Research Techniques Made Simple: An Introduction to Use and Analysis of Big Data in Dermatology

Mackenzie R. Wehner[1], Katherine A. Levandoski[2], Martin Kulldorff[3] and Maryam M. Asgari[2]

Big data is a term used for any collection of datasets whose size and complexity exceeds the capabilities of traditional data processing applications. Big data repositories, including those for molecular, clinical, and epidemiology data, offer unprecedented research opportunities to help guide scientific advancement. Advantages of big data can include ease and low cost of collection, ability to approach prospectively and retrospectively, utility for hypothesis generation in addition to hypothesis testing, and the promise of precision medicine. Limitations include cost and difficulty of storing and processing data; need for advanced techniques for formatting and analysis; and concerns about accuracy, reliability, and security. We discuss sources of big data and tools for its analysis to help inform the treatment and management of dermatologic diseases.

Journal of Investigative Dermatology (2017) **137,** e153–e158; doi:10.1016/j.jid.2017.04.019

CME Activity Dates: 20 July 2017
Expiration Date: 19 July 2018
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: Maryam Asgari received research grant support from Pfizer, Inc and Valeant Pharmaceuticals. All other authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationships with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:

- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of William Beaumont Hospital and the Society for Investigative Dermatology. William Beaumont Hospital is accredited by the ACCME to provide continuing medical education for physicians.
William Beaumont Hospital designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the Journal of Investigative Dermatology website: http://www.jidonline.org/current. Tests for CME credits may only be submitted online at https://beaumont.cloud-cme.com/RTMS-August17 – click 'CME on Demand' and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

WHAT ARE BIG DATA?

Big data are commonly defined as data so large or complex that traditional data processing and analytic approaches are inadequate. The *3 Vs* that characterize big data are volume (amount of data), velocity (speed at which data are generated and processed), and variety (types of data) (Laney, 2001), all of which have been growing rapidly (Figure 1). Although there is no predefined threshold for volume, in general, anything 1 petabyte (10^{15} bytes, or the approximate size of 1 million human genomes) or greater is considered big data (Figure 2). The ability to monitor, record, and store information from large populations from

[1]*Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA;* [2]*Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA; and* [3]*Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA*

Correspondence: Maryam M. Asgari, Department of Dermatology, Massachusetts General Hospital, 50 Staniford Street, Suite 230A, Boston, Massachusetts 02114, USA. E-mail: harvardskinstudies@partners.org

SUMMARY POINTS

- Big data describes any collection of datasets whose size and complexity exceeds the capabilities of traditional data processing applications.
- Big data has the potential to help inform the treatment and management of dermatologic diseases through improved risk assessment, surveillance, diagnosis, and treatment methods.
- While big data presents spectacular research opportunities, there are important limitations to consider, including storage costs, processing challenges, and concerns about accuracy, reliably, and security.

sources including electronic medical records, insurance claims, surveys, disease registries, biospecimens, apps and social media, the internet, and personal monitoring devices has shepherded the era of big data into use in health care. The volume of health care data in the United States in 2017 is rapidly approaching zettabyte levels (iHT2, 2013). This wealth of structured and unstructured data has the potential to substantially affect health care delivery through improved risk assessment, surveillance, diagnosis, and treatment methods.

WHAT ARE SOME BIG DATA SOURCES IN HEALTH CARE?

