Potential pitfalls in the use of real-world data for studying long COVID

Harrison G Zhang, Jacqueline P Honerlaw, Monika Maripuri, Malarkodi Jebathilagam Samayamuthu, Brendin R Beaulieu-Jones, Huma S Baig, Sehi L'yi, Yuk-Lam Ho, Michele Morris, Vidul Ayakulangara Panickan, et al.

Potential pitfalls in the use of real world data to study Long

HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2023 May 24.

Published in final edited form as:

Nat Med. 2023 May ; 29(5): 1040–1043. doi:10.1038/s41591-023-02274-y.

†Corresponding author: Gabriel A Brat, MD, MPH. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States.

*These authors contributed equally

The Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

James R Aaron MHA[39], Giuseppe Agapito PhD[40], Adem Albayrak[41], Giuseppe Albi MS[28], Mario Alessiani MD, FACS[42], Anna Alloni PhD[37], Danilo F Amendola MSc[43], François Angoulvant MD,PhD[44], Li L.L.J Anthony[45], Bruce J Aronow PhD[32], Fatima Ashraf MS[46], Andrew Atz MD[47], Paul Avillach MD, PhD[1], Paula S Azevedo MD, PhD[48], James Balshi[49], Brett K Beaulieu-Jones PhD[1], Douglas S Bell[23], Antonio Bellasi MD, PhD[50], Riccardo Bellazzi MS, PhD[28], Vincent Benoit PhD[24], Michele Beraghi MS[51], José Luis Bernal-Sobrino MS[52], Mélodie Bernaux[53], Romain Bey[24], Surbhi Bhatnagar PhD[32], Alvar Blanco-Martínez MS[52], Clara-Lea Bonzel MSc[1], John Booth MSc[54], Silvano Bosari Prof.[35], Florence T Bourgeois MD, MPH[11], Robert L Bradford[55], Gabriel A Brat MD[1], Stéphane Bréant[56], Nicholas W Brown MEng[1], Raffaele Bruno MD[57], William A Bryant PhD[54], Mauro Bucalo MS[37], Emily Bucholz MD, PhD, MPH[58], Anita Burgun[59], Tianxi Cai ScD[1], Mario Cannataro M.Sc.[60], Aldo Carmona[61], Charlotte Caucheteux[62], Julien Champ[63], Jin Chen PhD[64], Krista Y Chen BS[65], Luca Chiovato MD, PhD[31], Lorenzo Chiudinelli PhD[66], Kelly Cho PhD, MPH[29], James J Cimino MD[67], Tiago K Colicchio PhD, MBA[67], Sylvie Cormont[56], Sébastien Cossin[30], Jean B Craig PhD[68], Juan Luis Cruz-Bermúdez PhD[52], Jaime Cruz-Rojo MD[52], Arianna Dagliati MS, PhD[2], Mohamad Daniar MSIS[69], Christel Daniel[70], Priyam Das PhD[1], Batsal Devkota[71], Audrey Dionne MD[58], Rui Duan PhD[3], Julien Dubiel[56], Scott L DuVall PhD[72], Loic Esteve[73], Hossein Estiri PhD[15], Shirley Fan[74], Robert W Follett BS[23], Thomas Ganslandt MD[75], Noelia García-Barrio MS[52], Lana X Garmire PhD[76], Nils Gehlenborg[1], Emily J Getzen MS[77], Alon Geva MD, MPH[78], Tobias Gradinger MD, BSc[75], Alexandre Gramfort[62], Romain Griffier[30], Nicolas Griffon[70], Olivier Grisel[62], Alba Gutiérrez-Sacristán PhD[1], Larry Han PhD[3], David A Hanauer MD, MS[8], Christian Haverkamp MD[79], Derek Y Hazard MSc[80], Bing He PhD[76], Darren W Henderson BS[39], Martin Hilka[56], Yuk-Lam Ho MPH[22], John H Holmes MS, PhD[9,10], Chuan Hong PhD[1,81], Kenneth M Huling HS[1], Meghan R Hutch BS[82], Richard W Issitt DClinP[54], Anne Sophie Jannot[83], Vianney Jouhet MD,PhD[30], Ramakanth Kavuluru PhD[25], Mark S Keller[1], Chris J Kennedy PhD[84], Daniel A Key BEng[54], Katie Kirchoff MSHI[85], Jeffrey G Klann MEng, PhD[15], Isaac S Kohane MD, PhD[1], Ian D Krantz[86], Detlef Kraska Dr.