smits, M. G. Netea and A. van der Ven. 'Non-specific Effects of Vaccines and Stunting: Timing May Be Essential'. In: *EBioMedicine* 8 (2016), pp. 341–348.

[27] R. B. Stricker and L. Johnson. 'Lyme disease: the promise of Big Data, companion diagnostics and precision medicine'. In: *Infection and drug resistance* 9 (2016), p. 215.

[28] J. Avasarala and P. Zachariah. 'Vitamin D deficiency in multiple sclerosis: Should testing and treatment be based on racial background?' In: *Journal of the neurological sciences* 358.1 (2015), pp. 417–418.

[29] Y. Hayashi, M. Saito and T. Yajima. 'Estimation of the Health Status of People in the Vicinity of Pharmacies Using Pharmacy Big Data'. In: *Yakugaku zasshi: Journal of the Pharmaceutical Society of Japan* 136.2 (2016), pp. 265–271.

[30] S. Zhang et al. 'Characterizing and differentiating task-based and resting state fMRI signals via two-stage sparse representations'. In: *Brain imaging and behavior* 10.1 (2016), pp. 21–32.

[31] P. L. Kubben. 'Programming for physicians: A free online course'. In: *Surgical neurology international* 7 (2016).

[32] A. Rezza et al. 'Signaling networks among stem cell precursors, transit-amplifying progenitors, and their niche in developing hair follicles'. In: *Cell reports* 14.12 (2016), pp. 3001–3018.

[33] P. M. Dawson, M. Werkman, E. Brooks-Pollock and M. J. Tildesley. 'Epidemic predictions in an imperfect world: modelling disease spread with partial data'. In: *Proc. R. Soc. B*. Vol. 282. 1808. The Royal Society. 2015, p. 20150205.

[34] A. Cambiaghi, M. Ferrario and M. Masseroli. 'Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration'. In: *Briefings in bioinformatics* 18.3 (2016), pp. 498–510.

[35] B. Hao et al. 'Accelerate Healthcare Data Analytics: An Agile Practice to Perform Collaborative and Reproducible Analyses.' In: *Studies in health technology and informatics* 228 (2016), pp. 552–556.

[36] P. R. Gordon and M. A. Sephton. 'Organic Matter Detection on Mars by Pyrolysis-FTIR: An Analysis of Sensitivity and Mineral Matrix Effects'. In: *Astrobiology* 16.11 (2016), pp. 831–845.

[37] D. Chalmers et al. 'Has the biobank bubble burst? Withstanding the challenges for sustainable biobanking in the digital era'. In: *BMC medical ethics* 17.1 (2016), p. 39.

[38] M. C. Tremblay, G. J. Deckard and R. Klein. 'Health informatics and analyticsbuilding a program to integrate business analytics across clinical and administrative disciplines'. In: *Journal of the American Medical Informatics Association* 23.4 (2016), pp. 824–828.

[39] E. W. Deutsch et al. 'Tiered human integrated sequence search databases for shotgun proteomics'. In: *Journal of proteome research* 15.11 (2016), pp. 4091–4100.

[40] J. H. Moore and J. H. Holmes. 'The golden era of biomedical informatics has begun'. In: *BioData mining* 9.1 (2016), p. 15.

[41] F. Cheng, H. Hong, S. Yang and Y. Wei. 'Individualized network-based drug repositioning infrastructure for precision oncology in the panomics era'. In: *Briefings in bioinformatics* (2016), bbw051.

[42] T. Sherif et al. 'CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research'. In: *Frontiers in neuroinformatics* 8 (2014).

[43] C. Holmes, F. McDonald, M. Jones and J. Graham. 'Knowledge translation: Moving proteomics science to innovation in society'. In: *Omics: a journal of integrative biology* 20.6 (2016), pp. 352–361.