There are many big data sources in health care. OptumLabs (https://www.optumlabs.com), an open collaborative research center, provides de-identified clinical data from electronic health records and claims data for over 100 million insured members (Borah, 2016). Sentinel (https://www.sentinelinitiative.org), a US Food and Drug Administration initiative, uses data from electronic health records, insurance claims, and registries to monitor postmarketing, real-world safety of medicines. Sentinel data were used to estimate the validity of *International Classification of Diseases—Ninth Revision* codes (Centers for Disease Control, 1998) for ascertaining Stevens-Johnson syndrome and toxic epidermal necrolysis in 12 collaborating research units, covering almost 60 million people (Davis et al., 2015). UK Biobank and Kaiser Permanente Biobank are examples of medical data and tissue samples collected for research purposes. UK Biobank (www.ukbiobank.ac.uk) is a cohort of 500,000 participants in the UK who have provided baseline information and blood, urine, and saliva samples and who are being followed prospectively through their regular care. The Kaiser Permanente Research Biobank (https://www.dor.kaiser.org/external/DORExternal/rpgeh) is composed of 220,000 health plan members who have contributed genetic and electronic health record data. This was recently used in a large genome-wide association study of cutaneous squamous cell carcinoma, which identified 10 single-nucleotide polymorphisms associated with cutaneous squamous cell carcinoma at genome-wide significance and provided new insights into the genetics of heritable cutaneous squamous cell carcinoma risks (Asgari et al., 2016). For genomic data, such as those found in biobanks, the National Center for Biotechnology Information has developed the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo), which acts as a public archive and repository of microarray, next-generation sequencing, and high-throughput functional genomic data. Geographic information systems, such as the National Cancer Institute Geographic Information Systems and Science for Cancer Control (https://gis.cancer.gov), capture geographic data that allow for mapping of disease trends. Solar UV radiation data are available through this system, and the association between cutaneous melanoma incidence rates and county-level UV exposure has been examined (Richards et al., 2011).

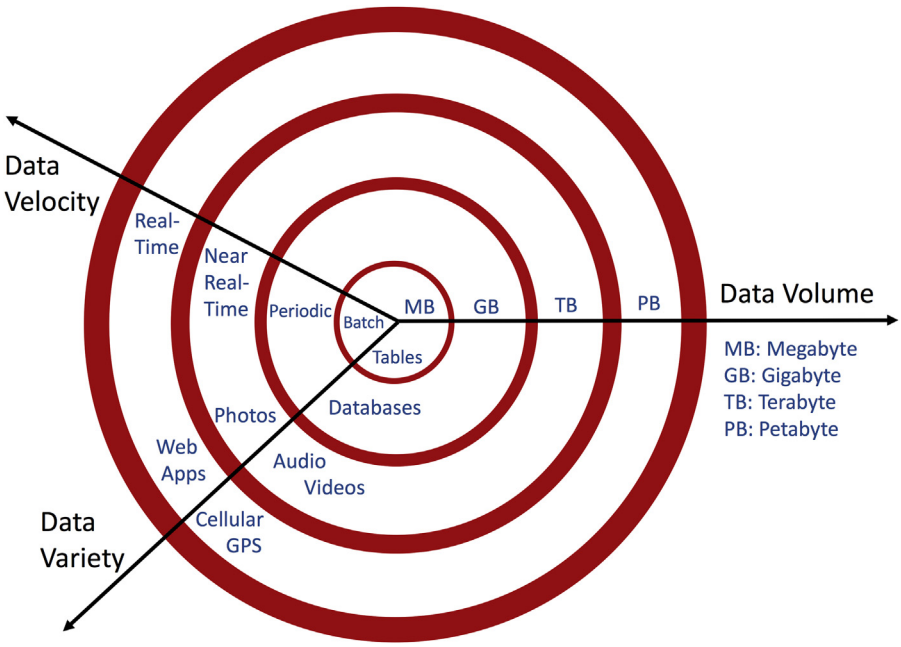

Figure 1. The *3* Vs of big data. The *3* Vs of big data are volume (amount of data), velocity (speed at which data is generated), and variety (number of types of data), all of which have been growing rapidly. After "The 3Vs That Define Big Data," Diya Soubra, Data Science Central, http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data. GPS, global positioning system.