[87], Ashok K Krishnamurthy PhD[88], Sehi L'Yi PhD[1], Trang T Le PhD[9], Judith Leblanc[89], Guillaume Lemaitre[62], Leslie Lenert MD, MS[68], Damien Leprovost[90], Molei Liu PhD[91], Ne Hooi Will Loh MBBS[92], Qi Long PhD[93], Sara Lozano-Zahonero PhD[20], Yuan Luo PhD[82], Kristine E Lynch PhD[72], Sadiqa Mahmood[41], Sarah E Maidlow AA[12], Adeline Makoudjou MD[20], Alberto Malovini PhD[27], Kenneth D Mandl MD, MPH[65], Chengsheng Mao PhD[82], Anupama Maram MS[94], Patricia Martel[95], Marcelo R Martins MSc[96], Jayson S Marwaha MD[97], Aaron J Masino PhD[98], Maria Mazzitelli PhD[99], Arthur Mensch[100], Marianna Milano PhD[101], Marcos F Minicucci MD, PhD[102], Bertrand Moal MD, PhD[13], Taha Mohseni Ahooyi PhD[103], Jason H Moore PhD[104], Cinta Moraleda MD,PhD[105], Jeffrey S Morris[106], Michele Morris BA6, Karyn L Moshal[107], Sajad Mousavi PhD[1], Danielle L Mowery PhD[9], Douglas A Murad[23], Shawn N Murphy MD, PhD[14], Thomas P Naughton BA[108], Carlos Tadeu Breda Neto[43], Antoine Neuraz MD, PhD[16], Jane Newburger MD, MPH[58], Kee Yuan Ngiam MBBS, FRCS[109], Wanjiku FM Njoroge MD[110], James B Norman[1], Jihad Obeid MD, FAMIA[68], Marina P Okoshi PhD[102], Karen L Olson PhD[111], Gilbert S. Omenn MD, PhD[19], Nina Orlova[56], Brian D Ostasiewski BS[112], Nathan P Palmer PhD[1], Nicolas Paris[56], Lav P Patel MS[7], Miguel Pedrera-Jiménez MS[52], Emily R Pfaff PhD[113], Ashley C Pfaff MD[114], Danielle Pillion MS[1], Sara Pizzimenti MS[35], Hans U Prokosch[115], Robson A Prudente PhD[116], Andrea Prunotto PhD[20], Víctor Quirós-González MS[52], Rachel B Ramoni[117], Maryna Raskin[41], Siegbert Rieg MD[118], Gustavo Roig-Domínguez MS[52], Pablo Rojo MD,PhD[119], Paula Rubio-Mayo MS[52], Paolo Sacchi MD[57], Carlos Sáez PhD[120], Elisa Salamanca[56], Malarkodi Jebathilagam Samayamuthu MD[6], L. Nelson Sanchez-Pinto MD, MBI[121], Arnaud Sandrin[56], Nandhini Santhanam MSc[75], Janaina C.C Santos MS[122], Fernando J Sanz Vidorreta[23], Maria Savino MS[123], Emily R Schriver MS[124], Petra Schubert MPH[22], Juergen Schuettler[125], Luigia Scudeller MD, MSc[35], Neil J Sebire MD, FRCPath[54], Pablo Serrano-Balazote MD,MS[52], Patricia Serre[56], Arnaud Serret-Larmande MD[126], Mohsin Shah MSc[54], Zahra Shakeri Hossein Abad PhD[1], Domenick Silvio[127], Piotr Sliz[65], Jiyeon Son MD[128], Charles Sonday[129], Andrew M South MD, MS[38], Anastasia Spiridou PhD[54], Zachary H. Strasser MD[15], Amelia LM Tan BSc, PhD[1], Bryce W.Q. Tan MBBS[5], Byorn W.L. Tan MBBS[5], Suzana E Tanni PhD[102], Deanne M Taylor PhD[130], Ana I Terriza-Torres MS[52], Valentina Tibollo MS[27], Patric Tippmann MSc[80], Emma MS Toh[33], Carlo Torti PhD[99], Enrico M Trecarichi PhD[99], Yi-Ju Tseng PhD[131], Andrew K Vallejos[132], Gael Varoquaux[133], Margaret E Vella BS[1], Guillaume Verdy MSc[13], Jill-Jênn Vie[134], Shyam Visweswaran MD, PhD[6], Michele Vitacca MD, PhD[135], Kavishwar B Wagholikar MBBS, PhD[36], Lemuel R Waitman[136], Xuan Wang PhD[1], Demian Wassermann[62], Griffin M Weber MD, PhD[1], Martin Wolkewitz PhD[80], Scott Wong[5], Zongqi Xia MD, PhD[4], Xin Xiong MS[3], Ye Ye BMED, MSPH, PhD[6], Nadir Yehya MD, MSCE[137], William Yuan PhD[1], Alberto Zambelli[138], Harrison G Zhang BA[1], Daniela Zöller PhD[20], Valentina Zuccaro MD[57], Chiara Zucco PhD[101]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, United States