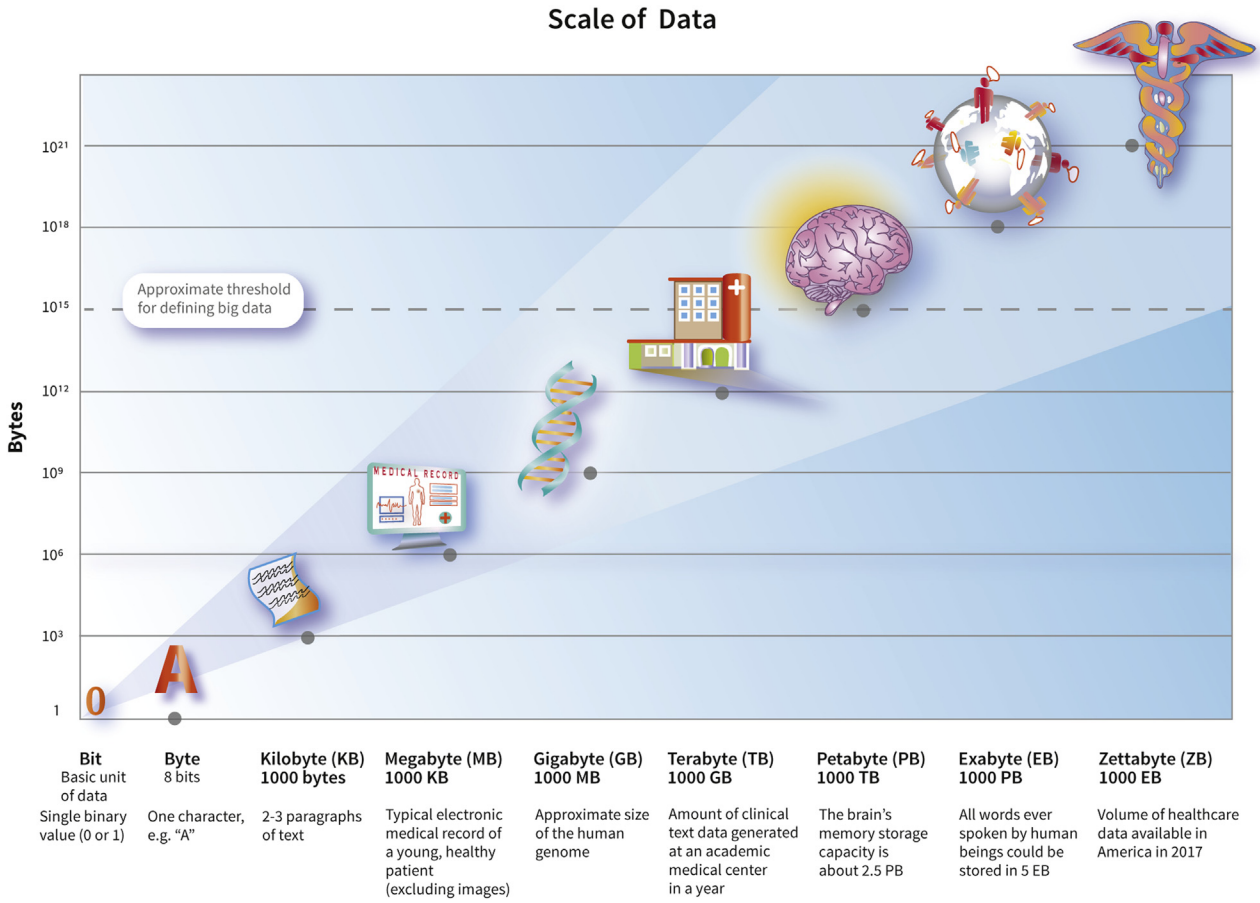

Figure 2. Logarithmic scale depicting volume of big data. The relative scale of different datasets is depicted. There is no predefined threshold for volume that defines big data, but in general, anything one petabyte or greater is considered big data.

Computer-based geographic information systems, web-based geospatial technologies such as global positioning systems in smartphones, and geospatial modeling can be used to follow disease trends and to examine mobility and social networks and their impact on disease (Birch, 2016; Ray et al., 2016).

To enhance the utility of biomedical big data from these diverse sources, the National Institutes of Health established Big Data to Knowledge (https://datascience.nih.gov/bd2k). It aims to make digital data "findable, accessible, interoperable, and reusable (FAIR)," with the following specific goals: (i) to improve the ability to find and use big data, (ii) to develop analysis tools for big data, (iii) to increase training in data science, and (iv) to establish centers of excellence in data science (Margolis et al., 2014). Big Data to Knowledge has funding opportunities in many areas, including curating, coordinating, and organizing big data, developing big data educational curricula, and improving big data standards (https://www.nlm.nih.gov/ep/BD2KGrants.html).