[2]Department of Electrical Computer and Biomedical Engineering, University of Pavia, Pavia, Italy.

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, United States.

[4]Department of Neurology, University of Pittsburgh, Pittsburgh, United States.

[5] Department of Medicine, National University Hospital, Singapore, Singapore, Singapore.

[6]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, United States.

[7]Department of Internal Medicine, Division of Medical Informatics, University Of Kansas Medical Center, Kansas City, United States.

[8]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, United States.

[9]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States.

[10]Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States.

[11]Department of Pediatrics, Harvard Medical School, Boston, United States.

[12]Michigan Institute for Clinical and Health Research (MICHR) Informatics, University of Michigan, Ann Arbor, United States.

[13]IAM unit, Bordeaux University Hospital, Bordeaux, France.

[14]Department of Neurology, Massachusetts General Hospital, Boston, United States.

[15]Department of Medicine, Massachusetts General Hospital, Boston, United States.
of Paris, Paris, France.
[17]Department of Biomedical informatics, WiSDM, National University Health Systems Singapore, Singapore, Singapore.
[18]Department of Anaesthesia, National University Health Systems Singapore, Singapore, Singapore.
[19]Dept of Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, United States.
[20]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany.
[21]Department of Ophthalmology, Mayo Clinic, Rochester, United States.
[22]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, United States.
[23]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, United States.
[24]IT Department, Innovation & Data, APHP Greater Paris University Hospital, Paris, France.
[25]Division of Biomedical Informatics (Department of Internal Medicine), University of Kentucky, Lexington, United States.
[26]Department of Preventive Medicine, Northwestern University, Chicago, USA.
[27]Laboratory of Informatics and Systems Engineering for Clinical Research, Istituti Clinici Scientifici Maugeri SpA SB IRCCS, Pavia, Italy.
[28]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy.
[29]Population Health and Data Science, MAVERIC, VA Boston Healthcare System, Boston, United States.
[30]IAM unit, INSERM Bordeaux Population Health ERIAS TEAM, Bordeaux University Hospital / ERIAS - Inserm U1219 BPH, Bordeaux, France.
[31]Unit of Internal Medicine and Endocrinology, Istituti Clinici Scientifici Maugeri SpA SB IRCCS, Pavia, Italy.
[32]Departments of Biomedical Informatics, Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, United States.
[33]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.
[34]Department of Medicine, National University Health Systems Singapore, Singapore, Singapore.
[35]Scientific Direction, IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy.
[36]Department of Medicine, Massachusetts General Hospital, Boston, USA.
[37]BIOMERIS (BIOMedical Research Informatics Solutions), Pavia, Italy.
[38]Department of Pediatrics-Section of Nephrology, Brenner Children's, Wake Forest School of Medicine, Winston Salem, United States.
[39]Department of Biomedical Informatics, University of Kentucky, Lexington, United States.
[40]Department of Legal, Economic and Social Sciences, University Magna Graecia of Catanzaro, Catanzaro, Italy.
[41]Health Catalyst, INC., Cambridge, United States.
[42]Department of Surgery, ASST Pavia, Lombardia Region Health System, Pavia, Italy.
[43]Clinical Research Unit of Botucatu Medical School, São Paulo State University, Clinical Research Unit of Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[44]Pediatric emergency Department, Hôpital Necker-Enfants Malades, Assistance Public-Hôpitaux de Paris, Paris, France. [45]National Center for Infectious Diseases, Tan Tock Seng Hospital, Singapore, Singapore.
[46]BIG-ARC, The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, United States.
[47]Department of Pediatrics, Medical University of South Carolina, Charleston, United States.
[48]Internal Medicine Department, Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[49]Department of Surgery, St. Luke's University Health Network, Bethlehem, United States.
[50]Department of Medicine, Division of Nephrology, Ente Ospedaliero Cantonale, Lugano, Switzerland.
[51]IT Department, ASST Pavia, Voghera, Italy.
[52]Health Informatics, Hospital Universitario 12 de Octubre, Madrid, Spain.
[53]Strategy and Transformation Department, APHP Greater Paris University Hospital, Paris, France.
[54]Digital Research, Informatics and Virtual Environments (DRIVE), Great Ormond Street Hospital for Children, UK, London, United Kingdom.
[55]North Carolina Translational and Clinical Sciences (NC TraCS) Institute, UNC Chapel Hill, Chapel Hill, United States.
[56]IT department, Innovation & Data, APHP Greater Paris University Hospital, Paris, France.
[57]Division of Infectious Diseases I, Fondazione I.R.C.C.S. Policlinico San Matteo, Pavia, Italy.
[58]Department of Cardiology, Boston Children's Hospital, Harvard Medical School, Boston, United States.
[59]Department of Biomedical Informatics, HEGP, APHP Greater Paris University Hospital, Paris, France.
[60]Department of Medical and Surgical Sciences, Data Analytics Research Center, University Magna Graecia of Catanzaro, Catanzaro, Italy.
[61]Department of Anesthesia, St. Luke's University Health Network, Bethlehem, United States.
[62]Université Paris-Saclay, Inria, CEA, Palaiseau, France.
[63]INRIA Sophia-Antipolis – ZENITH team, LIRMM, Montpellier, France.
[64]Department of Internal Medicine, University of Kentucky, Lexington, United States.
[65]Computational Health Informatics Program, Boston Children's Hospital, Boston, United States.
[66]UOC Ricerca, Innovazione e Brand reputation, ASST Papa Giovanni XXIII, Bergamo, Bergamo, Italy.
[67]Informatics Institute, University of Alabama at Birmingham, Birmingham, United States.
[68]Biomedical Informatics Center, Medical University of South Carolina, Charleston, United States.
[69]Clinical Research Informatics, Boston Children's Hospital, Boston, United States.
[70]IT department, Innovation & Data (APHP), UMRS1142 (INSERM), APHP Greater Paris University Hospital, INSERM, Paris, France.
[71]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, United States.