HOW DO ANALYTIC TECHNIQUES FOR BIG DATA DIFFER FROM THOSE FOR TRADITIONAL DATA?

Although big data can be used for traditional hypothesis testing and can be especially valuable for research on rare diseases or exposures, big data analyses are often hypothesis generating. Rather than test a hypothesis, they can provide evidence for new hypotheses that can later be tested with traditional techniques. Big data analyses often center on identifying patterns. Unlike traditional predictive modeling based on a small number of covariates, big data predictive modeling often involves variables that are not preselected. Thus, compared with traditional data analysis, big data analysis has the potential to be more exploratory. Given the multiplicity inherent in the many potential patterns evaluated, such big data analyses benefit from special statistical methods that account for this multiple testing using *P*-value adjustments or false discovery rates.

ANALYTIC TECHNIQUES FOR BIG DATA

There are many computational and statistical methods used to analyze big data. *Data mining* is a process through which data are analyzed from different perspectives to identify unsuspected patterns. Using insurance claims, data mining with TreeScan software was used to explore unsuspected adverse reactions associated with antifungal drug exposure (Kulldorff et al., 2013). TreeScan is free data mining software available for download online (TreeScan, Boston, MA; https://www.treescan.org). *Cluster analysis* focuses on grouping similar patients or observations by demographics, medical history, genetics, or geography. For example, the spatial scan statistic was used to detect geographic clusters of basal cell carcinomas in a Northern California population with the goal of targeting screening and prevention efforts

(Ray et al., 2016). Another example is cluster analysis of different quality-of-life scoring systems in psoriasis patients, which showed lack of correlation of disease severity with psychological distress instruments (Sampogna et al., 2004).

Machine learning allows algorithms to learn from a training dataset to make predictive models without specifying the model in advance. Machine learning is currently being explored to track pigmented lesions over time and identify lesions at higher risk for malignancy (Li et al., 2016). Machine learning was recently used to develop a diagnosis algorithm for skin cancer based on clinical images (Esteva et al., 2017). The algorithm, which uses only pixels and disease labels as inputs, matches the performance of dermatologists in identifying cancerous and noncancerous lesions (Esteva et al., 2017). Deployable on mobile devices, machine learning algorithms that train computers to make reliable diagnoses directly from clinical images hold the potential to make a significant clinical impact by extending the reach of dermatologists beyond the clinic (Esteva et al., 2017). *Decision tree learning* is a type of machine learning in which the independent variables are used to create a hierarchical tree structure with leaves and branches, which can predict an outcome (see Figure 3 for example). There are two main types of decision tree analyses: *classification tree analysis,* where the predicted outcome is dichotomous such as for melanoma mortality, and *regression tree analysis,* where the predicted outcome is a continuous variable such as age at melanoma diagnosis. Both classification and regression tree analyses were used to identify histological features of melanoma associated with CDKN2A germline mutations (Sargen et al., 2015). *Bayesian networks* are another type of machine learning that use probabilistic graphs to explore relationships between, for example, symptoms and disease, to be used in clinical decision making or diagnosis. *Cognitive computing* is a type of machine learning that tries to mimic the functioning of the human brain. Natural language processing algorithms allow computers to extract useful information from text, such as electronic health records, well enough to yield meaningful data. Such algorithms can identify mentions of a risk factor or of an outcome disease in clinic notes, recognizing that the same exposure or diagnosis can be expressed in many different ways and with potential misspellings and distinguishing a positive diagnosis from a rule-out diagnosis. Natural language processing has been used in dermatology research to find nonmelanoma skin cancer diagnoses in electronic pathology reports (Eide et al., 2012).