COVID
[72]VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, United States.
[73]SED/SIERRA, Inria Centre de Paris, Paris, France.
[74]Health Information Technology & Services, University of Michigan, Ann Arbor, United States.
[75]Heinrich-Lanz-Center for Digital Health, University Medicine Mannheim, Heidelberg University, Mannheim, Germany.
[76]Department of Computational Biology and Bioinformatics, University of Michigan, Ann Arbor, United States.
[77]Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, United States.
[78]Department of Anesthesiology, Critical Care, and Pain Medicine and Computational Health Informatics Program, Boston Children's Hospital, Boston, United States.
[79]Institute of Digitalization in Medicine, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany.
[80]Institute of Medical Biometry and Statistics, Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Freiburg, Germany.
[81]Department of Biostatistics and Bioinformatics, Duke University, Durham, United States.
[82]Department of Preventive Medicine, Northwestern University, Chicago, United States.
[83]Department of Biomedical Informatics, HEGP, APHP Greater Paris University Hospital, Paris, France.
[84]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, United States.
[85]Medical University of South Carolina, Charleston, United States.
[86]Department of Pediatrics, Division of Human Genetics, The Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, United States.
[87]Center for Medical Information and Communication Technology, University Hospital Erlangen, Germany.
[88]Renaissance Computing Institute/Department of Computer Science, University of North Carolina, Chapel Hill, Chapel Hill, United States.
[89]Clinical Research Unit, Saint Antoine Hospital, APHP Greater Paris University Hospital, Paris, France.
[90]Clevy.io, Paris, France.
[91]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, United States.
[92]Department of Anaesthesia, National University Health Systems, Singapore, Singapore, Singapore.
[93]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States.
[94]Harvard Catalyst, Harvard Medical School, Boston, United States.
[95]Clinical Research Unit, Paris Saclay, APHP Greater Paris University Hospital, Boulogne-Billancourt, France.
[96]Medical Informatics Center, Hospital das Clínicas, Faculty of Medicine of Botucatu, Clinical Research Unit of Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[97]Department of Surgery, Beth Israel Deaconess Medical Center, Boston, United States.
[98]Department of Anesthesiology and Critical Care, Children's Hospital of Philadelphia, Philadelphia, United States.
[99]Department of Medical and Surgical Sciences, Infectious and Tropical Disease Unit, University Magna Graecia of Catanzaro, Catanzaro, Italy.
[100]ENS, PSL University, Paris, France.
[101]Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, Catanzaro, Italy.
[102]Internal Medicine Department of Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[103]Department of Biomedical Health Informatics, Children's Hospital of Philadelphia, Philadelphia, United States.
[104]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, United States.