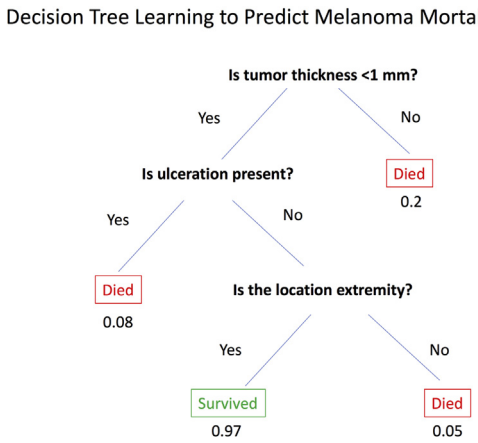

Figure 3. Decision-tree learning to predict melanoma mortality (hypothetical). Hypothetical example illustrating the utility of decision-tree learning for melanoma mortality prediction showing "leaves" (independent variables) such as tumor thickness, ulceration, and tumor location, and probability of survival (outcome).

ANALYTIC PLATFORMS FOR BIG DATA

There are two approaches to analytic platforms for big data: (i) a divide-and-conquer approach (distributed data) and (ii) a centralized approach using a platform that provides both database storage and analytics in a centralized fashion, such as SAP HANA (SAP, Walldorf, Germany; http://www.sap.com/product/technology-platform/hana.html). SAP HANA is a computing platform that offers tools for storing, managing, and analyzing big data. When big data are in different physical locations, distributed data analysis can be used with some of the analysis conducted locally on the complete data while the final analysis occurs centrally using summary data from each site. The advantage of distributed data for medical information is that data remain at local sites, minimizing storage costs and maximizing data integrity and patient privacy.

SUMMARY AND FUTURE DIRECTIONS IN DERMATOLOGY

The term *big data* is more than just very large data or a large number of data sources but encompasses a new approach to complex data. It offers a new, hypothesis-generating framework to conduct research and requires novel analysis methods. It has significant advantages but also has limitations (Table 1), and traditional data analytics are still

Table 1. Advantages and limitations of big data

Advantages	Limitations
• Large sample size • Data can be inexpensive to collect and acquire: in many cases the data have already been collected through routine clinical care (electronic health records) or through the participants themselves (internet searches or personal monitoring devices) • Both retrospective and prospective approaches are often available • Multiple data points from different sources can be combined, leveraging the advantages of different collection sources or smaller datasets	• Storage: datasets can require considerable resources to store • Formatting and data cleaning: advanced computer science can be required before the data is analyzable • Quality control: can be difficult and often has to be done through small representative samples • Security and privacy concerns: often more complex than for traditional datasets • Accuracy and consistency of methods: many approaches are relat ively new and imperfect, although these may continue to improve over time

MULTIPLE CHOICE QUESTIONS

1. What are the *3 Vs* that characterize big data?
   a. Value, viability, and variety
   b. Volume, velocity, and viability
   c. Volume, velocity, and variety
   d. Volume, value, and variety

2. What distinguishes big data analyses from traditional data analyses?
   a. They can be used to both test and generate hypotheses.
   b. Variables are often not preselected for prediction modeling.
   c. They often center around identifying and evaluating patterns.
   d. All of the above

3. What analytic technique focuses on grouping similar patients by characteristics such as demographics, genetics, or geography and can be used to inform geographically targeted screening and prevention efforts?
   a. Cluster analysis
   b. Decision-tree learning
   c. Bayesian networks
   d. Cognitive computing

4. Which of the following is NOT a limitation of big data?
   a. Storage may require considerable resources.
   b. Formatting and analysis may require advanced computer science.
   c. Big data can be used only for retrospective analyses.
   d. Big data have more complex security and information privacy concerns than traditional datasets.

5. Which of the following is NOT a potential application of big data?
   a. Improve risk prediction for very rare diseases
   b. Identify distinct disease phenotypes in heterogeneous diseases that may merit different therapies
   c. Identify causal associations
   d. Perform drug and medical device surveillance

crucially important. In dermatology, big data can be used to improve risk prediction models, support targeted screening for high-risk individuals (e.g., targeted skin cancer screening), optimize management of a variety of skin diseases, and offer clinical decision support (e.g., assistance in deciding whether to biopsy a pigmented lesion). We can further investigate the genetics of skin disease (e.g., genome-wide association studies) (Asgari et al., 2016; Frelinger, 2015) and examine distinct disease phenotypes within heterogeneous diseases that could benefit from tailored therapies (e.g., in psoriasis or eczema). Big data may be an excellent way to perform surveillance and evaluate safety of medications and devices, especially for rarer outcomes. Big data in dermatology present spectacular opportunities, allowing researchers to maximize the potential of existing data sources and opening up new, efficient, and powerful methods for future research.