[105]Pediatric Infectious Disease Department, Hospital Universitario 12 de Octubre, Madrid, Spain.
[106]Department of Biostatistics, Epidemiology, and Informatics, Institute for Biomedical Informatics,, University of Pennsylvania Perelman School of Medicine, Berwyn, United States.
[107]Department of Infectious Diseases, Great Ormond Street Hospital for Children, UK, London, United Kingdom.
[108]Harvard Catalyst | The Harvard Clinical and Translational Science Center, Harvard Medical School, Boston, United States.
[109]Department of Biomedical informatics, WiSDM, National University Health System Singapore, Singapore, Singapore.
[110]Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, United States.
[111]Computational Health Informatics Program and Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, United States.
[112]CTSI, WFBMI, Wake Forest School of Medicine, Winston Salem, United States.
[113]NC TraCS Institute, UNC Chapel Hill, Chapel Hill, United States.
[114]Department of Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, United States.
[115]Department of Medical Informatics, University of Erlangen-Nürnberg, Erlangen, Germany.
[116]Clinical Research Unit São Paulo State University, Brazil, Clinical Research Unit of Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[117]Office of Research and Development, Department of Veterans Affairs, Washington, DC, United States.
[118]Division of Infectious Diseases, Department of Medicine II, Medical Center – University of Freiburg, Faculty of Medicine, Freiburg, Germany.
[119]Pediatric Infectious Disease Deparment, Hospital Universitario 12 de Octubre, Madrid, Spain.
[120]Biomedical Data Science Lab, ITACA Institute, Universitat Politècnica de València, Spain, Valencia, Spain.
[121]Department of Pediatrics (Critical Care), Northwestern University Feinberg School of Medicine, Chicago, United States.
[122]Nurse departament of FMB - medicine school of Botucatu, Clinical Research Unit of Botucatu Medical School, São Paulo State University, Botucatu, Brazil.
[123]ASST Pavia, Lombardia Region Health System, Management Engineer, Direction, Pavia, Italy.
[124]Data Analytics Center, University of Pennsylvania Health System, Philadelphia, United States.
[125]Department of Anesthesiology, University Hospital Erlangen, FAU Erlangen-Nürnberg, Erlangen, Germany.
[126]Hôpital Saint Louis, Department of Biostatistics and Bioinformatics, APHP Greater Paris University Hospital, Paris, France.
[127]MICHR Informatics, University of Michigan, Ann Arbor, United States.

Harrison G Zhang[1,2,5], **Jacqueline P Honerlaw**[2], **Monika Maripuri**[2], **Malarkodi Jebathilagam Samayamuthu**[3], **Brendin R Beaulieu-Jones**[1], **Huma S Baig**[4], **Sehi L'Yi**[1], **Yuk-Lam Ho**[2], **Michele Morris**[3], **Vidul Ayakulangara Panickan**[1], **Xuan Wang**[1], **Griffin M Weber**[1], **Katherine P Liao**[2,5], **Shyam Visweswaran**[3], **Bryce W.Q. Tan**[6], **William Yuan**[1], **Nils Gehlenborg**[1], **Sumitra Muralidhar**[7], **Rachel B Ramoni**[7], **The Consortium for Clinical Characterization of COVID-19 by EHR (4CE)**[1], **Isaac S Kohane**[1,*], **Zongqi Xia**[8,*], **Kelly Cho**[2,*], **Tianxi Cai**[1,2,*], **Gabriel A Brat**[1,*,†]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States.