CONFLICT OF INTEREST

MA has received research funding to her institution from Pfizer, Inc. and Valeant Pharmaceuticals, but these associations have not influenced our work on this article. The authors have no other potential conflicts of interest to disclose.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants R01CA166672 (MA) and K24AR069760 (MA). We would like to acknowledge Susan Gruber for her assistance with reviewing the content of this manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Asgari MM, Wang W, Ioannidis NM, Itnyre J, Hoffmann T, Jorgenson E, et al. Identification of susceptibility loci for cutaneous squamous cell carcinoma. J Invest Dermatol 2016;136:930–7.

Birch P. Powering geospatial analysis: public geo datasets now on Google Cloud, https://cloudplatform.googleblog.com/2016/10/powering-geospatial-analysis-public-geo-datasets-now-on-Google-Cloud.html; 2016 (accessed 6 January 2017).

Borah BJ. Optum Labs overview, https://http://www.allianceforclinicaltrialsinoncology.org/main/cmsfile?cmsPath=/Public/Annual Meeting/files/Prevention-Optum Labs Overview.pdf; 2016 (accessed 14 December 2016).

Centers for Disease Control. International Classification of Diseases–Ninth Revision. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD-9/ucod.txt. Published April 9, 1998. Accessed 22 June 2017.

Davis RL, Gallagher MA, Asgari MM, Eide MJ, Margolis DJ, Macy E, et al. Identification of Stevens-Johnson syndrome and toxic epidermal necrolysis in electronic health record databases. Pharmacoepidemiol Drug Safe 2015;24:684–92.

Eide MJ, Tuthill JM, Krajenta RJ, Jacobsen GR, Levine M, Johnson CC. Validation of claims data algorithms to identify nonmelanoma skin cancer. J Invest Dermatol 2012;132:2005–9.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

Frelinger JA. Big data, big opportunities, and big challenges. J Investig Dermatol Symp Proc 2015;17:33–5.

iHT2. Transforming health care through big data, http://c4fd63cb482ce6861463-bc6183f1c18e748a49b87a25911a0555.r93.cf2.rackcdn.com/iHT2_BigData_2013.pdf; 2013 (accessed 14 December 2016).

Kulldorff M, Dashevsky I, Avery TR, Chan AK, Davis RL, Graham D, et al. Drug safety data mining with a tree-based scan statistic. Pharmacoepidemiol Drug Saf 2013;22:517–23.

Laney D. 3D data management: controlling data volume, variety and velocity. Application Delivery Strategies 2001;6 Feb:949.

Li Y, Esteva A, Kuprel B, Novoa R, Ko J, Thrun S. Skin cancer detection and tracking using data synthesis and deep learning. arXiv 2016:161201074.

Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 2014;21:957–8.

Ray GT, Kulldorff M, Asgari MM. Geographic clusters of basal cell carcinoma in a northern California health plan population. JAMA Dermatol 2016;152:1218–24.

Richards TB, Johnson CJ, Tatalovich Z, Cockburn M, Eide MJ, Henry KA, et al. Association between cutaneous melanoma incidence rates among white US residents and county-level estimates of solar ultraviolet exposure. J Am Acad Dermatol 2011;65:S50–7.

Sampogna F, Sera F, Abeni D. IDI Multipurpose Psoriasis Research on Vital Experiences (IMPROVE) Investigators. Measures of clinical severity, quality of life, and psychological distress in patients with psoriasis: a cluster analysis. J Invest Dermatol 2004;122:602–7.

Sargen MR, Kanetsky PA, Newton-Bishop J, Hayward NK, Mann GJ, Gruis NA, et al. Histologic features of melanoma associated with CDKN2A genotype. J Am Acad Dermatol 2015;72:496–507.e7.