[2]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA, United States.

[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States.

[4]Department of Surgery, Beth Israel Deaconess Medical Center, Boston, MA, United States.

[5]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA, United States.

[6]Department of Medicine, National University Hospital, Singapore, Singapore, Singapore.

[7]Office of Research and Development, U.S. Department of Veterans Affairs, Washington, D.C., United States.

[8]Department of Neurology, University of Pittsburgh, Pittsburgh, PA, United States

The value of large-scale real-world data such as that from electronic health records (EHRs) has been used to establish vaccine efficacy, elucidate the genetic etiologies of diseases, and advance epidemiological research.[1–3] Real-world data also has the potential to capture the wide spectrum of clinical features attributed to post-acute sequalae of SARS-CoV-2, also called long COVID, in diverse patient populations.[4]

We are an international consortium that has operationalized definitions of long COVID using health agency guidelines and established a chart review procedure based on these definitions.[5] During this process, we identified 3 major challenges in using real world data to study long COVID: ambiguity and heterogeneity in clinical coding of long COVID; inadequacy of diagnostic codes in capturing the constellation of symptoms; and biases in

[128]Department of Neurology, University of Pittsburgh Medical Center, Pittsburgh, United States.
[129]Critical Care Medicine, Department of Medicine, St. Luke's University Health Network, Bethlehem, United States.
[130]Department of Biomedical Health Informatics and the Department of Pediatrics, The Children's Hospital of Philadelphia and the University of Pennsylvania Perelman Medical School, Philadelphia, United States.
[131]Department of Information Management, National Central University, Taoyuan, Taiwan.
[132]Clinical & Translational Science Institute, Medical College of Wisconsin, Milwaukee, United States.
[133]Université Paris-Saclay, Inria, CEA, Montréal Neurological Institute, McGill University, Palaiseau, France.
[134]SequeL, Inria Lille, Villeneuve-d'Ascq, France.
[135]Respiratory Department, ICS S. Maugeri IRCCS Pavia Italy, Lumezzane (BS), ITALY.
[136]Department of Health Management and Informatics, University of Missouri, Columbia, Columbia, United States.
[137]Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Philadelphia and University of Pennsylvania, Philadelphia, United States.
[138]Department of Oncology, ASST Papa Giovanni XXIII, Bergamo, Bergamo, Italy

Competing interests
The authors declare no competing interests

EHR data arising from variability in the number and kind of contacts with the healthcare system. These challenges warrant special attention if the clinical community wishes to arrive at a robust understanding of long COVID using evidence derived from real world data.

We performed a manual medical record review of 300 randomly sampled patients infected with SARS-CoV-2 and assigned an International Classification of Diseases (ICD)-10 code (U09.9) for long COVID at the Beth Israel Deaconess Medical Center, University of Pittsburgh Medical Center, and national U.S. Veterans Health Administration.[5] These three health systems collectively serve over 15 million patients each year.

We evaluated the extent to which patients with the ICD-10 code for this condition met our operationalized definitions from the World Health Organization's (WHO) and the US Centers for Disease Control.[5–7] Our long COVID definition based on WHO guidelines required that a patient present with at least two new-onset persistent symptoms lasting for 60 days after infection, whereas our definition based on CDC guidelines required that a patient present with at least one new-onset persistent symptom lasting for 30 days.[5]

A comparison of real-world EHR and administrative data with manually extracted clinical information (obtained through chart review of patients with the U09.9 code) found that functional definitions of long COVID varied widely by provider, which led to inconsistencies in coding practice and adherence to clinical definitions. Among patients assigned the U09.9 code, an average of 40.2% met the more stringent WHO definition, 58.3% had a single symptom that met the WHO definition, and 65.4% met the least stringent CDC definition.[5] This shows that the ICD-10 code is an unreliable surrogate of long COVID disease status in research. Research and policy efforts are needed to converge on a definition that will standardize coding practices and improve the ICD-10 code reliability.

Coding is further obfuscated by the potential for misclassification of long COVID with long-lasting complications from acute hospitalization, which are not specific to COVID-19.[8] We found an average of 42.3% patients assigned the U09.9 code were hospitalized after infection and an average of 12.3% received intensive care, both of which can produce long lasting symptoms that overlap with long COVID.5 Physical and physiological effects of hospitalization or critical care are important patient-level factors that should not be misattributed to SARS-CoV-2 infection.

Capturing long COVID symptomology using diagnosis codes is difficult, as the syndrome encompasses a constellation of non-specific symptoms including pain, fatigue, and brain fog that are not well represented by coding schemes such as ICD-10.[4–7] Leveraging textual data from EHRs may improve the ability to capture symptomatology. When we examined the data capture of symptoms by ICD-10 codes and natural language processing of clinical narratives, such as clinician notes and discharge summaries, we found that the incorporation of narrative data significantly improved identification of symptoms, compared to using diagnosis codes alone.[5] This shows the potential use of natural language processing techniques to ascertain a more complete representation of a patient's health.

A further challenge is the definition of long COVID patient cohorts, as the syndrome is defined by a time to presentation and therefore requires an index date from which to observe

clinical outcomes. The index date is usually an initial infection date, which may become increasingly difficult to ascertain with the use of at-home testing, the results of which are inconsistently reported in EHRs. Researchers should therefore perform routine quality controls (such as checking the time period between initial infection and input of the ICD-10 code for long COVID), to better understand biases present in the data. Researchers should also allow for some flexibility in defining index dates, such as considering an infection time period rather than a single date, which helps to account for delays in billing or data processing.

Researchers must remain cognizant of potential patient selection bias in real-world data; using visits to a long COVID clinic as a proxy for true disease status is problematic. We found that, on average, only 24.0% of patients assigned the U09.9 code visited a long COVID clinic, suggesting that the majority of patients sampled were being coded by physicians who do not work at these clinics.[5] Among patients who met the WHO definition of long COVID, only an average of 35.6% visited a long COVID clinic, suggesting that many patients are not being seen at these specialty care facilities.[5]

Studies should also account for differential data density and healthcare utilization. We found that patients who visited a long COVID clinic were on average annotated with more new-onset conditions when compared to patients who never visited a long COVID clinic.[5] Physicians working at long COVID clinics could be more experienced with the syndrome and therefore document the disease more thoroughly. This contributes to a difference in data density and granularity, which can confound findings if not properly addressed.

Studies of long COVID using real-world data must be based on robust and comprehensive clinical datasets. The incorporation of narrative data obtained using natural language processing techniques should better capture symptoms, and researchers should take caution when using only the ICD-10 code or a visit to a long COVID clinic as surrogates for disease status.

Computational phenotypes (where data elements are combined using machine learning algorithms to describe a particular disorder) have the potential to account for the longitudinal persistence of symptoms while avoiding the misattribution of conditions that existed prior to initial infection.[9] Semi-supervised machine learning algorithms are resistant to some of these challenges, and so may be powerful tools to capture complex underlying temporal patterns in the data using a small number of manually curated labels. Rule-based algorithms may be less suited for the inherent complexity of long COVID.[10]

Real-world data has an important role in supporting long COVID research, but these pitfalls should be considered so that the most equitable clinical and policy decisions can be informed by population-level studies.

REFERENCES

1. Haas EJ et al. The Lancet 397, 1819–1829 (2021).
2. McCarty CA et al. BMC Med Genomics 4, 13 (2011). [PubMed: 21269473]
3. Weber GM et al. J Med Internet Res 23, e31400 (2021). [PubMed: 34533459]

4. Zhang HG et al. NPJ Digit Med 5, 81 (2022). [PubMed: 35768548]
5. Zhang HG et al. 2023.02.12.23285701 Preprint at 10.1101/2023.02.12.23285701 (2023).
6. Soriano JB et al. Lancet Infect Dis 22, e102–e107 (2022). [PubMed: 34951953]
7. CDC. Post-COVID Conditions. Centers for Disease Control and Prevention https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (2022).
8. Clift AK et al. JAMA Psychiatry 79, 690–698 (2022). [PubMed: 35544272]
9. Liao KP et al. BMJ 350, h1885 (2015). [PubMed: 25911572]
10. Zhang Y et al. Nat Protoc 14, 3426–3444 (2019). [PubMed: 31748751]